A

AALEN'S ADDITIVE RISK MODEL.

See Additive Risk Model, Aalen's

ABAC

A graph from which numerical values may be read off, usually by means of a grid of lines corresponding to argument values.

See also Nomograms.

ABACUS

A simple instrument to facilitate numerical computation. There are several forms of abacus. The one in most common use at present is represented diagramatically in Fig. 1. It consists of a rectangular framework *ABCD* with a cross-piece *PQ* parallel to the longer sides, *AB* and *CD*, of the rectangle. There are a number (at least eight, often more) of thin rods or wire inserted in the framework and passing through *PQ*, parallel to the shorter sides, *AD* and *BC*. On each rod there are threaded four beads between *CD* and *PQ*, and one bead between *PQ* and *AB*.

Analogously to the meaning of position in our number system, the extreme right-hand rod corresponds to units; the next to the left, tens; the next to the left, hundreds; and so on. Each bead in the lower rectangle (*PQCD*) counts for 1, when moved up, and each bead in the upper rectangle (*ABQP* counts for 5. The number shown in Fig. 2 would be 852 if beads on all rods except the three extreme right-hand ones are as shown for the three extreme left-hand rods (corresponding to "zero").

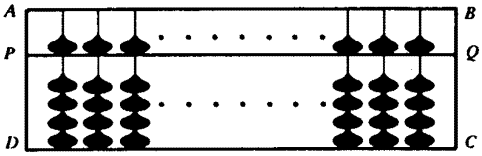


Figure 1. Diagrammatic representation of the form of abacus presently in common use.

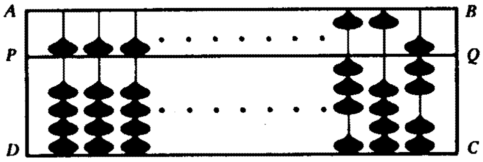


Figure 2. Abacus that would be showing the number 852 if beads on all rods except the three extreme right-hand ones are as shown for the three extreme left-hand rods (corresponding to "zero").

The Roman abacus consisted of a metal plate with two sets of parallel grooves, the lower containing four pebbles and the upper one pebble (with a value five times that of each pebble in the corresponding groove of the lower set). The Japanese and Chinese abacus (still in use) consists of a frame with beads on wires. The Russian abacus, which originated in the sixteenth century (the modern version in the eighteenth century), is also still in use.

BIBLIOGRAPHY

Dilson, J. (1968). *The Abacus: A Pocket Computer*. St. Martin's Press, New York.

Gardner, M. (1979). *Mathematical Circus*. Alfred A. Knopf, New York, Chap. 18.

Pullan, J. (1969). *The History of the Abacus*. F. A. Praeger, New York.

ABBE, ERNST

Born: January 23, 1840, in Eisenach, Germany.

Died: January 14, 1905, in Jena, Germany.

Contributed to: theoretical and applied optics, astronomy, mathematical statistics.

The recognition of Abbe's academic talent by those in contact with him overcame a childhood of privation and a financially precarious situation very early in his academic career, when he completed "On the Law of Distribution of Errors in Observation Series," his inaugural dissertation for attaining a lectureship at Jena University at the age of 23 [1]. This dissertation, partly motivated by the work of C. F. Gauss*, seems to contain his only contributions to the probability analysis of observations subject to error. These contributions constitute a remarkable anticipation of later work in distribution theory and time-series* analysis, but they were overlooked until the late 1960s [5,8], and almost none of the early bibliographies on probability and statistics (a notable exception being ref. 10) mention this work. In 1866, Abbe was approached by Carl Zeiss, who asked him to establish a scientific basis for the construction of microscopes; this was the beginning of a relationship that lasted throughout his life, and from this period on his main field of activity was optics [9] and astronomy*.

Abbe shows, first, that the quantity $\Delta = \sum_{i=1}^{n} Z_i^2$, where Z_i, $i = 1, \ldots, n$, are n independently and identically distributed $N(0, 1)$ random variables, is described by a chi-square* density with n degrees of freedom [5,8], although this discovery should perhaps be attributed to I. J. Bienaymé* [4]. Second, again initially by means of a "discontinuity factor" and then by complex variable methods, Abbe obtains the distribution of $\Theta = \sum_{j=1}^{n} (Z_j - Z_{j+1})^2$, where $Z_{n+1} = Z_1$, and ultimately that of Θ/Δ, a ratio of quadratic forms* in $Z_1, \ldots, Z_n$ very close in nature to the definition of what is now called the first circular serial correlation* coefficient, and whose distribution under the present conditions is essentially that used to test the null hypothesis of Gaussian white noise* against a first-order autoregression alternative, in time-series* analysis [3]. (The distribution under such a null hypothesis was obtained by R. L. Anderson in 1942.) Knopf [6] expresses Abbe's intention in his dissertation as being to seek a numerically expressible criterion to determine when differences between observed and sought values in a series of observations are due to chance alone.

REFERENCES

1. Abbe, E. (1863). *Über die Gesetzmässigkeit der Vertheilung der Fehler bei Beobachtungsreihen*. Hab. schrift., Jena (Reprinted as pp. 55–81 of ref. 2.)
2. Abbe, E. (1906). *Gesammelte Abhandlungen*, Vol. 2. G. Fischer, Jena.
3. Hannan, E. J. (1960). *Time Series Analysis*. Methuen, London, pp. 84–86.
4. Heyde, C. C. and Seneta, E. (1977). *I. J. Bienaymé: Statistical Theory Anticipated*. Springer-Verlag, New York, p. 69.
5. Kendall, M. G. (1971). *Biometrika*, **58**, 369–373. (Sketches Abbe's mathematical reasoning in relation to the contributions to mathematical statistics.)
6. Knopf, O. (1905). *Jahresber. Dtsch. Math.-Ver.*, 14, 217–230. [One of several obituaries by his associates; nonmathematical, with a photograph. Another is by S. Czapski (1905), in *Verh. Dtsch. Phys. Ges.*, 7, 89–121.]
7. Rohr, L. O. M. von (1940). *Ernst Abbe*. G. Fischer, Jena. (Not seen.)
8. Sheynin, O. B. (1966). *Nature (Lond.)*, **211**, 1003–1004. (Notes Abbe's derivation of the chi-square density.)
9. Volkman, H. (1966). *Appl. Opt.*, **5**, 1720–1731. (An English-language account of Abbe's life and contributions to pure and applied optics; contains two photographs of Abbe, and further bibliography.)
10. Wölffing, E. (1899). *Math. Naturwiss. Ver. Württemberg* [Stuttgart], *Mitt.*, (2) **1**, 76–84. [Supplements the comprehensive bibliography given by E. Czuber (1899), *in Jahresber. Dtsch. Math.-Ver.*, **7** (2nd part), 1–279.]

See also CHI-SQUARE DISTRIBUTION; QUADRATIC FORMS; SERIAL CORRELATION; and TIME SERIES ANALYSIS AND FORECASTING SOCIETY.

E. SENETA

ABEL'S FORMULA

(Also known as the Abel identity.) If each term of a sequence of real numbers $\{a_i\}$ can be represented in the form $a_i = b_i c_i, i = 1, \ldots, n$, then $a_1 + a_2 + \cdots + a_n$ can be expressed as

$$s_1(b_1 - b_2) + s_2(b_2 - b_3) + \cdots + s_{n-1}(b_{n-1} - b_n),$$

where $s_i = c_1 + \cdots + c_i$. Equivalently,

$$\sum_{k=n}^{m} b_k c_k = B_m c_{m+1} - B_{n-1} c_n + \sum_{k=n}^{m} B_k (c_k - c_{k+1}),$$

where $B_k = \sum_{l=1}^{k} b_l$.

This representation is usually referred to as Abel's formula, due to Norwegian mathematician Niels Henrik Abel (1802–1829). (The continuous analog of this formula is the formula of integration by parts.) It is useful for manipulations with finite sums.

BIBLIOGRAPHY

Knopp, K. (1951). *Theory and Application of Infinite Series*, 2nd ed. Blackie, London/Dover, New York.

ABSOLUTE ASYMPTOTIC EFFICIENCY (AAE). See ESTIMATION, CLASSICAL

ABSOLUTE CONTINUITY

Absolute continuity of measures, the Radon–Nikodym theorem*, and the Radon–Nikodym derivative* are subjects properly included in any basic text on measure and integration. However, both the mathematical theory and the range of applications can best be appreciated when the measures are defined on an infinite-dimensional linear topological space. For example, this setting is generally necessary if one wishes to discuss hypothesis testing* for stochastic processes with infinite parameter set. In this article we first define basic concepts in the area of absolute continuity, state general conditions for absolute continuity to hold, and then specialize to the case where the two measures are defined on either a separable Hilbert space or on an appropriate space of functions. Particular attention is paid to Gaussian measures.

The following basic material is discussed in many texts on measure theory*; see, e.g., ref. 23. Suppose that (Ω, β) is a measurable space, and that μ_1 and μ_2 are two probability measures on (Ω, β). μ_1 is said to be *absolutely continuous* with respect to μ_2 ($\mu_1 \ll \mu_2$) if A in β and $\mu_2(A) = 0$ imply that $\mu_1(A) = 0$. This is equivalent to the following: $\mu_1 \ll \mu_2$ if and only if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\mu_2(A) < \delta$ implies that $\mu_1(A) \leqslant \epsilon$. Similar definitions of absolute continuity can be given for nonfinite signed measures; this article, however, is restricted to probability measures. When $\mu_1 \ll \mu_2$, the Radon–Nikodym theorem* states that there exists a real-valued β-measurable function f such that $\mu_1(A) = \int_A f\, d\mu_2$ for all A in β. The function f, which belongs to $L_1[\Omega, \beta, \mu_2]$ and is unique up to μ_2-equivalence, is called the Radon–Nikodym derivative of μ_1 with respect to μ_2, and is commonly denoted by $d\mu_1/d\mu_2$. In statistical and engineering applications $d\mu_1/d\mu_2$ is usually called the likelihood ratio*, a term that has its genesis in maximum likelihood estimation*.

Absolute continuity and the Radon–Nikodym derivative have important applications in statistics. For example, suppose that $X : \Omega \to \mathbb{R}^N$ is a random vector. Suppose also that under hypothesis H_1 the distribution function of X is given by $F_1 = \mu_1 \circ X^{-1}[F_1(x) = \mu_1\{\omega : X(\omega) \leqslant x\}]$, whereas under H_2, X has the distribution function $F_2 = \mu_2 \circ X^{-1}$. F_i defines a Borel measure on $\mathbb{R}^1$; one says that F_i is induced from μ_i by X. A statistician observes one realization (sample path) of X, and wishes to design a statistical test to optimally decide in favor of H_1 or H_2. Then, under any of several classical decision criteria of mathematical statistics (e.g., Bayes risk, Neyman–Pearson*, minimum probability of error), an optimum decision procedure when $\mu_1 \ll \mu_2$ is to form the test statistic* $\Lambda(X) = [dF_1/dF_2](X)$ and compare its value with some constant, C_0; the decision is then to accept H_2 if $\Lambda(X) \leqslant C_0$, accept H_1 if $\Lambda(X) > C_0$. The value of C_0 will depend on the properties of F_1 and F_2 and on the optimality criterion. For more details, see HYPOTHESIS TESTING*.[1]

Two probability measures μ_1 and μ_2 on (Ω, β) are said to be equivalent ($\mu_1 \sim \mu_2$) if $\mu_1 \ll \mu_2$ and $\mu_2 \ll \mu_1$. They are orthogonal, or extreme singular ($\mu_1 \perp \mu_2$) if there exists a set A in β such that $\mu_2(A) = 0$ and $\mu_1(A) = 1$. For the hypothesis-testing problem discussed above, orthogonal induced measures permit one to discriminate perfectly between H_1 and H_2. In many practical applications, physical considerations rule out perfect discrimination. The study of conditions for absolute continuity then becomes important from the aspect of verifying that the mathematical model is valid.

In the framework described, the random vector has range in $\mathbb{R}^N$. However, absolute continuity, the Radon–Nikodym derivative, and their application to hypothesis-testing problems are not limited to such finite-dimensional cases. In fact, the brief comments above on hypothesis testing apply equally well when X takes its value in an infinite-dimensional linear topological space, as when $X(\omega)$ represents a sample path* from a stochastic process* (X_t), $t \in [a, b]$. (The infinite-dimensional case does introduce interesting mathematical complexities that are not present in the finite-dimensional case.)

GENERAL CONDITIONS FOR ABSOLUTE CONTINUITY

We shall see later that special conditions for absolute continuity can be given when the two measures involved have certain specialized properties, e.g., when they are both Gaussian. However, necessary and sufficient conditions for absolute continuity can be given that apply to any pair of probability measures on any measurable space (Ω, β). Further, if (Ω, β) consists of a linear topological space Ω and the smallest σ-field β containing all the open sets (the Borel σ-field), then additional conditions for absolute continuity can be obtained that apply to any pair of probability measures on (Ω, β). Here we give one well-known set of general necessary and sufficient conditions. First, recall that if (Ω, β, P) is a probability space and F a collection of real random variables on (Ω, β), then F is said to be uniformly integrable with respect to P [23] if the integrals $\int_{\{\omega:|f(\omega)|\geqslant c\}} |f(\omega)| dP(\omega), c > 0, f$ in F, tend uniformly to zero as $c \to \infty$. An equivalent statement is the following: F is uniformly integrable (P) if and only if

(a) $$\sup_F \int_\Omega |f(\omega)| dP(\omega) < \infty$$

and

(b) For every $\epsilon > 0$ there exists $\delta > 0$ such that $P(A) < \delta$ implies that

$$\sup_F \int_A |f(\omega)| dP(\omega) \leqslant \epsilon.$$

Theorem 1. Suppose that μ_1 and μ_2 are two probability measures on a measurable space (Ω, β). Suppose that $\{\mathscr{F}_n, n \geqslant 1\}$ is an increasing family of sub-σ-fields of β such that β is the smallest σ-field containing $\cup_n \mathscr{F}_n$. Let μ_i^n be the restriction of μ_i to $\mathscr{F}_n$. Then $\mu_1 \ll \mu_2$ if and only if

(a) $$\mu_1^n \ll \mu_2^n \quad \text{for all } n \geqslant 1,$$

and

(b) $$\{d\mu_1^n / d\mu_2^n, n \geqslant 1\}$$

is uniformly integrable (μ_2).

When $\mu_1 \ll \mu_2$, then $d\mu_1/d\mu_2 = \lim_n d\mu_1^n / d\mu_2^n$ almost everywhere (a.e.) $d\mu_2$.

Condition (a) of Theorem 1 is obviously necessary. The necessity of (b) follows from the fact that $\{d\mu_1^n/d\mu_2^n, \mathscr{F}_n : n \geqslant 1\}$ is a martingale* with respect to μ_2. This property, and the martingale convergence theorem, yield the result that $d\mu_1/d\mu_2 = \lim_n d\mu_1^n/d\mu_2^n$ a.e. $d\mu_2$. Sufficiency of (a) and (b) follows from the second definition of uniform integrability given above and the assumption that β is the smallest σ-field containing $\cup_n \mathscr{F}_n$.

Conditions (a) and (b) of Theorem 1 are also necessary and sufficient for $\mu_1 \ll \mu_2$ when the family of increasing σ-fields $(\mathscr{F}_t)$ has any directed index set.

A number of results frequently used to analyze absolute continuity can be obtained from Theorem 1. This includes, for example, Hájek's divergence criterion [20] and Kakutani's theorem on equivalence of infinite product measures [29] (a fundamental result in its own right).

The conditions of Theorem 1 are very general. However, in one respect they are somewhat unsatisfactory. They usually require that one specify an infinite sequence of Radon–Nikodym derivatives $\{d\mu_1^n/d\mu_2^n, n \geqslant 1\}$. It would be preferable to have a more direct method of determining if absolute continuity holds. One possible alternative when the measures are defined on a separable metric space involves the use of characteristic functions*. The characteristic function of a probability measure defined on the Borel σ-field of a separable metric space completely and uniquely specifies the measure [38]. Thus in such a setting, two characteristic functions contain all the information required to determine whether absolute continuity exists between the associated pair of measures. The use of characteristic functions offers a method for attacking the following problem. For a given measure μ on (Ω, β) determine the set $\mathscr{P}_\mu$ of all probability measures on (Ω, β) such that $\nu \ll \mu$ for all ν in $\mathscr{P}_\mu$. Some results on this problem are contained in ref. 3; further progress, especially detailed results for the case of a Gaussian measure μ on Hilbert space, would be useful in several important applications areas (detection of signals in noise, stochastic filtering, information theory*).

PROBABILITY MEASURES ON HILBERT SPACES

There has been much activity in the study of probability measures on Banach spaces [1,4,5,31]. Here we restrict attention to the case of probabilities on Hilbert spaces; this is the most important class of Banach spaces for applications, and the theory is relatively well developed in this setting.

Let $\mathbf{H}$ be a real separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and Borel σ-field Γ (*see* SEPARABLE SPACE). Let μ be a probability measure on Γ. For any element y in $\mathbf{H}$, define the distribution function F_y by $F_y(a) = \mu\{x : \langle y, x\rangle \leqslant a\}$, a in $(-\infty, \infty)$. μ is said to be *Gaussian* if F_y is Gaussian for all y in $\mathbf{H}$. It can be shown that for every Gaussian μ there exists a self-adjoint trace-class nonnegative linear operator R_μ in $\mathbf{H}$ and an element m_μ in $\mathbf{H}$ such that

$$\langle y, m_\mu\rangle = \int_{\mathbf{H}} \langle y, x\rangle d\mu(x) \tag{1}$$

and

$$\langle R_\mu, v\rangle = \int_{\mathbf{H}} \langle y - m_\mu, x\rangle \langle v - m_\mu, x\rangle \, d\mu(x) \tag{2}$$

for all y and v in $\mathbf{H}$. R_μ is called the *covariance (operator)* of μ, and m_μ is the *mean (element)*. Conversely, to every self-adjoint nonnegative trace-class operator R_μ and element m in $\mathbf{H}$ there corresponds a unique Gaussian measure μ such that relations (1) and (2) are satisfied. Non-Gaussian measures μ may also have a covariance operator R_μ and mean element m_μ satisfying (1) and (2); however, the covariance R_μ need not be trace-class. For more details on probability measures on Hilbert space, see refs. 17, 38, and 53.

Elegant solutions to many problems of classical probability theory (and applications) have been obtained in the Hilbert space framework, with methods frequently making use of the rich structure of the theory of linear operators. Examples of such problems include Sazanov's solution to obtaining necessary and sufficient conditions for a complex-valued function on $\mathbf{H}$ to be a characteristic function* [49]; Prohorov's conditions for weak compactness of families of probability measures, with applications to convergence of stochastic processes [43]; the results of Mourier on laws of large numbers* [34]; the results of Fortét and Mourier on the central limit theorem* [15,34]; and conditions for absolute continuity of Gaussian measures. The latter problem is examined in some detail in the following section. The study of probability theory in a Hilbert space framework received much of its impetus from the pioneering work of Fortét and Mourier (see refs. 15 and 34, and the references cited in those papers). Their work led not only to the solution of many interesting problems set in Hilbert space, but also to extensions to Banach spaces and more general linear topological spaces [1,4,5,15,31,34].

The infinite-dimensional Hilbert spaces $\mathbf{H}$ most frequently encountered in applications are $L_2[0, T]$ ($T < \infty$) and l_2. For a discussion of how Hilbert spaces frequently arise in engineering applications, *see* COMMUNICATION THEORY, STATISTICAL. In particular, the interest in Gaussian measures on Hilbert space has much of its origin in hypothesis testing and estimation problems involving stochastic processes: detection and filtering of signals embedded in Gaussian noise. For many engineering applications, the noise can be realistically modeled as a Gaussian stochastic process with sample paths almost surely (a.s.) in $L_2[0, T]$ or a.s. in l_2. When $\mathbf{H}$ is $L_2[0, T]$, a trace-class covariance operator can be represented as an integral operator whose kernel is a covariance function. Thus suppose that (X_t), $t \in [0, T]$, is a measurable zero-mean stochastic process on (Ω, β, P), inducing the measure μ on the Borel σ-field of $L_2[0, T]$; $\mu(A) = P\{\omega : X(\omega) \in A\}$. Then $E\int_0^T X_t^2(\omega)\,dt < \infty$ if and only if μ has a trace-class covariance operator R_μ defined by $[R_\mu f](t) = \int_0^T R(t, s)f(s)\,ds, f$ in $L_2[0, T]$, where R is the covariance function of (X_t). If R_μ is trace-class, then $E\int_0^T X_t^2(\omega)\,dt =$ trace R_μ.

ABSOLUTE CONTINUITY OF PROBABILITY MEASURES ON HILBERT SPACE

If **H** is finite-dimensional and μ_1 and μ_2 are two zero-mean Gaussian measures on Γ, it is easy to see that μ_1 and μ_2 are equivalent if and only if their covariance matrices have the same range space. However, if **H** is infinite-dimensional, this condition (on the ranges of the covariance operators) is neither necessary nor sufficient for $\mu_1 \sim \mu_2$. The study of conditions for absolute continuity of two Gaussian measures on function space has a long and active history. Major early contributions were made by Cameron and Martin [6,7] and by Grenander [18]. The work of Cameron and Martin was concerned with the case when one measure is Wiener measure (the measure induced on $C[0, 1]$ by the Wiener process*) and the second measure is obtained from Wiener measure by an affine transformation. Grenander obtained conditions for absolute continuity of a Gaussian measure (induced by a stochastic process with continuous covariance) with respect to a translation. Segal [50] extended the work of Cameron and Martin to a more general class of affine transformations of Wiener measure. Segal also obtained [50] conditions for absolute continuity of Gaussian "weak distributions." These necessary and sufficient conditions can be readily applied to obtain sufficient conditions for equivalence of any pair of Gaussian measures on **H**; they can also be used to show that these same conditions are necessary. Complete and general solutions to the absolute continuity problem for Gaussian measures were obtained by Feldman [12] and Hájek [21]. Their methods are quite different. The main result, in each paper, consists of two parts: a "dichotomy theorem," which states that any two Gaussian measures are either equivalent or orthogonal; and conditions that are necessary and sufficient for equivalence.

The following theorem for Gaussian measures on Hilbert space is a modified version of Feldman's result [12]; several proofs have been independently obtained (Kallianpur and Oodaira [30], Rao and Varadarajan [44], Root [45]).

Theorem 2. Suppose that μ_1 and μ_2 are two Gaussian measures on Γ, and that μ_i has covariance operator R_i and mean $m_i, i = 1, 2$. Then:

1. either $\mu_1 \sim \mu_2$ or $\mu_1 \perp \mu_2$;
2. $\mu_1 \sim \mu_2$ if and only if all the following conditions are satisfied:
   (a) range $(R_1^{1/2}) =$ range $(R_2^{1/2})$;
   (b) $R_1 = R_2^{1/2}(I + T)R_2^{1/2}$, where I is the identity on **H** and T is a Hilbert–Schmidt operator in **H**:
   (c) $m_1 - m_2$ is in range $(R_1^{1/2})$.

Various specializations of Theorem 2 have been obtained; see the references in refs. 8 and 47. Two of the more interesting special cases, both extensively analyzed, are the following: (1) both measures induced by stationary Gaussian stochastic processes; (2) one of the measures is Wiener measure. In the former case, especially simple conditions can be given when the two processes have rational spectral densities; see the papers by Feldman [13], Hájek [22], and Pisarenko [40,41]. In this case, when the two measures have the same mean function, $\mu_1 \sim \mu_2$ if and only if $\lim_{|\lambda|\to\infty} f_1(\lambda)/f_2(\lambda) = 1$, where f_i is the spectral density* of the Gaussian process inducing μ_i. Moreover, this occurs if and only if the operator T appearing in Theorem 2 is also trace-class [22]. For the case where one of the measures is Wiener measure, see the papers by Shepp [51], Varberg [54,55], and Hitsuda [24].

The problem of determining the Radon–Nikodym derivative for two equivalent Gaussian measures on a Hilbert space has been studied, especially by Rao and Varadarajan [44]. For convenience, we use the notation of Theorem 2 and assume now that all covariance operators are strictly positive. In the case where the Hilbert space is finite-dimensional, the log of the Radon–Nikodym derivative $d\mu_1/d\mu_2$ (log-likelihood ratio*) is easily seen to be a quadratic-linear form; that is, $\log \Lambda(X) = \langle x, Wx\rangle + \langle x, b\rangle + \text{constant}$, where the linear operator $W = \frac{1}{2}(R_2^{-1} - R_1^{-1})$, $b = R_1^{-1}m_1 - R_2^{-1}m_2$, and $\log \equiv \log_e$. However, when $\mathbf{H}$ is infinite-dimensional, the log-likelihood ratio need not be a quadratic-linear form defined by a bounded linear operator. This holds true even if the operator T of Theorem 2 is not only Hilbert–Schmidt, but is also trace class. However, when T is Hilbert–Schmidt, one can always express the log of the Radon–Nikodym derivative as an almost surely convergent series [44]. The essential difficulty in characterizing the likelihood ratio for infinite-dimensional Hilbert space is that the operators R_1 and R_2 cannot have *bounded* inverses and these two inverses need not have the same domain of definition. Even if range $(R_1) =$ range (R_2), so that $R_2^{-1} - R_1^{-1}$ is defined on range (R_1), it is not necessary that $R_2^{-1} - R_1^{-1}$ be bounded on range (R_1).

In the finite-dimensional case, if $R_1 = R_2$, then $\log \Lambda(X) = \langle x, b\rangle +$ constant, with b defined as above, so that the log-likelihood ratio is a bounded linear form. This need not be the case for infinite-dimensional Hilbert space; in general, $\log \Lambda(X)$ will be a bounded linear form (when $R_1 = R_2$) if and only if $m_1 - m_2$ is in the range of R_1. As can be seen from Theorem 1, this condition is strictly stronger than the necessary and sufficient condition for $\mu_1 \sim \mu_2$, which (with $R_1 = R_2$) is that $m_1 - m_2$ be in range $(R_1^{1/2})$.

If the two measures are induced by stationary Gaussian processes with rational spectral densities, expressions for the likelihood ratio can be given in terms of the spectral densities; see the papers by Pisarenko [41] and Hájek [22].

In many applications, only one of the two measures can be considered to be Gaussian. For this case, a useful sufficient condition for absolute continuity is given in ref. 2. This condition can be applied when the two measures are induced by stochastic processes (X_t) and (Y_t), where (Y_t) is a function of (X_t) and a process (Z_t) that is independent of (X_t). In particular, if (X_t) is Gaussian and $(Y_t) = (X_t + Z_t)$, then conditions for absolute continuity can be stated in terms of sample path properties of the (Z_t) process (absolute continuity, differentiability, etc.). Such conditions can often be verified in physical models by knowledge of the mechanisms generating the observed data, when the distributional properties of the (Z_t) process are unknown. When (X_t) is the Wiener process on [0, T], conditions for absolute continuity of the induced measures on $L_2[0, T]$ can be obtained from the results of refs. 10, 27, and 28. Some of these results do not require independence of (X_t) and (Z_t).

Other results on absolute continuity of measures on Hilbert space have been obtained for infinitely divisible measures [16], measures induced by stochastic processes with independent increments [16], admissible translations of measures [42,52], and for a fixed measure and a second measure obtained from the first measure by a nonlinear transformation [16]. With respect to admissible translates, Rao and Varadarajan [44] have shown that if μ is a zero-mean measure having a trace-class covariance operator, R, then the translate of μ by an element y is orthogonal to μ if y is not in range $(R^{1/2})$. A number of these results are collected in the book by Gihman and Skorohod [17], which also contains much material on basic properties of probability measures on Hilbert space, and on weak convergence*. The book by Kuo [33] contains not only basic material on probability measures on Hilbert spaces (including absolute continuity), but also an introduction to some topics in probability on Banach spaces.

ABSOLUTE CONTINUITY OF MEASURES INDUCED BY STOCHASTIC PROCESSES

Many problems involving stochastic processes are adequately modeled in the framework of probability measures on Hilbert space, provided that the sample paths of each process of interest belong almost surely to some separable Hilbert space. However, this condition is not always satisfied; even when it is satisfied, one may prefer conditions for absolute continuity stated in terms of measures on $\mathbb{R}^T$ (the space of real-valued functions on T), where T is the parameter set of the process. For example, a class of stochastic processes frequently considered are those having almost all paths in D [0, 1]. D [0, 1] is the set of all real-valued functions having limits from both left and right existing at all points of (0, 1), with either left-continuity or right-continuity at each point of (0, 1), and with a limit from the left (right) existing at 1(0). D [0, 1] is a linear metric space* under the Skorohod metric [38], but this metric space is not a Hilbert space.

The general conditions for absolute continuity stated in Theorem 1 apply in any setting. Moreover, necessary and sufficient conditions for equivalence of measures (most frequently on $\mathbb{R}^T$) induced by two Gaussian stochastic processes can be stated in a number of ways: The reproducing kernel Hilbert space (r.k.H.s.) of the two covariance functions [30,37,39]; operators and elements in an L_2 space of real-valued random functions [12]; operators and elements in an L_2-space of random variables [46]; and tensor products [35]. Hájek's conditions for absolute continuity in terms of the divergence [21] apply to the general case. Sato [48] has stated conditions for absolute continuity in terms of a representation for all Gaussian processes whose induced measure on $\mathbb{R}^T$ is equivalent to the measure induced by a given Gaussian process. Several of these results are presented in [8]. Many other papers on absolute continuity for measures induced by two Gaussian processes have appeared; space does not permit an attempt at a complete bibliography.

Use of the r.k.H.s. approach to study linear statistical problems in stochastic processes was first explicitly and systematically employed by Parzen; the r.k.H.s. approach was also implicit in the work of Hájek (see the papers by Hájek [22] and Parzen [39] and their references).

For non-Gaussian processes, results on absolute continuity have been obtained for Markov processes* [16,32], diffusion processes* [36], locally infinitely divisible processes [16], semimartingales* [25], point processes* [26], and non-Gaussian processes equivalent to the Wiener process [9,10,27,28].

Dudley's result [9] is of particular interest to researchers interested in Gaussian measures. Suppose that (W_t) is the Wiener process on [0,1] with zero mean and unity variance parameter, and that $\beta(\cdot,\cdot)$ is a continuous real-valued function on $\mathbb{R} \times [0, 1]$. Let $Y_t = \beta(W_t, t)$. Dudley shows in ref. 9 that the measure on function space induced by (Y_t) is absolutely continuous with respect to Wiener measure if and only if $\beta(u, t) = u + \phi(t)$ or $\beta(u, t) = -u + \phi(t)$, where ϕ is in the r.k.H.s. of the Wiener covariance min(t, s). The methods used to prove this result rely heavily on some of the special properties of the Wiener process, such as the fact that (W_t) has the strong Markov property, and laws of the iterated logarithm* for the Wiener process (obtained in ref. 9). A characterization of admissible β's for other Gaussian processes with continuous paths would be of much interest; such characterizations would necessarily require a different approach, and this problem is very much open at present.

The absolute continuity problem discussed in refs. 10, 27, and 28 has received much attention, partly because of its connection to signal detection* and nonlinear filtering*. One considers a measurable process (Y_t) defined by $Y_t = \int_0^t h_s\,ds + W_t, 0 \leqslant t \leqslant T$, where (W_t) is a zero-mean Wiener process and (h_s) is a stochastic process with sample paths a.s. in $L_1[0, T]$. Let μ_Y and μ_W be the measures induced by (Y_t) and (W_t) on the space of continuous functions on [0, 1]. Conditions for $\mu_Y \ll \mu_W, \mu_Y \sim \mu_W$, and results on the Radon–Nikodym derivative have been obtained in refs. 10, 27, and 28. In the special case where (h_s) is independent of (W_t), a sufficient condition for $\mu_Y \sim \mu_W$ is that $\int_0^T h_s^2 ds < \infty$ for almost all sample paths of (h_s). This condition is also sufficient for $\mu_Y \ll \mu_W$ if the process (h_s) is only assumed independent of future increments of (W_t).

Finally, we mention a result of Fortét [14], who has obtained a sufficient condition for orthogonality of two measures when one is Gaussian, expressed in terms of the r.k.H.s. of the two covariances. Suppose that μ_i is a probability measure on $\mathbb{R}^T, T = [0, 1]$, with r.k.H.s. H_i and mean function m_i. Then if μ_1 is Gaussian, μ_1 and μ_2 are orthogonal unless both the following conditions are satisfied: (a) $H_1 \subset H_2$; and (b) $m_1 - m_2 \in H_2$.

NOTE

1. The ubiquitous nature of the Radon–Nikodym derivative in various hypothesis-testing applications can be attributed to its being a necessary and sufficient statistic* [11].

REFERENCES

1. Aron, R. M. and Dineen, S., eds. (1978). Vector Space Measures and Applications, I. *Lect. Notes Math.*, 644. (Contains most of the papers on probability theory presented at the Conference on Vector Space Measures and Applications, Dublin, 1978.)
2. Baker, C. R. (1973). *Ann. Prob.*, **1**, 690–698.
3. Baker, C. R. (1979). *Lect. Notes Math.*, **709**, 33–44.
4. Beck, A., ed. (1976). Probability in Banach spaces. *Lect. Notes Math.*, 526. (Collection of papers on various topics; Proceedings of First International Conference on Probability in Banach Spaces, Oberwolfach, 1975.)
5. Beck, A., ed. (1979). Probability in Banach Spaces, II. *Lect. Notes Math.* 709. (Proceedings of Second International Conference on Probability in Banach Spaces, Oberwolfach, 1978.)
6. Cameron, R. H. and Martin, W. T. (1944). *Ann. Math.*, **45**, 386–396.
7. Cameron, R. H. and Martin, W. T. (1945). *Trans. Amer. Math. Soc.*, **58**, 184–219.
8. Chatterji, S. D. and Mandrekar, V. (1978). In *Probabilistic Analysis and Related Topics*, Vol. 1, A. T. Bharucha-Reid, ed. Academic Press, New York, pp. 169–197. (Extensive bibliography.)
9. Dudley, R. M. (1971). *Zeit. Wahrscheinlichkeitsth.*, **20** 249–258; correction: *ibid.*, **30** (1974), 357–358.
10. Duncan, T. E. (1970). *Inf. Control*, **16**, 303–310.
11. Dynkin, E. B. (1951). *Uspehi Mat. Nauk* (N.S.), **6**, 68–90; translation: *Select. Transl. Math. Statist. Prob.* **1**, 17–40, (1961).
12. Feldman, J. (1958). *Pacific J. Math.*, **8**, 699–708; correction: *ibid.*, **9**, 1295–1296 (1959).
13. Feldman, J. (1960). *Pacific J. Math.*, **10**, 1211–1220.
14. Fortét, R. (1973). *Ann. Inst. Henri Poincaré*, **9**, 41–58.
15. Fortét, R. and Mourier, E. (1955). *Studia Math.*, **15**, 62–79.
16. Gihman, I. I. and Skorohod, A. V. (1966). *Russ. Math. Surv. (Uspehi Mat. Nauk.)*, **21**, 83–156.
17. Gihman, I. I. and Skorohod, A. V. (1974). *The Theory of Stochastic Processes*, Vol. 1. Springer-Verlag, New York. (Basic material on probability on metric spaces; many results on absolute continuity; also has extensive treatment of weak convergence of probability measures, and applications to convergence of stochastic processes.)
18. Grenander, U. (1950). *Ark. Mat.*, **1**, 195–277.
19. Grigelonis, B. (1973). *Lect. Notes Math.*, **330**, 80–94.
20. Hájek, J. (1958). *Czech. Math. J.*, **8**, 460–463.
21. Hájek, J. (1958). *Czech. Math. J.*, **8**, 610–618.
22. Hájek, J. (1962). *Czech. Math. J.*, **12**, 404–444.
23. Hewitt, E. and Stromberg, K. (1965). *Real and Abstract Analysis*. Springer-Verlag, New York.
24. Hitsuda, M. (1968). *Osaka J. Math.*, **5**, 299–312.
25. Jacod, J. and Mémin, J. (1976). *Zeit Wahrscheinlichkeitsth.*, **35**, 1–37.
26. Kabanov, Yu. M., Liptser, R. S., and Shiryayev, A. N. (1973). *Lect. Notes Math.*, **550**, 80–94.
27. Kadota, T. T. and Shepp, L. A. (1970). *Zeit. Wahrscheinlichkeitsth.*, **16**, 250–260.
28. Kailath, T. and Zakai, M. (1971). *Ann. Math. Statist.*, **42**, 130–140.
29. Kakutani, S. (1948). *Ann. Math.*, **49**, 214–224.
30. Kallianpur, G. and Oodaira, H. (1963). In *Time Series Analysis* (Proc. 1962 Symp.), M. Rosenblatt, ed. Wiley, New York, pp. 279–291.
31. Kuelbs, J., ed. (1978). *Probability on Banach Spaces*. Marcel Dekker, New York. (Five papers in areas of central limit theorem, Gaussian measures, martingales.)
32. Kunita, H. (1976). *Lect. Notes Math.* **511**, 44–77.
33. Kuo, H. -H. (1975). Gaussian Measures in Banach Spaces. *Lect. Notes Math.*, 463. (Basic material on Gaussian measures on Hilbert space, including absolute continuity; also contains results on abstract Wiener spaces for Gaussian measures on Banach space.)
34. Mourier, E. (1953). *Ann. Inst. Henri Poincaré*, **13**, 161–244.
35. Neveu, J. (1968). *Processus aléatoires gaussiens*. University of Montreal Press, Montreal. (Results on Gaussian processes, including absolute continuity. Extensive bibliography.)
36. Orey, S. (1974). *Trans. Amer. Math. Soc.*, **193**, 413–426.
37. Pan Yi-Min (1972). *Select. Transl. Math. Statist. Prob.*, **12**, 109–118; translation of *Shuxue Jinzhan*, **9**, 85–90 (1966).
38. Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York. (Excellent collection of results available

through the mid-1960s in various areas of probability on metric spaces (not including absolute continuity), plus foundations.)

39. Parzen, E. (1963). In *Time Series Analysis* (Proc. Symp. 1962), M. Rosenblatt, ed. Wiley, New York, pp. 155–169.
40. Pisarenko, V. (1961). *Radio Eng. Electron.*, **6**, 51–72.
41. Pisarenko, V. (1965). *Theory Prob. Appl.*, **10**, 299–303.
42. Pitcher, T. S. (1963). *Trans. Amer. Math. Soc.*, **108**, 538–546.
43. Prohorov, Yu. V. (1956). *Theory Prob. Appl.*, **1**, 157–214.
44. Rao, C. R. and Varadarajan, V. S. (1963). *Sankhya A*, **25**, 303–330.
45. Root, W. L. (1963). In *Time Series Analysis* (Proc. 1962 Symp.), M. Rosenblatt, ed. Wiley, New York, 292–346.
46. Rosanov, Yu. V. (1962). *Theory Prob. Appl.*, **7**, 82–87.
47. Rosanov, Yu. V. (1971). Infinite-Dimensional Gaussian Distributions. *Proc. Steklov Inst. Math.*, No. 108. [Monograph containing results on absolute continuity, measurable linear functionals (including a zero-one law for linear manifolds). Extensive bibliography of papers on absolute continuity published prior to 1968.]
48. Sato, H. (1967). *J. Math. Soc. Japan*, **19**, 159–172.
49. Sazanov, V. V. (1958). *Theory Prob. Appl.*, **3**, 201–205.
50. Segal, I. E. (1958). *Trans. Amer. Math. Soc.*, **88**, 12–41.
51. Shepp, L. A. (1966). *Ann. Math. Statist.*, **37**, 321–354.
52. Skorohod, A. V. (1970). *Theory Prob. Appl.*, **15**, 557–580.
53. Skorohod, A. V. (1974). *Integration in Hilbert Space*. Springer-Verlag, New York. (Basic results concerning measures on Hilbert space; many results on absolute continuity.)
54. Varberg, D. E. (1964). *Trans. Amer. Math. Soc.*, **113**, 262–273.
55. Varberg, D. E. (1966). *Notices Amer. Math. Soc.*, **13**, 254.

See also COMMUNICATION THEORY, STATISTICAL; GAUSSIAN PROCESSES; LIKELIHOOD RATIO TESTS; MARTINGALES; MEASURE THEORY IN PROBABILITY AND STATISTICS; RADON–NIKODYM THEOREM; SEPARABLE SPACE; and STOCHASTIC PROCESSES.

CHARLES R. BAKER

ABSOLUTE DEVIATION

The numerical value of the difference between two quantities regardless of its sign. If $\hat{\theta}$ is an estimate of θ, its absolute deviation from θ is $|\theta - \hat{\theta}|$.

See also CHEBYSHEV'S INEQUALITY; MEAN DEVIATION; and MOMENTS.

ABSOLUTE MOMENT

The expected value of the modulus (absolute value) of a random variable X, raised to power r is its rth absolute (crude) moment

$$\nu'_r = E[|X|^r].$$

The quantity

$$\nu_r = E[|X - E[X]|^r]$$

is the rth absolute central moment of X.

ν_1 is the mean deviation*.

ACCELERATED LIFE TESTING

NOTATIONS

Accelerated life models relate the lifetime distribution to the explanatory variable (stress, covariate). This distribution can be defined by the survival function. But the sense of accelerated life models is best seen if they are formulated in terms of the hazard rate function.

Suppose at first that the explanatory variable $x(\cdot) \in E$ is a deterministic time function:

$$x(\cdot) = (x_1(\cdot), \ldots, x_m(\cdot))^T : [0, \infty) \to B \in \mathbf{R}^m,$$

where E is a set of all possible stresses, $x_i(\cdot)$ is univariate explanatory variable. If $x(\cdot)$ is constant in time, $x(\cdot) \equiv x = const$, we shall write x instead of $x(\cdot)$ in all formulas. We note E_1 a set of all constant in time stresses, $E_1 \subset E$.

Denote informally by $T_{x(\cdot)}$ the failure time under $x(\cdot)$ and by

$$S_{x(\cdot)}(t) = \mathbf{P}\{T_{x(\cdot)} \geqslant t\}, \quad t > 0, \quad x(\cdot) \in E,$$

the *survival function* of $T_{x(\cdot)}$. Let $F_{x(\cdot)}(t) = 1 - S_{x(\cdot)}(t)$ be the *cumulative distribution function* of $T_{x(\cdot)}$. The *hazard rate function* of $T_{x(\cdot)}$ under $x(\cdot)$ is

$$\alpha_{x(\cdot)}(t) = \lim_{h\downarrow 0} \frac{1}{h} \mathbf{P}\{T_{x(\cdot)} \in [t, t+h) \mid T_{x(\cdot)} \geqslant t\} = -\frac{S'_{x(\cdot)}(t)}{S_{x(\cdot)}(t)}.$$

Denote by

$$A_{x(\cdot)}(t) = \int_0^t \alpha_{x(\cdot)}(u)du = -ln\{S_{x(\cdot)}(t)\}$$

the *cumulative hazard function* of $T_{x(\cdot)}$.

Each specified accelerated life model relates the hazard rate (or other function) to the explanatory variable in some particular way. To be concise the word stress will be used here for explanatory variable.

We say that a stress $x_2(\cdot)$ is higher than a stress $x_1(\cdot)$ and we write $x_2(\cdot) > x_1(\cdot)$, if for any $t \geqslant 0$ the inequality $S_{x_1(\cdot)}(t) \geqslant S_{x_2(\cdot)}(t)$ holds and exists $t_0 > 0$ such that $S_{x_1(\cdot)}(t_0) > S_{x_2(\cdot)}(t_0)$. We say also that the stress $x_2(\cdot)$ is *accelerated* with respect to the stress $x_1(\cdot)$. It is evident that by the same way one can consider *decelerated* stresses.

At the end we note that if the stress is a stochastic process $X(t)$, $t \geqslant 0$, and $T_{X(\cdot)}$ is the failure time under $X(\cdot)$, then denote by

$$S_{x(\cdot)}(t) = \mathbf{P}\{T_{X(\cdot)} \geqslant t | X(s) = x(s),\ 0 \leqslant s \leqslant t\},$$

the conditional survival function. In this case the definitions of models should be understood in terms of these conditional function.

STRESSES IN ALT

In accelerated life testing (ALT) the most used types of stresses are: constant in time stresses, step-stresses, progressive (monotone) stresses, cyclic stresses and random stresses (see, for example, Duchesne & Lawless (2000, 2002), Duchesne & Rosenthal (2002), Duchesne (2000), Gertsbach & Kordonsky (1997), Miner (1945), Lawless (1982), Nelson (1990), Meeker & Escobar (1998), Nikulin & Solev (2000), Shaked & Singpurwalla (1983)).

The most common case is when the stress is unidimensional, for example, pressure, temperature, voltage, but then more then one accelerated stresses may be used. The monotone stresses are used, for example, to construct the so-called collapcible models, the accelerated degradation models, etc.

The mostly used time-varying stresses in ALT are step-stresses: units are placed on test at an initial low stress and if they do not fail in a predetermined time t_1, the stress is increased. If they do not fail in a predetermined time $t_2 > t_1$, the stress is increased once more, and so on. Thus step-stresses have the form

$$x(u) = \begin{cases} x_1, & 0 \leqslant u < t_1, \\ x_2, & t_1 \leqslant u < t_2, \\ \cdots & \cdots \\ x_m, & t_{m-1} \leqslant u < t_m, \end{cases} \tag{1}$$

where $x_1, \ldots, x_m$ are constant stresses. Sets of step-stresses of the form (1) will be denoted by E_m, $E_m \subset E$. Let E_2, $E_2 \subset E$, be a set of step-stresses of the form

$$x(u) = \begin{cases} x_1, & 0 \leqslant u < t_1, \\ x_2, & u \geqslant t_1, \end{cases} \tag{2}$$

where $x_1, x_2 \in E_1$.

SEDYAKIN'S PRINCIPLE AND MODEL

Accelerated life models could be at first formulated for constant explanatory variables. Nevertheless, before formulating them, let us consider a method for generalizing such models to the case of time-varying stresses.

In 1966 N. Sedyakin formulated his famous *physical principle in reliability* which states that for two identical populations of units functioning under different stresses x_1 and x_2, two moments t_1 and t_2 are equivalent if the probabilities of survival until these moments are equal:

$$\mathbf{P}\{T_{x_1} \geqslant t_1\} = S_{x_1}(t_1) = S_{x_2}(t_2) = \mathbf{P}\{T_{x_2} \geqslant t_2\}, \qquad x_1, x_2 \in E_1.$$

If after these equivalent moments the units of both groups are observed under the same stress x_2, i.e. the first population is observed under the step-stress $x(\cdot) \in E_2$ of the form (2) and the second all time under the constant stress x_2, then for all $s > 0$

$$\alpha_{x(\cdot)}(t_1 + s) = \alpha_{x_2}(t_2 + s). \tag{3}$$

Using this idea of Sedyakin, we considered some generalisation of the model of Sedyakin to the case of any time-varying stresses by supposing that the hazard rate $\alpha_{x(\cdot)}(t)$ at any moment t is a function of the value of the stress at this moment and of the probability of survival until this moment. It is formalized by the following definition.

Definition 1. We say that *Sedyakin's model (SM) holds on a set of stresses E if there exists on $E \times \mathbf{R}^+$ a positive function g such that for all $x(\cdot) \in E$*

$$\alpha_{x(\cdot)}(t) = g\left(x(t), A_{x(\cdot)}(t)\right). \tag{4}$$

The fact that the SM does not give relations between the survival under different constant stresses is a cause of non-applicability of this model for estimation of reliability under the design (usual) stress from accelerated experiments. On the other hand, restrictions of this model when not only the rule (4) but also some relations between survival under different constant stresses are assumed, can be considered. These narrower models can be formulated by using models for constant stresses and the rule (4). For example, it can be shown that the well known and mostly used accelerated failure time model for time-varying stresses is a restriction of the SM when the survival functions under constant stresses differ only in scale.

MODEL OF SEDYAKIN FOR STEP-STRESSES

The mostly used time-varying stresses in accelerated life testing are step-stresses (2) or (1). Let us consider the meaning of the rule (4) for these step-stresses. Namely, we shall show that in the SM the survival function under the step-stress is obtained from the survival functions under constant stresses by the *rule of time-shift*.

Proposition 1. *If the SM holds on E_2 then the survival function and the hazard rate under the stress $x(\cdot) \in E_2$ satisfy the equalities*

$$S_{x(\cdot)}(t) = \begin{cases} S_{x_1}(t), & 0 \leqslant t < t_1, \\ S_{x_2}(t - t_1 + t_1^*), & t \geqslant t_1, \end{cases} \tag{5}$$

and

$$\alpha_{x(\cdot)}(t) = \begin{cases} \alpha_{x_1}(t), & 0 \leqslant t < t_1, \\ \alpha_{x_2}(t - t_1 + t_1^*), & t \geqslant t_1, \end{cases} \tag{6}$$

respectively; the moment t_1^* is determined by the equality $S_{x_1}(t_1) = S_{x_2}(t_1^*)$.

From this proposition it follows that for any $x(\cdot) \in E_2$ and for all $s \geqslant 0$

$$\alpha_{x(\cdot)}(t_1 + s) = \alpha_{x_2}(t_1^* + s). \tag{7}$$

It is the model on E_2, proposed by Sedyakin (1966).

Let us consider a set E_m of stepwise stresses of the form (1). Set $t_0 = 0$. We shall show that the *rule of time-shift* holds for the SM on E_m.

Proposition 2. *If the SM holds on E_m then the survival function $S_{x(\cdot)}(t)$ satisfies the equalities:*

$$S_{x(\cdot)}(t) = S_{x_i}(t - t_{i-1} + t_{i-1}^*), \quad \textit{if} \quad t \in [t_{i-1}, t_i),\ (i = 1, 2, \ldots, m), \tag{8}$$

where t_i^* satisfy the equations

$$S_{x_1}(t_1) = S_{x_2}(t_1^*), \ldots, S_{x_i}(t_i - t_{i-1} + t_{i-1}^*) = S_{x_{i+1}}(t_i^*),\ (i = 1, \ldots, m-1). \tag{9}$$

From this proposition it follows that for all $t \in [t_{j-1}, t_j)$ we have

$$A_{x(\cdot)}(t) = A_{x_j}(t - t_{j-1} + t_{j-1}^*).$$

In the literature on ALT (see Bhattacharyya & Stoejoeti (1989), Nelson (1990), Meeker & Escobar (1998)) the model (8) is also called the *basic cumulative exposure model*.

We note here that the SM can be not appropriate in situations of periodic and quick change of the stress level or when switch-up of the stress from one level to the another can imply failures or shorten the life, see Bagdonavičius & Nikulin (2002).

ACCELERATED FAILURE TIME MODEL

Accelerated life models describing dependence of the lifetime distribution on the stresses will be considered here. The considered models are used in survival analysis and reliability theory analyzing results of accelerated life testing. A number of such models was proposed by engineers who considered physics of failure formation process of certain products or by statisticians (see, for example, Bagdonavičius, Gerville-Reache, Nikulin (2002), Cox & Oakes (1984), Chen (2001), Cooke & Bedford (1995), Crowder, Kimber, Smith & Sweeting (1991), Duchesne & Rosenthal (2002), Duchesne Lawless (2002), Gertsbach & Kordonskiy (1969, 1997), Gerville-Reache & Nikoulina (2002), Kartashov (1979), Lawless (2000), LuValle (2000), Meekers & Escobar (1998), Nelson (1990), Nelson & Meeker (1991), Singpurwalla & Wilson (1999), Viertl (1988), Xu & Harrington (2001),...) To introduce in this topic we consider at first the most famous Accelerated Failure Time (AFT) model. We start from the definition of this model for constant stresses.

Suppose that under different constant stresses the survival functions differ only in scale:

$$S_x(t) = S_0\{r(x)\,t\} \quad \text{for any} \quad x \in E_1, \qquad (10)$$

where the survival function S_0, called the *baseline survival function*, does not depend on x. For any fixed x the value $r(x)$ can be interpreted as the *time-scale change constant* or the acceleration (deceleration) constant of survival functions.

Applicability of this model in accelerated life testing was first noted by Pieruschka (1961), see also Sedyakin (1966). It is the most simple and the most used model in failure time regression data analysis and ALT (see, Cooke & Bedford (1995), Nelson (1990), Meeker & Escobar (1998), Chen (2001), Hu & Harrington (2001), etc. . . .)

Under the AFT model on E_1 the distribution of the random variable

$$R = r(x)T_x$$

does not depend on $x \in E_1$ and its survival function is S_0. Denote by m and σ^2 the mean and the variance of R, respectively. In this notations we have

$$\mathbf{E}(T_x) = m/r(x), \quad \mathbf{Var}(T_x) = \sigma^2/r^2(x),$$

and hence the coefficient of variation

$$\frac{\mathbf{E}(T_x)}{\sqrt{\mathbf{Var}(T_x)}} = \frac{m}{\sigma}$$

does not depend on x.

The survival functions under any $x_1, x_2 \in E_1$ are related in the following way:

$$S_{x_2}(t) = S_{x_1}\{\rho(x_1, x_2)\,t\},$$

where the function $\rho(x_1, x_2) = r(x_2)/r(x_1)$ shows the degree of scale variation. It is evident that $\rho(x, x) = 1$.

Now we consider the definition of the AFT model for time-varying stresses, using the Sedyakin's prolongation of the model (10) on E_1 to another model on E.

Definition 2. *The AFT model holds on E if there exists on E a positive function r and on* $[0, \infty)$ *a survival function* S_0 *such that*

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t r\{x(u)\}\,du\right) \quad \text{for any} \quad x(\cdot) \in E. \qquad (11)$$

The most used way of application of AFT model is the following. The baseline survival function S_0 is taken from some class of parametric distributions, such as Weibull, lognormal, loglogistic, or from the class of regression models such as Cox proportional hazards model, frailty model, linear transformation model, additive hazards model, etc. . . .

The AFT model in this form (11) was studied in Bagdonavičius (1990), Bagdonavičius & Nikulin (2002), Cox & Oakes (1984), Chen (2001), Meeker & Escobar (1998), Nelson (1990), Sedyakin (1966), Viertl (1988), Xu & Harrington (2001), etc.

AFT MODEL FOR STEP-STRESSES

The next two proposals give the forms of the survival functions in AFT model on E_m.

Proposition 3. *If the AFT model holds on E_2 then the survival function under any stress $x(\cdot) \in E_2$ of the form (2) verifies the equality*

$$S_{x(\cdot)}(t) = \begin{cases} S_{x_1}(t), & 0 \leqslant \tau < t_1, \\ S_{x_2}(t - t_1 + t_1^*), & \tau \geqslant t_1, \end{cases} \tag{12}$$

where

$$t_1^* = \frac{r(x_1)}{r(x_2)} t_1. \tag{13}$$

Proposition 4. *If the AFT model holds on E_m then the survival function $S_{x(\cdot)}(t)$ verifies the equalities*:

$$\begin{aligned} S_{x(\cdot)}(t) &= S_0 \left\{ \sum_{j=1}^{i-1} r(x_j)(t_j - t_{j-1}) \right. \\ &\qquad \left. + r(x_i)(t - t_{i-1}) \right\} \\ &= S_{x_i} \left\{ t - t_{i-1} + \frac{1}{r(x_i)} \right. \\ &\qquad \left. \times \sum_{j=1}^{i-1} r(x_j)(t_j - t_{j-1}) \right\}, \\ &\qquad t \in [t_{i-1}, t_i), \quad (i = 1, 2, \ldots, m). \end{aligned}$$

So the AFT model in the form (11) verifies the Sedyakin's principle on E_m. The proofs of these propositions one can find in Bagdonavičius and Nikulin (1995, 2002).

Remark 1. Suppose that $z_0(\cdot)$ is a fixed (for example, *usual* or *standard*) stress and $S_0 = S_{z_0(\cdot)}$, $z_0(\cdot) \in E$. In this case the AFT model is given by (11). Denote

$$\begin{aligned} f_{x(\cdot)}(t) &= S_0^{-1}(S_{x(\cdot)}(t)) \\ &= \int_0^t r\{x(u)\} du, \quad x(\cdot) \in E, \\ &\qquad \text{with} \quad f_{x(\cdot)}(0) = 0. \end{aligned} \tag{14}$$

This relation shows that the moment t under any stress $x(\cdot) \in E$ is equivalent to the moment $f_{x(\cdot)}(t)$ under the usual stress $z_0(\cdot)$, therefore $f_{x(\cdot)}(t)$ is called *the resource used till the moment t under the stress $x(\cdot)$*.

PARAMETRIZATION OF THE AFT MODEL

Let $x(\cdot) = (x_0(\cdot), x_1(\cdot), \ldots, x_m(\cdot))^T \in E$ be a possibly time-varying and multidimensional explanatory variable; here $x_0(t) \equiv 1$ and $x_1(\cdot), \ldots, x_m(\cdot)$ are univariate explanatory variables.

Under the AFT model the survival function under $x(\cdot)$ is given by (11). Often the function r is parametrized in the following form:

$$r(x) = e^{-\beta^T x}, \tag{15}$$

where $\beta = (\beta_0, \ldots, \beta_m)^T$ is a vector of unknown regression parameters. In this case the parametrized AFT model is given by the next formula:

$$S_{x(\cdot)}(t) = S_0 \left(\int_0^t e^{-\beta^T x(\tau)} d\tau \right), \quad x(\cdot) \in E. \tag{16}$$

Here $x_j(\cdot)$ $(j = 1, \ldots, m)$ are not necessary the observed explanatory variables. They may be some specified functions $z_j(x)$. Nevertheless, we use the same notation x_j for $z_j(x)$.

If the explanatory variables are constant over time then the model (16), or (10), is written as

$$S_x(t) = S_0 \left(e^{-\beta^T x} t \right), \quad x \in E, \quad x_j \in E_1, \tag{17}$$

and the logarithm of the failure time T_x under x may be written as

$$\ln\{T_x\} = \beta^T x + \varepsilon, \tag{18}$$

where the survival function of the random variable ε is $S(t) = S_0(\ln t)$. It does not depend on x. Note that if the failure-time distribution is lognormal, then the distribution of the random variable ε is normal, and we have the *standard multiple linear regression model*.

Let us consider, following Nelson (1990), Meeker & Escobar (1998), Viertl (1988), some examples. For this we suppose at first that the explanatory variables are interval-valued (load, temperature, stress, voltage, pressure).

Suppose at first that x is one-dimensional, $x \in E_1$.

Example 1. Let

$$r(x) = e^{-\beta_0 - \beta_1 x}. \tag{19}$$

It is the *log-linear model*.

Example 2. Let

$$r(x) = e^{-\beta_0 - \beta_1 \log x} = \alpha_1 x^{\beta_1}. \tag{20}$$

It is the *power rule model*.

Example 3. Let

$$r(x) = e^{-\beta_0 - \beta_1 / x} = \alpha_1 e^{-\beta_1 / x}. \tag{21}$$

It is the *Arrhenius model*.

Example 4. Let

$$r(x) = e^{-\beta_0 - \beta_1 \ln \frac{x}{1-x}} = \alpha_1 \left(\frac{x}{1-x} \right)^{-\beta_1}, \quad 0 < x < 1. \tag{22}$$

It is the *Meeker-Luvalle* model (1995).

The Arrhenius model is used to model product life when the explanatory variable is the temperature, the power rule model—when the explanatory variable is voltage, mechanical loading, the log-linear model is applied in endurance and fatigue data analysis, testing various electronic components (see Nelson (1990)). The model of Meeker-Luvalle is used when x is the proportion of humidity.

The model (16) can be generalized. One can suppose that $\ln r(x)$ is a linear combination of some specified functions of the explanatory variable:

$$r(x) = \exp\left\{ -\beta_0 - \sum_{i=1}^{k} \beta_i z_i(x) \right\}, \tag{23}$$

where $z_i(x)$ are specified functions of the explanatory variable, $\beta_0, \ldots, \beta_k$ are unknown (possibly not all of them) parameters.

Example 5. Let

$$r(x) = e^{-\beta_0 - \beta_1 \log x - \beta_2 / x} = \alpha_1 x e^{-\beta_2 / x}, \tag{24}$$

where $\beta_1 = -1$. It is the *Eyring model*, applied when the explanatory variable x is the temperature.

INTERPRETATION OF THE REGRESSION COEFFICIENTS IN AFT MODEL

Suppose that the stresses are constant over time. Then under the AFT model (17) the p-quantile of the failure time T_x is

$$t_p(x) = e^{\beta^T x} S_0^{-1}(1-p), \quad x \in E_1, \tag{25}$$

so the logarithm

$$\ln\{t_p(x)\} = \beta^T x + c_p, \quad x \in E_1, \tag{26}$$

is a linear function of the regression parameters; here $c_p = \ln(S_0^{-1}(1-p))$.

Let $m(x) = \mathbf{E}\{T_x\}$ be the *mean life* of units under x. Then

$$m(x) = e^{\beta^T x} \int_0^\infty S_0(u) du, \quad x \in E_1, \tag{27}$$

and the logarithm

$$\ln\{m(x)\} = \beta^T x + c, \quad x \in E_1, \tag{28}$$

is also a linear function of the regression parameters; here

$$c = \ln\left\{ \int_0^\infty S_0(u) du \right\}.$$

Denote by

$$MR(x,y) = \frac{m(y)}{m(x)} \quad \text{and} \quad QR(x,y) = \frac{t_p(y)}{t_p(x)}, \quad x,y \in E_1, \tag{29}$$

the ratio of means and quantiles, respectively. For the AFT model on E_1 we have

$$MR(x,y) = QR(x,y) = e^{\beta^T (y-x)}, \tag{30}$$

and hence $e^{\beta^T (y-x)}$ *is the ratio of means, corresponding to the stresses x and y.*

TIME-DEPENDENT REGRESSION COEFFICIENTS

The AFT model usually is parametrized in the following form (16). In this case the resource used till the moment t under stress $x(\cdot)$ is given by (14), from which it follows that at any moment t the resource usage rate

$$\frac{\partial}{\partial t} f_{x(\cdot)}(t) = e^{-\beta^T x(t)}, \quad x(\cdot) \in E,$$

depends only on the value of the stress $x(\cdot)$ at the moment t. More flexible models can be obtained by supposing that the coefficients β are time-dependent, i.e. taking

$$\frac{\partial}{\partial t} f_{x(\cdot)}(t) = e^{-\beta^T(t)x(t)} = e^{-\sum_{i=0}^m \beta_i(t)x_i(t)},$$

$$x(\cdot) \in E,$$

If the function $\beta_i(\cdot)$ is increasing or decreasing in time then the effect of ith component of the stress is increasing or decreasing in time.

So we have the model

$$S_{x(\cdot)}(t) = S_0\left\{\int_0^t e^{-\beta^T(u)x(u)}du\right\}. \tag{31}$$

It is the AFT model with time-dependent regression coefficients. We shall consider the coefficients $\beta_i(t)$ in the form

$$\beta_i(t) = \beta_i + \gamma_i g_i(t), \quad (i = 1, 2, \ldots, m),$$

where $g_i(t)$ are some specified deterministic functions or realizations of predictable processes. In such a case the AFT model with time-dependent coefficients and constant or time-dependent stresses can be written in the usual form (16) with different interpretation of the stresses. Indeed, set

$$\begin{aligned}\theta &= (\theta_0, \theta_1, \ldots, \theta_{2m})^T\\ &= (\beta_0, \beta_1, \ldots, \beta_m, \gamma_1, \ldots, \gamma_m)^T,\\ z(\cdot) &= (z_0(\cdot), z_1(\cdot), \ldots, z_{2m}(\cdot))^T\\ &= (1, x_1(\cdot), \ldots, x_m(\cdot), x_1(\cdot)g_1(\cdot), \ldots,\\ &\qquad x_m(\cdot)g_m(\cdot))^T.\end{aligned} \tag{32}$$

Then

$$\begin{aligned}\beta^T(u)x(u) &= \beta_0 + \sum_{i=1}^m (\beta_i + \gamma_i g_i(t))x_i(t)\\ &= \theta^T z(u).\end{aligned}$$

So the AFT model with the time-dependent regression coefficients can be written in the standard form

$$S_{x(\cdot)} = S_0\left\{\int_0^t e^{-\theta^T z(u)}du\right\}, \tag{33}$$

where the unknown parameters and the explanatory variables are defined by (32).

PLANS OF EXPERIMENTS IN ALT

As it was said before the purpose of ALT is to give estimators of the main reliability characteristics: *the reliability function $S_0 = S_{x^{(0)}}$, the p-quantile $t_p(x^{(0)})$ and the mean value $m(x_0)$ under usual (design) stress $x^{(0)}$*, using data of accelerated experiments when units are tested at higher than usual stress conditions. Different plans of experiments are used in ALT with dynamic environment.

The first plan of experiments.

Denote by $x_0 = (x_{00}, x_{01}, \ldots, x_{0m})$, $x_{00} = 1$, the usual stress. Generally accelerated life testing experiments are done under an one-dimensional stress ($m = 1$), sometimes under two-dimensional ($m = 2$).

Let $x_1, \ldots, x_k$ be constant over time *accelerated stresses*:

$$x_0 < x_1 < \cdots < x_k;$$

here $x_i = (x_{i0}, x_{i1}, \ldots, x_{im}) \in E_m$, $x_{i0} = 1$. The usual stress x_0 *is not used* during experiments. According to the first plan of experiment *k groups of units are tested. The ith group of n_i units, $\sum_{i=1}^k n_i = n$, is tested under the stress x_i. The data can be complete or independently right censored.*

If the form of the function r is completely unknown and this plan of experiments is used, the function S_{x_0} can not be estimated even if it is supposed to know a parametric family to which belongs the distribution $S_{x_0}(t)$.

For example, if $S_0(t) = e^{-(t/\theta)^\alpha}$ then for constant stresses

$$S_x(t) = \exp\left\{-\left(\frac{r(x)}{\theta}t\right)^\alpha\right\}, \quad x \in E_1. \tag{34}$$

Under the given plan of experiments the parameters

$$\alpha,\ \frac{r(x_1)}{\theta}, \ldots, \frac{r(x_k)}{\theta}, \quad x_i \in E_1, \tag{35}$$

and the functions $S_{x_1}(t), \ldots, S_{x_k}(t)$ may be estimated. Nevertheless, the function $r(x)$ being completely unknown, the parameter $r(x_0)$ can not be written as a known function of these estimated parameters. So $r(x_0)$ and, consequently, $S_{x_0}(t)$ can not be estimated.

Thus, the function r must be chosen from some class of functions. Usually the model (17) is used.

The second plan of experiments. In step-stress accelerated life testing the second plan of experiments is as follows:

n units are placed on test at an initial low stress and if it does not fail in a predetermined time t_1, the stress is increased and so on. Thus, all units are tested under the step-stress $x(\cdot)$ of the form:

$$x(\tau) = \begin{cases} x_1, & 0 \leqslant \tau < t_1, \\ x_2, & t_1 \leqslant \tau < t_2, \\ \cdots & \cdots \\ x_k, & t_{k-1} \leqslant \tau < t_k; \end{cases} \tag{36}$$

where $x_j = (x_{j0}, \ldots, x_{jm})^T \in E_m$, $x_{j0} = 1$, $t_0 = 0, t_k = \infty$.

In this case the function $r(x)$ should be also parametrized because, even when the usual stress is used until the moment t_1, the data of failures occurring after this moment do not give any information about the reliability under the usual stress when the function $r(x)$ is unknown. Thus, the model (16) should be used. From the Proposition 2 it follows that for step-stresses the form (36) we have: if $t \in [t_{i-1}, t_i), i = 1, \ldots, k$

$$S_{x(\cdot)}(t) = S_0\left\{\mathbf{1}_{\{i>1\}}\sum_{j=1}^{i-1} e^{-\beta^T x_j}(t_j - t_{j-1}) + e^{-\beta^T x_i}(t - t_{i-1})\right\}. \tag{37}$$

Now we consider *the third plan of experiments*. Application of the first two plans may not give satisfactory results because assumptions on the form of the function $r(x)$ are done. These assumptions can not be statistically verified because of lack of experiments under the usual stress.

If the function $r(x)$ is completely unknown, and the coefficient of variation (defined as the ratio of the standard deviation and the mean) of failure times is not too large, the following plan of experiments may be used.

The third plan of experiments. Suppose that the failure time under the usual stress x_0 takes large values and most of the failures occur after the moment t_2 given for the experiment. According to this plan two groups of units are tested:

a) the first group of n_1 units under a constant accelerated stress x_1;

b) the second group of n_2 units under a step-stress: time t_1 under x_1, and after this moment under the usual stress x_0 until the moment t_2, $x_1, x_2 \in E_1$, i.e. under the stress $x_2(\cdot)$ from E_2:

$$x_2(\tau) = \begin{cases} x_1, & 0 \leqslant \tau \leqslant t_1, \\ x_0, & t_1 < \tau \leqslant t_2. \end{cases} \tag{38}$$

Units use much of their resources until the moment t_1 under the accelerated stress x_1, so after the switch-up failures occur in the interval $[t_1, t_2]$ even under usual stress. The AFT model implies that

$$S_{x_1}(u) = S_{x_0}(ru),$$

where $r = r(x_1)/r(x_0)$, and

$$S_{x_2(\cdot)}(u) = \begin{cases} S_{x_0}(ru), & 0 \leqslant u \leqslant t_1, \\ S_{x_0}(rt_1 + u - t_1), & t_1 < u \leqslant t_2, \end{cases}$$

or, shortly,

$$S_{x_2(\cdot)}(t) = S_{x_0}(r(u \wedge t_1) + (u - t_1) \vee 0), \quad x_2(\cdot) \in E_2, \tag{39}$$

with $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

It will be shown in the next section that if the third plan is used, and both functions S_{x_0} and $r(x)$ are completely unknown, semiparametric estimation of S_{x_0} is possible.

The third plan may be modified. The moment t_1 may be chosen as random. The most natural is to choose t_1 as the moment when the failures begin to occur.

At the end we consider the *fourth plan of experiment*, which is applied when the failure-time distribution under the usual stress is *exponential*. According to this plan k *groups of units are observed. The i-th group of* n_i *units is tested under one-dimensional constant stress* $x^{(i)}$ *until the* r_i*-th failure* ($r_i \leqslant n$), (type two censoring). *The failure moments of the i-th group are*

$$T_{(i1)} \leqslant T_{(i2)} \leqslant \cdots \leqslant T_{(ir_i)}.$$

DATA

We suppose that n units are observed. The ith unit is tested under the value $x^{(i)}(\cdot) = (x_1^{(i)}(\cdot), \ldots, x_m^{(i)}(\cdot))^T$ of a possibly time-varying and multi-dimensional explanatory variable $x(\cdot)$, according to the plan of experiment. The data are supposed to be independently right censored.

Let T_i and C_i be the failure and censoring times of the ith unit,

$$X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \leqslant C_i\}}.$$

As it is well known the right censored data may be presented in the form

$$(X_1, \delta_1), \ldots, (X_n, \delta_n). \tag{40}$$

or

$$(N_1(t), Y_1(t), t \geqslant 0), \ldots, (N_n(t), Y_n(t), t \geqslant 0), \tag{41}$$

where

$$N_i(t) = \mathbf{1}_{\{X_i \leqslant t, \delta_i = 1\}}, \quad Y_i(t) = \mathbf{1}_{\{X_i \geqslant t\}}. \tag{42}$$

Here $N_i(t)$ is the number of observed failures of the ith unit in the interval $[0, t]$, and $Y_i(t)$ is the indicator of being at risk just prior to the moment t.

Using this presentation of data the statistical analysis of an appropriate accelerated life model can be done as is shown, for example, in Andersen, Borgan, Gill & Keiding (1993), Bagdonavičius and Nikulin (2002), Lawless (1982), Meeker & Escobar (1998).

ESTIMATION AND TESTING IN ALT

If the functions $r(\cdot)$ and $S_0(\cdot)$ are unknown we have a nonparametric model. The function $r(\cdot)$ can be parametrized. If the baseline function S_0 is completely unknown, in this case the we have a semiparametric model. Very often the baseline survival function S_0 is also taken from some class of parametric distributions, such as Weibull, lognormal, loglogistique, etc. In this case we have a parametric model and the maximum likelihood estimators of the parameters are obtained by almost standard way for any plans. Parametric case was studied by many people, see, for example, Bagdonavičius, Gerville-Réache, Nikoulina and Nikulin (2000), Gertsbakh & Kordonskiy (1969), Gerville-Réache & Nikoulina (2000), Glaser (1984), Hirose (1993), Iuculano & Zanini (1986), Lin & Ying (1995), LuValle (2000), Mazzuchi & Soyer (1992), Nelson (1990), Meeker & Escobar (1998), Sedyakin (1966), Sethuraman & Singpurwalla (1982), Schmoyer (1986), Shaked & Singpurwalla (1983), Viertl (1988). Nonparametric analysis of AFT model was considered by Basu & Ebrahimi (1982), Lin & Ying (1995), Lawless (1982), Robins & Tsiatis (1992), Schmoyer (1991), Ying (1993), Bagdonavičius and Nikulin (2000). Semiparametric case was considered by Tsiatis (1991), Lin & Ying (1995), Duchesne & Lawless (2002), Bagdonavičius and Nikulin (2002).

Tsiatis (1991), (constant stresses), Robins and Tsiatis (1992), Lin and Ying (1995) (time-dependent stresses) give asymptotic properties of the regression parameters for random right censored data. Lin and Ying (1995) give also semiparametric procedures for making inference about β (but not about the survival function and other reliability characteristics) which by pass the estimation of the covariance matrix of $\hat{\beta}$. All above mentioned papers boundedness of the density of the censoring variable is required. In the case of accelerated life testing when type one censoring is generally used, this condition is not true. In the case of accelerated life testing when type one censoring is generally used, this condition does not hold. Bagdonavičius and Nikulin (2002) give asymptotic properties of the estimators under the third plan of experiments. Using these properties Lin & Ying (1995) and Bagdonavičius & Nikulin (2002) studied tests for nullity of the regression coefficients, namely for testing the hypothesis

$$H_{k_1,k_2,\ldots,k_l} : \beta_{k_1} = \cdots = \beta_{k_l} = 0, \qquad (1 \leqslant k_1 < k_2 < \cdots < k_l). \tag{43}$$

MODELING IN ALT IN TERMS OF RESOURCE USAGE

The accelerated life (time transformation) models with dynamic environment give the possibility to include the effect of the usage history on the lifetime distributions of various units, subjects, items, populations, etc. Accelerated life models as time transformation models can be formulated using the notion of the resource introduced by Bagdonavičius and Nikulin (1995). This notion gives a general approach for construction time transformation models in terms of the rate of resource usage and it gives a simple physical interpretation of considered models.

Let Ω be a population of units and suppose that the failure-time of units under stress $x(\cdot)$ is a random variable $T_{x(\cdot)} = T_{x(\cdot)}(\omega), \omega \in \Omega$, with the survival function $S_{x(\cdot)}(t)$ and the cumulative distribution function $F_{x(\cdot)}(t)$. The moment of failure of a concrete item $\omega_0 \in \Omega$ is given by a nonnegative number $T_{x(\cdot)}(\omega_0)$.

The proportion $F_{x(\cdot)}(t)$ of units from Ω which fail until the moment t under the stress $x(\cdot)$ is also called the *uniform resource of population used until the moment t*. The same population of units Ω, observed under different stresses $x_1(\cdot)$ and $x_2(\cdot)$ use different resources until the same moment t if $F_{x_1(\cdot)}(t) \neq F_{x_2(\cdot)}(t)$. In sense of equality of used resource the moments t_1 and t_2 are equivalent if $F_{x_1(\cdot)}(t_1) = F_{x_2(\cdot)}(t_2)$.

The random variable

$$R^U = F_{x(\cdot)}(T_{x(\cdot)}) = 1 - S_{x(\cdot)}(T_{x(\cdot)})$$

is called the *uniform resource*. The distribution of the random variable R^U does not depend on $x(\cdot)$ and is uniform on $[0, 1)$. The uniform resource of any *concrete* item $\omega_0 \in \Omega$ is $R^U(\omega_0)$. It shows the proportion of the population Ω which fails until the moment of the unit's ω_0 failure $T_{x(\cdot)}(\omega_0)$.

The considered definition of the resource is not unique. Take any continuous survival function G such that the inverse $H = G^{-1}$ exists. In this case the distribution of the statistics $R^G = H(S_{x(\cdot)}(T_{x(\cdot)}))$ doesn't depend on $x(\cdot)$ and the survival function of R^G is G. The random variable R^G is called the *G-resource* and the number

$$f^G_{x(\cdot)}(t) = H(S_{x(\cdot)}(t)),$$

is called the *G-resource used until the moment t*. Note that in the case of the uniform resource $H(p) = 1 - p, p \in (0, 1]$.

Accelerated life models can be formulated specifying the way of resource usage, i.e. in terms of the rate of resource usage.

Note often the definitions of accelerated life models are formulated in terms of *exponential resource usage*, when $G(t) = e^{-t}, t \geqslant 0$, because *the exponential resource usage rate is nothing else but the hazard rate and the used resource is the cumulative hazard rate* respectively.

Let $\alpha_{x(\cdot)}(t)$ and $A_{x(\cdot)}(t)$ be the hazard rate and the cumulative hazard rate under $x(\cdot)$. The exponential resource is obtained by taking $G(t) = e^{-t}$, $t \geqslant 0$, and $H(p) = G^{-1}(p) = -\ln p$, so it is the random variable

$$R = A_{x(\cdot)}(T_{x(\cdot)})$$

with standard exponential distribution. For any t the number $A_{x(\cdot)}(t) \in [0, \infty)$ is the exponential resource used until the moment t under stress $x(\cdot)$. The rate of exponential resource usage is the hazard rate $\alpha_{x(\cdot)}(t)$.

All models can be formulated in terms of other resources than exponential. Let us consider at first one particular resource. Suppose that x_0 is a fixed (for example, usual) stress and $G = S_{x_0}$. For any $x(\cdot) \in E \supset E_1$ set

$$f_{x(\cdot)}(t) = S^{-1}_{x_0}(S_{x(\cdot)}(t)).$$

Then the moment t under any stress $x(\cdot) \in E$ is equivalent to the moment $f_{x(\cdot)}(t)$ under the usual stress x_0. The survival function of the resource R is S_{x_0}. As it was shown by (14) in the Remark 1 the AFT model (11) is determined by the S_{x_0}-resource usage rate by the next differential equation:

$$\frac{\partial}{\partial t} f_{x(\cdot)}(t) = r\{x(t)\}, \quad \text{for any} \quad x(\cdot) \in E,$$
$$\text{with} \quad f_{x(\cdot)}(0) = 0.$$

An natural generalization of the AFT model on E is obtained by changing, for example, the right part of this equation by the next way:

$$\frac{\partial f_{x(\cdot)}(t)}{\partial t} = r\{x(t)\}\, t^{\nu(x(t))-1} \quad \text{for any} \quad x(\cdot) \in E,$$
$$\text{with} \quad f_{x(\cdot)}(0) = 0, \qquad (44)$$

where ν is a positive function on E. This equality implies that

$$S_{x(\cdot)}(t) = S_{x_0}\left(\int_0^t r\{x(\tau)\}\tau^{\nu(x(\tau))-1} d\tau\right)$$
$$\text{for any} \quad x(\cdot) \in E. \qquad (45)$$

In this model variation of stress changes locally not only the scale but also the shape of distribution. This model is known as *the changing shape and scale model* (CHSS), see Bagdonavičius & Nikulin (2000).

To show the usefulness of the notion of the resource let us consider, following Bagdonavičius and Nikulin (1997), the so-called *generalized additive-multiplicative* (GAM) model on E, given in terms of the rate of resource usage by the next differential equation

$$\frac{\partial f^G_{x(\cdot)}(t)}{\partial t} = r[x(t)]\frac{\partial f^G_0(t)}{\partial t} + a(x(t)),$$
$$\text{with} \quad f^G_0(0) = f^G_{x(\cdot)}(0) = 0, \quad (46)$$

for some functions a and r (positive) on E, where $f^G_0(t) = H(S_0(t))$. In this model the stress influences the rate of resource using as multiplicatively as additively. The last equation implies that

$$S_{x(\cdot)}(t) = G\left\{\int_0^t r[x(\tau)]dH(S_0(\tau)) + \int_0^t a(x(\tau))d\tau\right\}. \qquad (47)$$

Consider some particular cases.

1. Taking $G(t) = e^{-t}\mathbf{1}_{\{t\geqslant 0\}}$ we obtain:

$$\frac{\partial f^G_{x(\cdot)}(t)}{\partial t} = -\frac{S'_{x(\cdot)}(t)}{S_{x(\cdot)}(t)} = \alpha_{x(\cdot)}(t),$$
$$\frac{\partial f^G_0(t)}{\partial t} = -\frac{S'_0(t)}{S_0(t)} = \alpha_0(t),$$

where $\alpha_{x(\cdot)}(t)$ is the hazard rate under the covariate $x(\cdot)$, $\alpha_0(t)$ is the baseline hazard rate. So we obtain the model:

$$\alpha_{x(\cdot)}(t) = r[x(t)]\alpha_0(t) + a(x(t)).$$

It is the *additive-multiplicative semiparametric model* (Lin & Ying (1996). If $a(x(t)) \equiv 0$, we obtain the *proportional hazards model* or *Cox model*:

$$\alpha_{x(\cdot)}(t) = r[x(t)]\alpha_0(t).$$

If $r[x(t)] \equiv 1$, we obtain the *additive hazards model*:

$$\alpha_{x(\cdot)}(t) = \alpha_0(t) + a(x(t)).$$

2. Taking $G(t) = \exp\{-\exp\{t\}\}, \quad t \in R^1$, we obtain

$$\frac{\partial f^G_{x(\cdot)}(t)}{\partial t} = \frac{\alpha_{x(\cdot)}(t)}{A_{x(\cdot)}(t)}, \quad \frac{\partial f^G_0(t)}{\partial t} = \frac{\alpha_0(t)}{A_0(t)},$$

where

$$A_{x(\cdot)}(t) = \int_0^t \alpha_{x(\cdot)}(\tau)d\tau, \quad A_0(t) = \int_0^t \alpha_0(\tau)d\tau$$

are the cumulated hazards rates. So we have the new model:

$$\frac{\alpha_{x(\cdot)}(t)}{A_{x(\cdot)}(t)} = r[x(t)]\frac{\alpha_0(t)}{A_0(t)} + a(x(t)).$$

3. Taking $G(t) = 1/(1+t)$, $t \geqslant 0$, we obtain

$$\frac{\partial f^G_{x(\cdot)}(t)}{\partial t} = \frac{\alpha_{x(\cdot)}(t)}{S_{x(\cdot)}(t)}, \quad \frac{\partial f^G_0(t)}{\partial t} = \frac{\alpha_0(t)}{S_0(t)}.$$

So we have the *generalized logistic regression* model:

$$\frac{\alpha_{x(\cdot)}(t)}{S_{x(\cdot)}(t)} = r[x(t)]\frac{\alpha_0(t)}{S_0(t)} + a(x(t)).$$

If $a(x(t)) \equiv 0$, we obtain the *logistic regression model* since we have:

$$\frac{1}{S_{x(\cdot)}(t)} - 1 = r[x(t)] \left(\frac{1}{S_0(t)} - 1 \right).$$

4. Taking $G(t) = 1/(1 + e^t)$, $t \in R^1$, we obtain

$$\frac{\partial f^G_{x(\cdot)}(t)}{\partial t} = \frac{\alpha_{x(\cdot)}(t)}{1 - S_{x(\cdot)}(t)}, \quad \frac{\partial f^G_0(t)}{\partial t} = \frac{\alpha_0(t)}{1 - S_0(t)}.$$

So we have the new model:

$$\frac{\alpha_{x(\cdot)}(t)}{1 - S_{x(\cdot)}(t)} = r[x(t)] \frac{\alpha_0(t)}{1 - S_0(t)} + a(x(t)).$$

5. Take $G(t) = \Phi(\ln t)$, $t \geqslant 0$, where $\Phi(\cdot)$ is the distribution function of the standard normal law. If $a(x(t)) \equiv 0$, then in terms of survival functions we obtain the model:

$$\Phi^{-1}(1 - S_{x(\cdot)}(t)) = \ln r[x(t)] + \Phi^{-1}(1 - S_0(t)).$$

It is the *generalized probit model*.

6. Taking $G = S_0$, we obtain

$$S_{x(\cdot)}(t) = S_0 \left\{ \int_0^t \sigma[x(\tau)] d\tau \right\}.$$

where $\sigma(x(t)) = r[x(t)] + a(x(t))$. It is the *accelerated life model*, given by (11). As one can see, the GAM model contains many interesting sub-models, which are well adapted to solve the statistical problems in ALT, and the notions of the resource and the rate of resource usage give a power instrument for modeling in ALT.

REFERENCES

1. Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
2. Bagdonavičius, V. (1990). Accelerated life models when the stress is not constant. *Kybernetika*, **26**, 289–295.
3. Bagdonavičius, V., Nikulin, M. (1995). *Semiparametric models in accelerated life testing*. Queen's Papers in Pure and Applied Mathematics, **98**. Kingston: Queen's University, Canada.
4. Bagdonavičius, V., Nikulin, M. (1997). Analysis of general semiparametric models with random covariates. *Revue Roumaine de Mathématiques Pures et Appliquées*, **42**, #5–6, 351–369.
5. Bagdonavičius, V., Nikulin, M. (1998). *Additive and multiplicative semiparametric models in accelerated life testing and survival analysis*. Queen's Papers in Pure and Applied Mathematics, **108**. Kingston: Queen's University, Canada.
6. Bagdonavičius, V., Gerville-Réache, L., Nikoulina, V., Nikulin, M. (2000). Expériences accélérées: analyse statistique du modèle standard de vie accélérée. *Revue de Statistique Appliquée*, XLVIII, 3, 5–38.
7. Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. New York: Methuen (Chapman and Hall).
8. Chen, Y. Q. (2001). Accelerated Hazards Regression Model and Its Adequacy for Censored Survival Data. *Biometrics*, **57**, 853–860.
9. Cooke R., Bedford, T. (1995). Analysis of Reliability Data Using Subsurvival Functions and Censoring Mosels. In: *Recent Advances in Life-Testing and Reliability*, (Ed. Balakrishnan, N.), Boca Raton: CRC Press, 12–41.
10. Duchesne, Th., Rosenthal, J. F. (2002). Stochastic Justification of Some Simple Reliability Models. Preprint of the Department of Statistics, University of Toronto, Toronto, ON, Canada, 18p.
11. Duchesne, Th., Lawless, J. (2002). Semiparametric Inference Methods for General Time Scale Models. *Lifetime Data Analysis*, **8**, 263–276.
12. Duchesne, T. (2000). Methods Common to Reliability and Survival Analysis. In: *Recent Advances in Reliability Theory. Methodology, Practice and Inference*, (Eds. N. Limnios and M. Nikulin), Boston: Birkhauser, 279–290.
13. Duchesne, T. and Lawless, J. (2000). Alternative Time Scale and Failure Time Models. *Lifetime Data Analysis*, **6**, 157–179.
14. Gertsbakh, L. B. and Kordonskiy, K. B. (1969). *Models of Failure*. New York: Springer-Verlag.
15. Gertsbakh, L. B. and Kordonskiy, K. B. (1997). Multiple time scale and the lifetime coefficient of variation: Engineering applications. *Lifetime Data Analysis*, **2**, 139–156.
16. Gerville-Réache, L. and Nikoulina, V. (1999). Analysis of Reliability Characteristics Estimators in Accelerated Life Testing, In: *Statistical and Probabilistic Models in Reliability*,

(Eds. D. Ionescu and N. Limnios), Boston: Birkhauser, 91–100.

17. Glaser, R. E. (1984). Estimation for a Weibull Accelerated Life Testing Model, *Naval Research Logistics Quarterly*, **31**, 4, 559–570.
18. Hirose, H. (1993). Estimation of Thresshold Stress in Accelerated Life-Testing, *IEEE Transaction on reliability*, **42**, 650–657.
19. Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons.
20. Kartashov, G. D. (1979). Methods of Forced (Augmented) Experiments (in Russian). Moscow: *Znaniye Press*.
21. Klein, J. P. and Basu, A. P. (1981). Weibull Accelerated Life Tests When There are Competing Causes of Failure, *Communications in Statistical Methods and Theory*, **10**, 2073–2100.
22. Klein, J. P. and Basu, A. P. (1982). Accelerated Life Testing under Competing Exponential Failure Distributions, *IAPQR Trans.*, **7**, 1–20.
23. Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.
24. Lawless, J. F. (2000). Dynamic Analysis of Failures in Repairable Systems and Software. In: *Recent Advances in Reliability Theory*. (Eds. Limnios, N., Nikulin, M), Boston: Birkhauser, 341–354.
25. Lin, D. Y. and Ying, Z. (1995). Semiparametric inference for accelerated life model with time dependent covariates. *Journal of Statist. Planning and Inference*, **44**, 47–63.
26. Lin, D. Y. and Ying, Z. (1996). Semiparametric analysis of the general additive-multiplicative hazards model for counting processes. *Ann. Statistics*, **23**, #5, 1712–1734.
27. LuValle, M. (2000). A Theoretical Framework for Accelerated Testing. In: *Recent Advances in Reliability Theory. Methodology, Practice and Inference*, (Eds. N. Limnios and M. Nikulin), Boston: Birkhauser, 419–434.
28. Miner, M. A. (1945). Cumulative Damage in Fatigue, *J. of Applied Mechanics*, **12**, A159–A164.
29. Meeker, W. Q., Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley and Sons.
30. Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis.* New York: John Wiley and Sons.
31. Nelson, W. and Meeker, W. (1991). Accelerated Testing: Statistical models, test plans, and data analysis. *Technometrics*, **33**, 236–238.
32. Nikulin, M. S., and Solev, V. N. (2002). Testing Problem for Increasing Function in a Model with Infinite Dimensional Nuisance Parameter. In: *Goodness-of-fit Tests and Validity of Models*, (Eds. C. Huber, N. Balakrishnan, M. Nikulin and M. Mesbah), Boston: Birkhauser.
33. Sethuraman, J. and Singpurwalla, N. D. (1982). Testing of hypotheses for distributions in accelerated life tests. *JASA*, **77**, 204–208.
34. Singpurwalla, N. (1995). Survival in Dynamic Environments. *Statistical Sciences*, **1**, #10, 86–103.
35. Viertl, R. (1988). *Statistical Methods in Accelerated Life Testing*. Göttingen: Vandenhoeck and Ruprecht.
36. Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.*, **18**, 354–72.
37. Robins, J. M., and Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time dependent covariates. *Biometrika*, **79**, #2, 311–319.

V. Bagdonavicius
M. Nikulin

ACCEPTABLE QUALITY LEVEL (AQL)

This is usually defined as the maximum percent defective (or the maximum number of defects per 100 units) that can be considered satisfactory for a process average.

BIBLIOGRAPHY

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd ed. Wiley, New York.

Johnson, N. L. and Leone, F. (1977). *Statistics and Experimental Design in Engineering and the Physical Sciences*, 2nd ed., Vol. 1. Wiley, New York, Chap. 10.

Juran, J. M., ed. (1964). *Quality-Control Handbook*, 2nd ed. McGraw-Hill, New York (first ed., 1953).

See also Acceptance Process Zone and Quality Control, Statistical.

ACCEPTANCE ERROR

A term used in the theory of testing hypotheses* to denote a decision to accept a hypothesis H_0 when that hypothesis is not valid. It is also called a Type I error*.

See also HYPOTHESIS TESTING and LEVEL OF SIGNIFICANCE.

ACCEPTANCE NUMBER

Given a sampling plan*, the acceptance number c denotes the maximum number of defective items that can be found in the sample without leading to rejection of the lot.

BIBLIOGRAPHY

Duncan, A. J. (1974). *Quality Control and Industrial Statistics*, 4th ed. Richard D. Irwin, Homewood, Ill.

Standards Committee of ASQC* (1971). ASQC Standard A2 (rev. 1978).

See also ACCEPTABLE QUALITY LEVEL (AQL); QUALITY CONTROL, STATISTICAL; and SAMPLING PLANS.

ACCEPTANCE PROCESS ZONE

The acceptance process level (APL) is a fundamental notion in quality control. It is the process level most remote from the standard that still yields product quality that we are willing to accept with high probability. Since most specifications are two-sided (i.e., requiring characteristics to lie within a specified tolerance band), it is usually appropriate to specify both an upper and a lower APL.

The band around the nominal value between the upper and lower acceptance process level (APL) values is called the *acceptance process zone*.

FURTHER READING

Duncan, A. J. (1986). *Quality Control and Industrial Statistics* (4th ed.). Irwin, Homewood, Ill.

Montgomery, D. C. (2001). *Introduction to Statistical Quality Control* (3rd ed.). Wiley, New York/Chichester.

See also QUALITY CONTROL, STATISTICAL; REJECTABLE PROCESS LEVEL (RPL); and SAMPLING PLANS.

ACCEPTANCE REGION (IN TESTING HYPOTHESES)

A hypothesis test* divides the space $\mathcal{T}$ of a test statistic* T into two complementary regions, C (the critical region) and $\mathcal{T} - C$. The region $\mathcal{T} - C$ is called the *acceptance region or nonrejection region*. This region is characterized by the property that if the *value* of the test statistic falls into this region the hypothesis under test is accepted.

See also CRITICAL REGION and HYPOTHESIS TESTING.

ACCEPTANCE SAMPLING

The term "acceptance sampling" relates to the acceptance or rejection of a product or process on the basis of sampling inspection*. It has been pointed out that "sampling inspection is the process of evaluating the quality of material by inspecting some but not all of it" [4]. Its methods constitute decision rules for the disposition or sentencing of the product sampled. In this sense it may be contrasted with survey sampling*, the purpose of which is largely estimation*.

Sampling plans*, which specify sample size and acceptance criteria, are fundamental to acceptance sampling. Such plans may be based on a simple dichotomous classification of conformance or nonconformance of a quality characteristic to specified criteria (*attributes plans*) or on a comparison of statistics computed from quantitative measurements to numerical criteria developed from the specifications and from assumptions about the shape and nature of the distribution of individual measurements (*variables plans*). An example of the former is the attributes plan: sample 50 items and accept the lot of material from which the sample was taken if two or fewer items are found nonconforming; reject otherwise. An example of the latter is the variables plan: sample 12 items and accept the lot if the sample mean is more than 2 standard deviations above the lower specification limit; reject otherwise.

When a process parameter such as the mean* or standard deviation* is specified, the sampling plan resolves itself simply into a test of hypothesis. *See* HYPOTHESIS TESTING. That is, the sampling plan might be a t-test* if the plan is imposed to assure that the process mean conforms to a specified value. These tests are called variables plans for process parameter and are commonly used in the sampling of bulk materials. *See* BULK SAMPLING. More complicated situations arise when it is necessary to test the proportion of product beyond or between specification limits through the use of a measurement criterion. Such tests are referred to as variables plans for proportion nonconforming.

The operating characteristic (OC) curve (complement of the power* curve) is the primary measure of performance of a sampling plan. It shows the probability of acceptance as a function of the value of the quality characteristic. As such, it provides a description of the protection afforded by the plan as well as a vehicle for comparison of various acceptance sampling plans and procedures. Two types of OC curves are distinguished. The type A operating characteristic curve relates to the inspection of individual lots of product and shows lot probability of acceptance as a function of lot quality. In attributes inspection, it is computed from the hypergeometric distribution*. The type B operating characteristic curve relates to the process that produced the product to be inspected and shows the proportion of lots accepted in a continuing series as a function of the process average. In attributes inspection, it is computed either from the binomial distribution* when inspection is for proportion nonconforming, or from the Poisson distribution* when inspection is for nonconformities per unit. Details of the nature and construction of type A and type B operating characteristic curves are given in ref. 2.

Often the performance of a plan is characterized by two points on the OC curve: a producer's quality level with high probability of acceptance and a consumer's quality level with low probability of acceptance. The corresponding risks are called the *producer's risk* (of rejection) and the *consumer's risk* (of acceptance). (The producer's risk is conventionally taken to be 0.05 and the consumer's risk is 0.10.) The ratio of the consumer's quality level to the producer's quality level is called the *operating* (or *discrimination*) *ratio*. It describes the steepness of the OC curve and hence the capability of the plan to distinguish between acceptable and unacceptable quality. So-called two-point plans can be derived from the producer's and consumer's quality levels through their operating ratios. A third point on the OC curve that is commonly referenced is the indifference quality level*, at which the probability of acceptance is 0.50. Sets of plans have been developed using the indifference quality level and the relative slope of the OC curve at that point.

Acceptance sampling procedures progress far beyond the simple single sampling plan to other, more complex procedures. Double-sampling* plans allow the possibility of two samples, a second sample being taken if the results of the first sample are not sufficiently definitive. This concept can be extended to multiple sampling plans, involving more than two samples. Sequential procedures* are applied in acceptance sampling to achieve the excellent discrimination and economy of sample size associated with such methods. These plans may be used in both attributes and variables inspection.

Various measures and associated curves have been developed to describe the properties of sampling plans. The *average sample number (ASN)* curve* describes the average sample size for various quality levels when using double*, multiple*, sequential*, or other procedures. The *average outgoing quality (AOQ)* curve* shows the average proportion of nonconforming product the consumer will receive if rejected lots are 100% inspected plotted against quality level. Its maximum is called the *average outgoing quality limit (AOQL)**. Under such a procedure, the average total inspection curve shows the total number of units inspected in both sampling and 100% inspection and can be used to estimate and compare inspection loads.

For a continuing sequence of lots sampling plans may be combined into sampling schemes consisting of two or more plans used together with switching rules that establish the procedure for moving from one of the plans to another. Schemes can be constructed to give protection for both the producer and the consumer which is superior to that of the constituent plans with a reduction of sample size for the protection afforded. Such schemes are usually specified by an *acceptable quality level (AQL)** which, when exceeded, will eventually lead to a switch to a tighter plan with consequent economic and psychological pressure on the supplier to improve the quality of the product. Sampling schemes have their own OC curves. Sometimes options for discontinuation of inspection are incorporated into the procedure. Sampling schemes may be combined into sampling systems that select specific schemes by prescribed rules. The most important sampling systems are military standards* MIL-STD-105D for attributes and MIL-STD-414 for variables. Their civilian counterparts are ANSI Z1.4 and ANSI Z1.9 in the United States and international standards ISO 2859 and ISO 3951, respectively.

The variety of approaches in acceptance sampling is almost limitless. Continuous sampling plans are used on streams of output where lots are difficult to define. Chain sampling* plans link the criteria for the immediate sampling plan to past results. Grand lot sampling procedures combine samples from lots that have been shown to be homogeneous, to achieve larger sample size and greater discrimination. Skip-lot plans* provide for the inspection of only a fraction of the lots submitted. Acceptance control charts* can be used to visually portray the results of inspection through the medium of the control chart. Bayesian plans introduce economic considerations and prior results and estimates into the sampling equation. Special plans have been developed for various areas of application, such as compliance testing, reliability and life testing*, and safety inspection. *See* SAMPLING PLANS.

The application and administration of sampling inspection demands a broad range of knowledge of statistical methodology, because the determination of what, where, when, how, and to what quality levels the inspection is to be carried out is largely empirical. Acceptance sampling procedures are an integral part of quality control* practice and serve to distinguish between quality levels as a vehicle for quality improvement. In application, however, they should, where possible, be supplemented and eventually supplanted by statistical techniques for process quality control (such as control charts*) in an effort to *prevent* the occurrence of nonconforming material rather than to *detect* it after it has been produced.

The history and development of acceptance sampling is described in detail by Dodge [1], who originated and developed many acceptance sampling procedures. The statistical methodology of acceptance sampling has been treated specifically by Schilling [5] and in the context of industrial statistics as a whole by Duncan [3].

REFERENCES

1. Dodge, H. F. (1969). *J. Quality Tech.*, **1**: Part I, Apr., 77–88; Part II, July, 155–162; Part III, Oct., 225–232; **2**: Part IV, Jan., 1970, 1–8.
2. Dodge, H. F. and Romig, H. G. (1959). *Sampling Inspection Tables—Single and Double Sampling*, 2nd ed. Wiley, New York.
3. Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 4th ed. Richard D. Irwin, Homewood, Ill.
4. Freeman, H. A., Friedman, M., Mosteller, F., and Wallis, W. A., eds. (1948). *Sampling Inspection Principles: Procedures and Tables for Single, Double and Sequential Sampling in Acceptance Inspection and Quality Control*. McGraw-Hill, New York.
5. Schilling, E. G. (1981). *Acceptance Sampling in Quality Control*. Marcel Dekker, New York.

See also AVERAGE OUTGOING QUALITY LIMIT (AOQL); CONTROL CHARTS; LOT TOLERANCE TABLES, DODGE–ROMIG; QUALITY CONTROL, STATISTICAL; and SAMPLING PLANS.

EDWARD G. SCHILLING

ACCURACY AND PRECISION

The *accuracy* of an observation or a statistic derived from a number of observations has to do with how close the value of the statistic is to a supposed “true value”.

In forecasting, accuracy is a measure of how close the forecast $\hat{Y}_t$ of an observation Y_t at time t is to Y_t; *see* PREDICTION AND FORECASTING. *See* also REGRESSION VARIABLES, SELECTION OF and FINAL PREDICTION ERROR CRITERIA, GENERALIZED for model choices in multiple regression aimed at reduction of error (and hence at improved accuracy).

In estimation theory accuracy measures how close an estimate $\hat{\theta}$ of a parameter θ is to the "true value" of θ. The accuracy of $\hat{\theta}$ can be measured, for example, in terms of the mean absolute error or of the mean squared error* (MSE) of $\hat{\theta}$.

Accuracy should be distinguished from *precision*. Precision of measurement indicates the resolving power of a measuring device and is frequently given by the number of decimal places reported in the measurements made with the device. The precision of an estimator $\hat{\theta}$, on the other hand, measures how tightly the distribution of $\hat{\theta}$ clusters about its center (say, its expected value) [1, Sec. 4.1]. One has

$$\mathrm{MSE}(\hat{\theta}) = \{\text{Variance of } \hat{\theta}\} + (\text{bias of } \hat{\theta})^2;$$

Here the accuracy of $\hat{\theta}$ can be measured via $\mathrm{MSE}(\hat{\theta})$ and its precision via $\mathrm{Var}(\hat{\theta})$. Carl Friedrich Gauss's measure of precision is

$$1/\{\sqrt{2} \times (\text{standard deviation of } \hat{\theta})\}$$

where the quantity $\sqrt{2}\times$ (standard deviation) is known as the *modulus*. Gauss's measure satisfies the intuitive notion that the precision increases as the standard deviation decreases.

NEYMAN AND WOLFOWITZ ACCURACY

Let T be a statistic* based on a sample from a population having an unknown parameter θ, and let $(L_1(T),L_2(T))$ be a confidence interval for θ. If

$$Q(\theta_0) = \Pr\,[L_1(T) \leqslant \theta_0 \leqslant L_2(T)|\theta],$$

then [2,4] $Q(\theta_0)$ is the *Neyman accuracy* of the confidence interval. It is a measure of the accuracy of $(L_1(T),\ L_2(T))$ in excluding the false value $\theta_0 \neq \theta$ of θ. The interval with the smaller Neyman accuracy is said to be *more selective* [5]; *see* CONFIDENCE INTERVALS AND REGIONS. If

$$W(a,b) = aE\{(L_1(T) - \theta)^2\} + bE\{(L_2(T) - \theta)^2\},$$

then [2] $W(\cdot,\cdot)$ is the *Wolfowitz accuracy* of the confidence interval [6], and measures how close the confidence limits L_1 and L_2 are to the true value of θ; see also [3], where the efficiency of competing confidence intervals is measured inter alia by the ratio of their Wolfowitz accuracies when $a = b = 1$.

REFERENCES

1. Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco.
2. Ghosh, B. K. (1975). A two-stage procedure for the Behrens-Fisher problem, *J. Amer. Statist. Ass.*, **70**, 457–462.
3. Harter, H. L. (1964). Criteria for best substitute interval estimators, with an application to the normal distribution, *J. Amer. Statist. Ass.*, **59**, 1133–1140.
4. Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philos. Trans. Royal. Soc. A*, **236**, 333–380.
5. Stuart, A. and Ord, J. K. (1991). *Kendall's Advanced Theory of Statistics*, Vol. 2 (5th ed.). Oxford University Press, Oxford, U.K. (Secs. 20.14, 20.15).
6. Wolfowitz, J. (1950). Minimax estimates of the mean of a normal distribution with known variance. *Ann. Math. Statist.*, **21**, 218–230.

See also FINAL PREDICTION ERROR CRITERIA, GENERALIZED; MEAN DEVIATION; MEAN SQUARED ERROR; PREDICTION AND FORECASTING; and VARIANCE.

ACHENWALL, GOTTFRIED

Born: October 20, 1719, in Elbing, Germany.

Died: May 1, 1772, in Göttingen, Germany.

Contributed to: *Staatswissenschaft* ("university statistics").

Achenwall was born into the family of a merchant. In 1738–1740 he acquired a knowledge of philosophy, mathematics, physics, and history at Jena; then he moved to Halle, where, without abandoning history, he studied the law and *Staatswissenschaft* (the science of the state; also known as "university statistics"). Apparently in 1742 Achenwall returned for a short time to Jena, then continued his education in Leipzig. In 1746 he became Docent at Marburg, and in 1748, extraordinary professor at Göttingen (ordinary professor of law and of philosophy from 1753), creating there the Göttingen school of statistics. Its most eminent member was A. L. Schlözer (1735–1809). Achenwall married in 1752, but his wife died in 1754, and he had no children.

Achenwall followed up the work of Hermann Conring (1606–1681), the founder of *Staatswissenschaft*, and was the first to present systematically, and in German rather than in Latin, the Conring tradition. According to both Conring and Achenwall, the aim of statistics was to describe the climate, geographical position, political structure, and economics of a given state, to provide an estimate of its population, and to give information about its history; but discovering relations between quantitative variables was out of the question. For Achenwall [1, p. 1], "the so-called statistics" was the *Staatswissenschaft* of a given country.

Since 1741, "statisticians" have begun to describe states in a tabular form, which facilitate the use of numbers, a practice opposed by Achenwall. Even in 1806 and 1811 [5, p. 670] the use of tabular statistics was condemned because numbers were unable to describe the spirit of a nation.

Nevertheless, Achenwall [4, Chap. 12] referred to Süssmilch,* advised state measures fostering the multiplication of the population, recommended censuses, and even [4, p. 187] noted that its "probable estimate" can be gotten by means of "yearly lists of deaths, births, and marriages." The gulf between statistics (in the modern sense) and *Staatswissenschaft* was not as wide as it is usually supposed to have been. Leibniz's manuscripts, written in the 1680s, present a related case. First published in 1866 and reprinted in 1977, they testify that he was both a political arithmetician and an early advocate of tabular description (both with and without the use of numbers) of a given state; see [10,222–227,255].

REFERENCES

1. Achenwall, G. (1748). *Vorbereitung zur Staatswissenschaft*. Göttingen. (An abridged version of this was included in his next contribution.)
2. Achenwall, G. (1749). *Abriβ der neuesten Staatswissenschaft der vornehmsten europäischen Reiche und Republicken zum Gebrauch in seinen academischen Vorlesungen*. Schmidt, Göttingen.
3. Achenwall, G. (1752). *Staatsverfassung der europäischen Reiche im Grundrisse*. Schmidt, Göttingen. This is the second edition of the *Abriβ*. Later editions: 1756, 1762, 1767, 1768, 1781–1785, 1790–1798. By 1768 the title had changed to *Staatsverfassung der heutigen vornehmsten europäischen Reiche and Völker*, and the publisher was Witwe Wanderhoeck.
4. Achenwall, G. (1763). *Staatsklugheit nach ihren Grundsätzen*. Göttingen. (Fourth ed., 1779).
5. John, V. (1883). The term "Statistics." *J. R. Statist. Soc.* **46**, 656–679. (Originally published in German, also in 1883.)
6. John, V. (1884). *Geschichte der Statistik*. Encke, Stuttgart.
7. Lazarsfeld, P. (1961). Notes on the history of quantification in sociology—trends, sources and problems, *Isis*, **52**, 277–333. Reprinted (1977) in *Studies in the History of Statistics and Probability*, Sir Maurice Kendall and R. L. Plackett. eds. Vol. 2, pp. 213–269. Griffin, London and High Wycombe
8. Leibniz, G. W. (1886). *Sämmtliche Schriften und Briefe*. Reihe 4, Bd. 3. Deutsche Akad. Wiss., Berlin.
9. Schiefer, P. (1916). *Achenwall und seine Schule*. München: Schrödl. (A Dissertation.)
10. Sheynin, O. B. (1977). Early history of the theory of probability. *Arch. Hist. Ex. Sci.*, **17**, 201–259.
11. Solf, H. H. (1938). *G. Achenwall. Sein Leben und sein Werk, ein Beitrag zur Göttinger Gelehrtengeschichte*. Mauser, Forchheim. Oberfranken. (A dissertation.)
12. Westergaard, H. (1932). *Contributions to the History of Statistics*. King, London. (Reprinted, New York, 1968, and The Hague, 1969.)
13. Zahn, F. and Meier, F. (1953). Achenwall, *Neue deutsche Biogr.* 1, 32–33. Duncker and Humblot, Berlin.

O. Sheynin

ACTUARIAL HEALTH STUDIES

Medico-actuarial studies originated in the United States in the 1890s from concerted efforts to improve the underwriting of life insurance risks [15]. The mortality investigations undertaken were aimed to isolate and measure the effects of selected risk factors, such as occupational hazards, medical conditions, and build. The underlying hypothesis was that each of the factors (or certain combinations of factors) influencing mortality could be regarded as an independent variable and the total mortality risk could be treated as a linear compound of a number of independent elements.

The first comprehensive study, known as the *Specialized Mortality Investigation*, was carried out by the Actuarial Society of America and published in 1903 [1]. It covered the experience of 34 life insurance companies over a 30-year period and focused attention on the mortality in 35 selected occupations, 32 common medical conditions, and several other factors affecting mortality. It was followed in 1912 by the *Medico-Actuarial Mortality Investigation* [2], sponsored jointly by the Actuarial Society of America and the Association of Life Insurance Company Medical Directors, which included a much wider variety of occupations, medical conditions, and other factors, among them abuse of alcohol. This study laid the broad lines on which such investigations have been conducted since.

The assessment of the long-term risk in life insurance was seen as requiring analysis of mortality by sex, age at issue, and duration since issue of insurance for policies issued under similar underwriting rules. Cohorts of policyholders were followed over long periods of time. Attention was focused on the mortality in the years following issue of insurance in order to trace the effects on mortality of the selection exercised by insurance companies through medical examinations and other screening for insurance, as well as effects of antiselection by applicants for insurance who withheld information relevant to their health. The extent of class selection, that is, the reflection of the underlying mortality in the segments of the population from which the insured lives were drawn, was brought out in the mortality experienced among the insured lives under study after many years since issue of insurance had elapsed. Most important, however, the patterns of the mortality experienced over longer periods of time indicated the incidence of the extra mortality by duration, which permitted classifying the long-term risk as one of decreasing extra mortality, relatively level extra mortality, or increasing extra mortality.

Analyses of mortality by cause shed light on the causes mainly responsible for excess mortality and also on the causes of death which could be controlled to some degree by screening applicants for life insurance. Successive medico-actuarial investigations permitted some conclusions regarding the trends in mortality associated with different factors affecting mortality, notably occupational hazards, build, blood pressure, various medical conditions, and changing circumstances.

METHODOLOGY

In selecting a particular cohort of policyholders for study, because of interest in some particular factor influencing mortality, it was the practice in medico-actuarial investigations to exclude individuals who also presented other kinds of risks. Specifically, all individuals were excluded from the study if they were also subject to any other kind of risk that would have precluded issuing insurance at standard premium rates. Consequently, such extra mortality as was found in the study could properly be associated with the factor of interest rather than with the combined effects of this factor and other elements at risk.

The findings of medico-actuarial investigations have been customarily expressed as ratios of actual deaths in the cohort of policyholders under study to the expected deaths, which were calculated on the basis of contemporaneous death rates among otherwise similar life insurance risks accepted at standard premium rates. Such mortality ratios usually were computed by sex, age groups at issue of the insurance, duration since issue of the insurance, and causes of death. The calculation of expected deaths involves accurate estimates of the exposed to risk. Because of the varying forms of records, individual and grouped, different tabulating procedures have been employed, as a rule considering deaths within a unit age interval as having occurred at the end of that interval.

Ratios of actual to expected mortality provide very sensitive measures of mortality and therefore may fluctuate widely in finer subdivisions of the experience level. They have the merit, however, of revealing even small departures from expected mortality in broad groupings. In some circumstances the patterns of excess mortality are more clearly perceived from the extra deaths per 1000 than from corresponding mortality ratios; this often is the case during the period immediately following surgery for cancer. It is important to keep in mind that mortality ratios generally decrease with age, so that the mortality ratios for all ages combined can be materially affected by the age composition of a population.

Proportions surviving a specified period of time, even though they provide absolute measures of longevity in a population, have rarely been used in medico-actuarial investigations, because they are relatively insensitive to appreciable changes in mortality; small differences in proportions surviving may be difficult to assess. Relative proportions surviving have been used occasionally in medico-actuarial studies where very high mortality rates occur, as among cancer patients.

In all mortality comparisons, but particularly in comparisons of mortality ratios and relative proportions surviving, it is the suitability of the basis for calculating expected deaths that makes the figures meaningful. If such a basis is regarded as a fixed yardstick, then the reliability of comparisons based on small numbers of deaths can be tested by determining whether an observed deviation from this basis is or is not significant in probability terms; if it is significant, then what are the limits in probability terms within which the “true” value of the observed deviation can be expected to lie [14]?

In medico-actuarial mortality investigations the numbers of deaths in most classifications have usually been quite large. The mortality ratios shown for such classifications have therefore been taken as reasonably reliable estimates of the “true” values of the mortality ratios in the underlying population. As a rule of thumb, when the number of policies terminated by death was 35 or greater and some doubt attached to the significance of the mortality ratio, confidence limits were calculated at the 95% confidence level on the assumption of a normal distribution; when the number of policies terminated by death was less than 35, confidence limits have been calculated on the assumption of a Poisson distribution.

INTERPRETING THE FINDINGS

Medico-actuarial investigations have been based on the experience among men and women insured under ordinary life insurance policies. These insured lives have been drawn predominantly from the middle-class and better-off segments of the population and have passed the screening for life insurance which results in the rejection of about 2% of all applicants and the charging of extra premiums to about 6% of all applicants. Initially at least more than 9 out of 10 persons are accepted for life insurance at standard premium rates and those issued insurance at standard premium rates are in ostensibly good health. In recent years the death rates of men aged 25 or older insured under standard ordinary policies have ranged from 25 to 35% of the corresponding population death rates in the first two policy years, from 40 to 50% of the corresponding population death rates at policy year durations 3–5, and from 55 to 75% of the corresponding population death rates after 15 or more years have elapsed since issue of insurance. The corresponding figures for women insured under standard ordinary policies have been similar to those of male insured lives at ages over 50, but were closer to population death rates at the younger ages.

Inasmuch as the underwriting rules determine which applicants are accepted for standard insurance and which for insurance at extra premium rates, the mortality experience in medico-actuarial investigations has occasionally been affected to a significant degree by changes in underwriting practices to more lenient or stricter criteria.

The mortality findings in medico-actuarial investigations relate to the status of individual at time of issue of the insurance. The experience therefore reflects not only the effects of some individuals becoming poorer risks with the passage of time, but also of some individuals becoming better risks (e.g., leaving employment in a hazardous occupation or benefiting from medical or surgical treatment) and withdrawing from the experience. Where the effects of employment in hazardous occupations or of certain physical impairments are deferred, it is essential that the study cover a sufficiently long period of time for the deferred mortality to become manifest. This is particularly important in the case of overweight and hypertension [13].

The results of medico-actuarial investigations have been relatively free from bias arising from failure to trace the experience among those withdrawing. Considerable evidence has been accumulated to show that insured lives who cease paying premiums and thus automatically remove themselves from observation are as a group subject to somewhat lower mortality [12].

It should also be kept in mind that the mortality ratios shown in medico-actuarial studies were computed on the basis of the number of policies (or amounts of insurance) and not on the basis of lives. In classifications involving small numbers of policies terminated by death it has been necessary to look into the data to determine whether the results had been affected by the death of a single individual with several policies (or with large amounts of insurance). This has usually been noted in the descriptive text.

The data in medico-actuarial investigations are very accurate with respect to reported ages and remarkably complete in the follow-up*. The information obtained on applications for life insurance with respect to past medical histories requires some qualification. The great majority of applicants for life insurance admit some physical impairment or medical history; if the impairment or history appears to be significant, attention is focused on it in the medical examination for insurance and a statement from the attending physician may be obtained. The medical examination on modest amounts of life insurance is not as comprehensive as a diagnostic examination in clinical practice, where the physician is in position to study a patient for a longer period of time, more intensively, and with the patient's full cooperation. Applicants for insurance not infrequently forget or try to conceal unfavorable aspects of their personal or family medical histories, particularly with respect to questionable habits. Even when reasonably complete details are elicited, there are usually practical limits on the extent to which it is feasible to check up on indefinite statements and vague diagnoses reported on applications for life insurance. Only on applications for large amounts of insurance would two or more medical examinations by different physicians be called for and intensive effort made to clarify obscure findings. Broadly speaking, the medical findings on life insurance medical examinations stand up very well, but the medical impairments studied in medico-actuarial mortality investigations often represent less differentiated conditions which cannot be characterized as precisely as is sometimes possible in clinical studies [13].

On the other hand, it has proved feasible on applications for life insurance to obtain fuller details of occupation and avocation (even approximate exposure to occupational hazards) than has been possible in many epidemiological* studies.

FINDINGS OF MEDICO-ACTUARIAL INVESTIGATIONS

The *Medico-Actuarial Mortality Investigation* of 1912 covered the period from 1885 to 1909 [2]. It produced tables of average weights for men and women by age and height which remained in general use as a weight standard until 1960. The mortality experienced according to variations in build indicated some extra mortality among underweights at ages under 35, associated with materially greater risk of tuberculosis and pneumonia, but at ages 35 and older the lowest mortality was found among those 5 to 10 pounds underweight. Overweight was found to be associated with increased death rates from heart disease, diabetes, and cerebral hemorrhage. The investigation also included 76 groups of medical impairments, 68 occupations, four categories of women studied according to marital status, and insured blacks and North American Indians.

The *Occupational Study 1926* dealt with some 200 occupations or groups of occupations, separately for those where occupational accidents were the dominant element of extra risk and those where nonoccupational accidents, pneumonia, cirrhosis of the liver, cancer, or other causes, were suspect as responsible for the extra risk [3].

The *Medical Impairment Study 1929* [4], which covered the period from 1909 to 1928, and its 1931 Supplement[11] broadly confirmed the findings of the Medico-Actuarial Investigation as to average weights by age and height at ages 25 and older and the effects on mortality of departures from average weight. The study centered on 122 groups of medical impairments, including a number of combinations of two impairments treated as a single element of risk. The more important findings related to the extra mortality on heart murmurs, elevated blood pressure, and albumin and sugar in the urine. The findings on elevated blood pressure indicated clearly that systolic blood pressures in excess of 140 mm were associated with significant extra mortality, which at the time was contrary to medical opinion [10].

Smaller investigations of the mortality among insured lives with various medical impairments followed, published under the titles *Impairment Study 1936* [6] and *Impairment Study 1938* [19]. Together they included 42 groups of medical impairments, among them persons with a history of cancer, gastric and duodenal ulcers, gall bladder disease, and kidney stone, including surgery for these conditions.

In 1938 an extensive investigation was also undertaken of the mortality according to variations in systolic and diastolic pressure. This study, known as the *Blood Pressure Study 1938*, covered the period from 1925 to 1938 [8,9]. It confirmed earlier findings that diastolic pressures in excess of 90 mm as well as systolic blood pressures in excess of 140 mm were associated with at least 25% extra mortality, and it brought out clearly that various minor impairments accompanying elevated blood pressure, notably overweight, increased the risk appreciably.

The *Occupational Study 1937* covered numerous occupations over the period 1925 to 1936 [7]. It developed the extra mortality among those employed in the manufacturing, distribution, and serving of alcoholic beverages. It also indicated some decline since the early 1920s in the accidental death rates in many occupations.

The *Impairment Study 1951*, which covered the period 1935 to 1950, reviewed the mortality experience for 132 medical impairment classifications on policies issued during the years 1935 through 1949 [16]. It showed lower extra mortality than that found in the Medical Impairment Study 1929 for medical impairments due to infections, for conditions treated surgically, for diseases of the respiratory system, and for some women's diseases. Because of the inclusion in the study of smaller groups of lives with specific impairments, greater use was made of confidence intervals based on the Poisson distribution.

The *Build and Blood Pressure Study 1959* covered the experience on about 4,500,000 policies over the period 1935 to 1953 [17]. It focused on changes in the mortality experienced among underweights and overweights, elevated blood pressures, and combinations of overweight and hypertension with other impairments. New tables of average weights for men and women by age and height were developed, which showed that men had gained weight while women had reduced their average weights since the 1920s. Moderate underweights showed very favorable mortality, while marked overweights recorded somewhat higher relative mortality. The mortality on slight, moderate, and marked elevations in blood pressure registered distinctly higher mortality than found in earlier investigations.

The *Occupational Study 1967* covered the period from 1954 to 1964 and was limited to occupations believed to involve some extra mortality risks [18]. Only the following occupations, on which there was substantial experience, recorded a crude death rate in excess of 1.5 per 1000:

Lumberjacks
Mining operators
Explosive workers
Construction crane workers
Shipbuilding operators
Structural iron workers
Railroad trainmen and switchmen
Taxi drivers
Marine officers and crew
Guards and watchmen
Marshals and detectives
Sanitation workers
Porters
Elevator operators
Persons selling, delivering, or
serving alcoholic beverage

The mortality in most occupations decreased from that reported in the *Occupational Study 1937*. Significant reductions occurred among mining officials and foremen, workers in metal industry, telecommunication linemen, longshoremen, firefighters, police officers, window cleaners, hotelkeepers, saloonkeepers and bartenders, and most laborers. Relative mortality increased for lumberjacks, railroad trainmen and switchmen, truck drivers, marine crew and guards, and watchmen.

The *Build Study 1979* [21] and the *Blood Pressure Study 1979* [22] each covered about 4,250,000 policies over the period 1954 to 1971. They showed that the average weights for men had continued to increase, as did the average weights for women under 30; women 30 and older registered decreases in average weights as compared with the *Build and Blood Pressure Study 1959*. The excess mortality among overweights was found to be substantially the same as in the earlier study, but somewhat higher mortality was recorded among moderate overweights. Nevertheless, the optimum weights (those associated with the lowest mortality) were again found to be in the range of 5 to 10% below average weight, even though the average weights for men had increased significantly. The excess mortality on elevated blood pressures was found to be distinctly lower than in the *Build and Blood Pressure Study 1959*. A cohort of 24,000 men who had been treated for hypertension exhibited virtually normal mortality among those whose blood pressures had been reduced to below 140 systolic and 90 diastolic after treatment. The study adduced the most convincing evidence thus far available that recent treatment for hypertension was highly effective for many years. In progress at this time is another medico-actuarial investigation of the mortality among insured lives with a wide variety of medical impairments, covering the period from 1955 through 1974.

REFERENCES

References 1, 3 to 10, and 16 to 21 are original reports.

1. Actuarial Society of America (1903). *Specialized Mortality Investigation*. New York.
2. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1912–1914). *Medico-Actuarial Mortality Investigation*, 5 vols. New York. (Original reports; some of these cover basic design of studies.)
3. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1926). *Occupational Study 1926*. New York.
4. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1929). *Medical Impairment Study 1929*. New York.
5. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1932). *Supplement to Medical Impairment Study 1929*. New York.
6. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1936). *Impairment Study 1936*. New York.
7. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1937). *Occupational Study 1937*. New York.
8. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1939). *Blood Pressure Study 1939*. New York.
9. Actuarial Society of America and Association of Life Insurance Medical Directors of America (1940). *Supplement to Blood Pressure Study 1939*. New York.
10. Association of Life Insurance Medical Directors of America and the Actuarial Society of America (1925). *Blood Pressure Study*. New York.
11. Batten, R. W. (1978). *Mortality Table Construction*. Prentice-Hall, Englewood Cliffs, N.J.
12. Benjamin, B. and Haycocks, H. W. (1970). *The Analysis of Mortality and Other Actuarial Statistics*. Cambridge University Press, London.
13. Lew, E. A. (1954). *Amer. J. Publ. Health*, **44**, 641–654. (Practical considerations in interpretation of studies.)
14. Lew, E. A. (1976). In *Medical Risks*. R. B. Singer and L. Levinson, eds. Lexington Books, Lexington, Mass., Chap. 3. (Limitations of studies.)
15. Lew, E. A. (1977). *J. Inst. Actuaries*, **104**, 221–226. (A history of medico-actuarial studies.)
16. Society of Actuaries (1954). *Impairment Study 1951*. New York. (A highly readable original report.)
17. Society of Actuaries (1960). *Build and Blood Pressure Study 1959*. Chicago. (A highly readable original report.)
18. Society of Actuaries (1967). *Occupational Study 1967*. New York.
19. Society of Actuaries and Association of Life Insurance Medical Directors of America (1938). *Impairment Study 1938*. New York.
20. Society of Actuaries and Association of Life Insurance Medical Directors of America

(1980). *Blood Pressure Study 1979*. Chicago. (Very readable.)

21. Society of Actuaries and Association of Life Insurance Medical Directors of America (1980). *Build Study 1979*. Chicago. (Highly readable.)

See also ACTUARIAL SCIENCE; CLINICAL TRIALS—II; EPIDEMIOLOGICAL STATISTICS—I; FOLLOW-UP; LIFE TABLES; RATES, STANDARDIZED; and VITAL STATISTICS.

EDWARD A. LEW

ACTUARIAL SCIENCE

Actuarial science is an applied mathematical and statistical discipline in which data-driven models are constructed to quantify and manage financial risk. The term "actuarial statistics" is not in common use because a well-defined set of statistical techniques useful to actuaries has not been established. This topic could also be viewed as a discussion of the types of data (mortality, morbidity*, accident frequency, and severity) collected by actuaries, but that will not be the focus here. This entry will concentrate on two particular statistical endeavors in which actuaries have played a major role—construction of mortality tables and credibility theory.

CONSTRUCTION OF MORTALITY TABLES

From the 1600s, governments sold annuities based on individual's lifetimes. To be useful as a fund-raising mechanism, the cost of the annuity needed to be greater than the expected cost of the benefit. Although not the first mortality table (or life table*), the work of Halley [10] combined the construction of a mortality table with the concept of expected present value. From the life table, for a person of current age x, it is possible to get the probability distribution of the number of years remaining, that is,

$$ {}_{k}|q_x = \text{Pr(death is between ages } x+k \text{ and } x+k+1),\ k=0,1,\ldots. $$

In addition, if the annuity is to pay one monetary unit at the end of each year, provided the annuitant is alive, the expected present value is

$$ a_x = \sum_{k=1}^{\infty} {}_{k}|q_x (v + v^2 + \cdots + v^k) $$

where $v = 1/(1+i)$ and i is the rate of interest.

A few years later, de Moivre* [19] introduced an approximation based on linear interpolation* between values in the life table (his table did not have survival probabilities at each integral age). This approximation continues to be used today and is referred to as the *uniform distribution of deaths* assumption [4], Chap. 3.

Life tables for actuarial use were constructed on an ad-hoc basis until the middle of the twentieth century when the so-called "actuarial method" was developed. It is loosely based on an assumption put forth by Balducci [2], viz.,

$$ \begin{aligned} &\text{Pr(a person age } x+t \text{ dies before age } x+1) \\ &\quad = (1-t)\,\text{Pr(a person age } x \\ &\qquad \text{dies before age } x+1), \end{aligned} $$

$0 < t < 1$ (*see* LIFE TABLES, BALDUCCI HYPOTHESIS). The result is an exposure-based formula that estimates the key life-table quantity as

$$ \begin{aligned} q_x &= \text{Pr(a person age } x \text{ dies before age } x+1) \\ &= \text{number of observed deaths / exposure.} \end{aligned} $$

For a life observed between ages x and $x+1$, the exposure contribution is the portion of the year the life was observed, except for deaths, for which the exposure is the time from first observation to age $x+1$.

This estimator is inconsistent [5]. However, it has one quality that made it extremely valuable. Given the types of records commonly kept by insurance companies, this formula was easy to implement by hand, or using mainframe computers prevalent through the 1980s. A good exposition of the actuarial method and its practical applications is Reference 3. Since then, actuaries have used the more accurate Kaplan-Meier* [13] and maximum likelihood estimation* procedures. These concepts are introduced in an actuarial setting in Reference 6.

Once the values of q_x have been obtained, a second actuarial contribution has been the smoothing of these values to conform with the *a priori* notion that from about age five onward the values should be smoothly increasing. The process of smoothing mortality rate estimates is called *graduation**. An introduction to all of the commonly used methods is given in Reference 17. Two of the more commonly used methods, interpolation*, and Whittaker, will be discussed here. Both methods create the graduated rates as a linear combination of surrounding values.

The interpolation method requires that the observations be grouped in a manner that creates estimates of q_x at every k (often 5) years of age. This is done by first aggregating the deaths (say, d_x) and exposures (say, e_x) at the surrounding ages, to create, for example, with $k = 5$,

$$d_x^* = d_{x-2} + d_{x-1} + d_x + d_{x+1} + d_{x+2},$$
$$e_x^* = e_{x-2} + e_{x-1} + e_x + e_{x+1} + e_{x+2}.$$

Because these series are often convex, an improved aggregated value can be found from King's pivotal point formula [14]:

$$d_x^{**} = -0.008d_{x-5}^* + 0.216d_x^* - 0.008d_{x+5}^*$$
$$e_x^{**} = -0.008e_{x-5}^* + 0.216e_x^* - 0.008e_{x+5}^*.$$

Finally, the mortality rate at age x is given by $q_x^{**} = d_x^{**}/u_x^{**}$.

The most commonly used interpolation formula is the Karup-King formula

$$\begin{aligned} q_{x+j} &= sq_{x+5}^{**} + 0.5s^2(s-1)\delta^2 q_{x+5}^{**} \\ &\quad +(1-s)q_x^{**} + 0.5(1-s)^2(-s)\delta^2 q_x^{**}, \\ &\qquad j = 0, 1, 2, 3, 4, 5, \end{aligned}$$

where $s = j/5$ and $\delta^2 q_x^{**} = q_{x+5}^{**} - 2q_x^{**} + q_{x-5}^{**}$ is the second central difference (*see* FINITE DIFFERENCES, CALCULUS OF). This formula uses four mortality rates and has the property that if those rates lie on a quadratic curve, the interpolated values will reproduce that curve. In addition, the cubic curves that connect consecutive mortality rates will have identical first derivatives where they meet.

Another popular formula is due to Jenkins [12] (see Eq. 7 in the entry GRADUATION). It requires fourth central differences and thus involves six points. It reproduces third-degree polynomials and adjacent curves will have identical first and second derivatives. To achieve these goals, the formula does not match the original mortality rates. That is, $q_{x+0} \neq q_x^{**}$.

The Whittaker method [22] can be derived by a Bayesian argument or from arguments similar to those used in creating smoothing splines (*see* SPLINE FUNCTIONS). Let q_x, $x = 0, \ldots, n$, be the original estimates; let v_x, $x = 0, \ldots, n$, be the graduated values; and let w_x, $x = 0, \ldots, n$, be a series of weights. Then, the graduated values are those that minimize the expression

$$\sum_{x=0}^{n} w_x(v_x - q_x)^2 + h\sum_{x=0}^{n-z}(\Delta^z v_x)^2$$

(*see* GRADUATION, WHITTAKER–HENDERSON). The weights are often chosen as either the exposure (sample size) at each age or the exposure divided by $q_x(1 - q_x)$, which would approximate using the reciprocal of the variance as the weight. The value of z controls the type of smoothing to be effected. For example, $z = 3$ leads to graduated values that tend to follow a quadratic curve. The choice of h controls the balance between fit (having the graduated values be close to the original values) and smoothing (having the graduated values follow a polynomial).

CREDIBILITY

Credibility theory is used by actuaries in the setting of premiums based on prior or corollary information. Two common situations are *experience rating* and *classification ratemaking*. An example of the former is workers compensation insurance. Suppose a particular employer had been charged a standard rate on the basis of expecting \$$x$ of claim payments per thousand dollars of payroll. In the previous year, the employer had claims of \$$y$ per thousand dollars of payroll, where $y < x$. The employer believes that a reduction in premium is warranted, while the insurer may claim that the result was simply good fortune.

A credibility procedure will base the next premium on the value $zy + (1 - z)x$, where $0 \leqslant z \leqslant 1$ and z is called the *credibility factor*. The magnitude of z is likely to depend on the sample size that produced y, the variance of y, and, perhaps, some measure of the accuracy of x.

With regard to classification ratemaking, consider setting premiums for automobile insurance. Separate rates may be needed for various combinations of gender, age, location, and accident history. Let y be an estimate based on the data for a particular combination of factors and let x be an estimate based on all the data. Because some combinations may occur infrequently, the reliability of y may be low. A credibility estimate using $zy + (1 - z)x$ may be more accurate (though biased). Credibility analysis succeeds for just that reason. By applying the factor $1 - z$ to an estimator that is more stable, the reduction in variance may offset the effect of bias, producing a smaller mean square error.

Two approaches to credibility have evolved. One, usually credited to Mowbray [20], has been termed *limited fluctuation credibility*. The question reduces to determining the sample size needed so that the relative error when estimating the mean will be less than $k\%$ with probability at least $p\%$. A normal or Poisson approximation along with a variance estimate is usually sufficient to produce the answer. If the sample size exceeds this number, then $z = 1$ (full credibility) is used. If not, z is customarily set equal to the square root of the ratio of the actual sample size to that needed for full credibility. Assuming no error in the quantity being multiplied by $1 - z$, the effect is to reduce the variance to equal that which would have been obtained with the full credibility sample size. The simplicity of this method causes it to remain popular. Its drawback is that it does not allow for the increased bias as z decreases, nor does it allow for any error in the quantity being multiplied by $1 - z$.

The second method has been termed *greatest accuracy credibility* and bears a strong resemblance to Bayesian analysis. It was introduced by Whitney [21] with a more thorough derivation produced by Bailey [1] and a modern derivation by Bühlmann [7]. This approach begins by assuming that a sample of size n is obtained from an individual. The observations are independent realizations of the random variable X with a distribution function that depends on the vector parameter θ. Define

$$E(X|\theta) = \mu(\theta) \text{ and } \mathrm{Var}(X|\theta) = v(\theta).$$

Further, assume that θ is unknown, but has been drawn at random from a random variable Θ with distribution function $F_\Theta(\theta)$. Finally, assume that $\mu(\theta)$ is to be estimated by a linear function of the observations, that is,

$$\widehat{\mu(\theta)} = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_n X_n.$$

The objective is to minimize

$$E_{\Theta,X_1,\ldots,X_n}\left\{\left[\widehat{\mu(\Theta)} - \mu(\Theta)\right]^2\right\}.$$

That is, the squared error should be minimized both over all possible observations and all possible parameter values. For a particular insured with a particular value of θ, the squared error may be larger than if the sample mean were used, but for others it will be smaller so that the overall error is reduced.

The solution is

$$\widehat{\mu(\theta)} = z\bar{x} + (1 - z)\mu;\ \ \mu = E[\mu(\Theta)],$$
$$z = \frac{n}{n + k},\ k = \frac{E[v(\Theta)]}{\mathrm{Var}[\mu(\Theta)]}.$$

It turns out to be the Bayesian (posterior mean) solution for certain common cases such as normal-normal and Poisson-gamma.

In practice, the indicated quantities must usually be estimated. An approach given in Bühlmann and Straub [8] provides an empirical Bayes* estimate, derived by a method of moments* approach. This is not unreasonable, because the distribution of Θ is not an *a priori* opinion, but rather a real, if unobservable, distribution of how characteristics vary from policyholder to policyholder or group to group. With data on several policyholders or groups, it is possible to estimate the needed moments. A true Bayesian model can be constructed by placing a prior distribution on the parameters of the distribution of Θ.

This is done for the normal-normal model in Reference 15.

Textbooks that develop these credibility topics and more (all include an English language version of the Bühlmann-Straub formula) include references 9, 11, 16, Chap. 5; and 18. A comprehensive list of book and article abstracts through 1982 is found in reference 23.

REFERENCES

1. Bailey, A. (1950). Credibility procedures. *Proc. Casualty Actuarial Soc.*, **37**, 7–23, 94–115.
2. Balducci, G. (1921). Correspondence. *J. Inst. Actuaries*, **52**, 184.
3. Batten, R. (1978). *Mortality Table Construction*. Prentice Hall, Englewood Cliffs, N.J.
4. Bowers, N., Gerber H., Hickman, J., Jones, D., and Nesbitt, C. (1997). *Actuarial Mathematics*, 2nd ed. Society of Actuaries, Schaumburg, Ill.
5. Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Stat.*, **2**, 437–453.
6. Broffitt, J. (1984). Maximum likelihood alternatives to actuarial estimators of mortality rates. *Trans. Soc. Actuaries*, **36**, 77–142.
7. Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bull.*, **4**, 199–207.
8. Bühlmann, H. and Straub, E. (1970). Glaubwürdigkeit für Schadensätze (credibility for loss ratios). *Mitteilungen der Vereinigung Schweizerisher Versicherungs-Mathematiker*, **70**, 111–133. (English translation (1972) in *Actuarial Research Clearing House*).
9. Dannenburg, D., Kass, R. and Goovaerts, M. (1996). *Practical Actuarial Credibility Models*. Ceuterick, Leuven, Belgium.
10. Halley, E. (1694). An estimate of the degrees of the mortality of mankind, drawn from curious tables of births and funerals at the city of Breslau; with an attempt to ascertain the price of annuities on lives. *Philos. Trans.*, **17**, 596–610.
11. Herzog, T. (1996). *Introduction to Credibility Theory*. Actex, Winsted, Conn.
12. Jenkins, W. (1927). Graduation based on a modification of osculatory interpolation. *Trans. Am. Soc. Actuaries*, **28**, 198–215.
13. Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.
14. King, G. (1887). Discussion: the graphic method of adjusting mortality tables (by T. Sprague). *J. Inst. Actuaries*, **26**, 114.
15. Klugman, S. (1987). Credibility for classification ratemaking via the hierarchical linear model. *Proc. Casualty Actuarial Soc.*, **74**, 272–321.
16. Klugman, S., Panjer, H. and Willmot, G. (1998). *Loss Models: From Data to Decisions*. Wiley, New York.
17. London, D. (1985). *Graduation: The Revision of Estimates*. Actex, Winsted, Conn.
18. Mahler, H. and Dean, C. (2001). *Credibility*. In *Foundations of Casualty Actuarial Science*, 4th ed. Casualty Actuarial Society, Arlington, Va.
19. de Moivre, A. (1725). *Annuities Upon Lives*. Fayram, Motte and Pearson, London.
20. Mowbray, A. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? *Proc. Casualty Actuarial Soc.*, **1**, 24–30.
21. Whitney, A. (1918). The theory of experience rating. *Proc. Casualty Actuarial Soc.*, **4**, 274–292.
22. Whittaker, E. and Robinson, G. (1924). *The Calculus of Observations*. Blackie and Sons, London.
23. de Wit, G., ed. (1986). Special issue on credibility theory. *Insurance Abstr. Rev.*, **2**(3).

See also Actuarial Health Studies; Demography; Graduation; Graduation, Whittaker–Henderson; Life Tables; Life Tables, Balducci Hypothesis; Morbidity; Multiple Decrement Tables; Population, Mathematical Theory of; Population Projection; Rates, Standardized; and Vital Statistics.

Stuart Klugman

ACTUARIAL STATISTICS. See Actuarial Science

ADAPTIVE IMPORTANCE SAMPLING (AIS). See Importance Sampling

ADAPTIVE METHODS

In adaptive statistical inference, we use the sample to help us select the appropriate type of statistical procedure needed for the situation under consideration. For a simple illustration of this, say that we use the sample kurtosis* K as a selector statistic [3]. One adaptive point estimator, T, for the center of a distribution would be

$$T = \begin{cases} \text{midrange}^*, & K \leqslant 2, \\ \text{arithmetic mean}^*, & 2 < K < 5, \\ \text{median}^*, & 5 \leqslant K. \end{cases}$$

That is, if the sample looks as if it arises from a short-tailed distribution, the average of the largest and smallest items of the sample is used as our estimator. If it looks like a long-tailed situation, the median is used. Otherwise, our estimate is the arithmetic mean (average) $\overline{x}$.

To generalize this illustration somewhat, suppose that we have a whole family (not necessarily finite) of possible distributions. Within this family of distributions, take a few representative ones, say $F_1, F_2, \ldots, F_k$. Now, for each of these k distributions, suppose that we can find a good statistic to make the inference under consideration. Let us say that these respective statistics are $T_1, T_2, \ldots, T_k$. We observe the sample from a distribution; and with a selector statistic, say Q, we determine which one of $F_1, F_2, \ldots, F_k$ seems closest to the underlying distribution from which the sample arises. If Q suggests that we have been sampling from F_i, then we would use the statistic T_i; or if Q suggests that we might be someplace between F_i and F_j, then we could use a combination of T_i and T_j; or more generally, Q could dictate a statistic that is a linear combination of all the statistics, $T_1, T_2, \ldots, T_k$: let us say

$$T = \sum_{i=1}^{k} W_i T_i, \quad \sum_{i=1}^{k} W_i = 1,$$

where the weights $W_1, W_2, \ldots, W_k$ are functions of the statistic Q. If it looks more like the sample arises from F_i, then, of course, the weight W_i would be large.

Consider a very simple example in which we are trying to choose the best of three types of concrete [10]. The compression strengths were tested after bars of each type had been exposed to severe environmental conditions for a period of 1 year. Seven ($n = 7$) observations were taken for each type of cement, where the observations are the breaking strengths of the bars measured in pounds per square inch. Let us denote the order statistics* of a sample by $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_7$. However, since we do not know from what underlying distribution these arose, we choose three representative distributions: the short-tailed uniform* using the midrange $(y_1 + y_7)/2$ as an estimate of center, the normal* using the average $\overline{x}$ as the estimate, and the long-tailed double exponential* with a modified median $(3y_3 + 8y_4 + 3y_5)/14$ as the statistic. These statistics were computed for each of the three samples and are given in Table 1 together with the original data.

Table 1.

Concrete	A	B	C
Ordered	5060	5625	4880
Observations	5398	6020	6030
	5820	6270	6290
	6131	6636	6372
	6400	6880	6920
	7527	7337	8320
	7560	8170	8581
Midrange	6310.0	6897.5	6730.5
Mean	6270.86	6705.43	6770.43
Modified median	6122.00	6609.86	6471.86

It is interesting to note that using the midrange or median statistics, concrete B looks to be the best, whereas $\overline{x}$ suggests concrete C. Accordingly, a selector statistic is needed, and we use

$$Q = \frac{(y_7 - y_1)/2}{\sum |y_i - M|/7},$$

the ratio of one-half of the range divided by the mean deviation* from the sample median M. (Q is defined somewhat differently when $n > 20$.) The average of the three Q values is computed to obtain $\overline{Q} = 1.876$. The midrange, average, or median is selected respectively, according to whether $\overline{Q}$ falls below, in between, or above

$$2.08 - (2/n) \quad \text{and} \quad 2.96 - (5.5/n);$$

that is, with $n = 7$, 1.794 and 2.174. (The formulas for these cutoffs have been determined empirically.) Since $\overline{Q} = 1.876$, it seems as if the distribution has fairly normal tails; thus the statistic $\overline{x}$ chooses concrete C as the best.

We must understand, however, that the inference under consideration is not necessarily a point estimate in the general situation. We could be considering a confidence interval* or a test of hypothesis*. Moreover, making an inference in this manner, that is, selecting the underlying distribution and then making the inference from the same data, can certainly destroy certain probabilities that are of interest in statistics. For example, if we are constructing a nominal 95% confidence interval, we can actually spoil the confidence coefficient* by such a procedure, so that it might actually be 0.80 or even 0.70. Or if we are making a test of a statistical hypothesis, the significant level might not be $\alpha = 0.05$, but 0.15 or 0.25. Despite this fact, however, the adaptive idea is useful in good data analysis; therefore, it is necessary for us to adjust our theories to the applications. That is, we want our theories to support the applications, not oppose them.

This forces us to look at some of the difficulties associated with the corresponding sampling distribution theory. Let us say that θ is the location parameter* and we are interested in testing the hypothesis $H_0 : \theta = \theta_0$ against the hypothesis $H_1 : \theta > \theta_0$. Again, suppose that we have a family of distributions for which θ is the location parameter of each member of the family. If we are sampling from F_i, say, we would reject the hypothesis H_0 and accept the alternative hypothesis H_1 if some statistic, say Z_i, was greater than or equal to c_i; i.e., $Z_i \geqslant c_i$, $i = 1, 2, \ldots, k$. Therefore, our adaptive test might be something like this: reject H_0 and accept H_1 if

$$Z = \sum_{i=1}^{k} W_i Z_i \geqslant c,$$

where $W_1, W_2, \ldots, W_k$ are functions of some selector statistic, say Q. The significance level of the test is then

$$\Pr \left| \sum_{i=1}^{k} W_i Z_i \geqslant c | H_0 \right| .$$

This probability is difficult to compute, so let us first consider a special and easier situation in which each of the W's is equal to 0 or 1. Of course, only one W_i can equal 1, and the rest must equal 0. Thus if Q suggests that F_i is the underlying distribution, then we will use Z_i. That is, if $Q \in R_i$, where $R_1, R_2, \ldots, R_k$ are appropriate mutually exclusive* and exhaustive sets, we will select Z_i for the test statistic. Under these conditions, the significance level would be

$$\sum_{i=1}^{k} \Pr[Q \in R_i \quad \text{and} \quad Z_i \geqslant c_i | H_0].$$

If each of the individual tests is made at the 0.05 significance level, it has been observed in practice that this significance level is frequently somewhat larger than that nominal significance level of 0.05.

There is a certain desirable element of model building in this entire procedure; that is, we observe the data and select the model that seems appropriate, and then we make the statistical inference* for the situation under consideration. However, there can be some cheating in doing this; that is, if we construct the model from given data and then make a test of hypothesis using those data, our nominal significance level is not necessarily the correct one. Moreover, even some researchers carry this to an extreme by selecting the test procedure that favors what they want (usually rejection of the null hypothesis*). They might then quote a significance level of 0.05, while the real α, for the overall selection and testing procedure might be higher than 0.25.

There is a method, however, of "legalizing" this cheating. Suppose that the selector statistic Q and each Z_i are independent under the null hypothesis H_0. Then the significance level is

$$\begin{aligned} &\sum_{i=1}^{k} \Pr[Q \in R_i \quad \text{and} \quad Z_i \geqslant c_i | H_0] \\ &\quad = \sum_{i=1}^{k} \Pr[Q \in R_i | H_0] \Pr[Z_i \geqslant c_i | H_0] \\ &\quad = \alpha \sum_{i=1}^{k} \Pr[Q \in R_i | H_0] = \alpha, \end{aligned}$$

provided that each individual test is made at the nominal significance level α. That is, this common significance level α is exactly the same as the overall significance level. The important feature of this is to make certain that the selector statistic Q is independent of the test statistic. One elegant way of achieving this is through distribution-free (nonparametric) methods* [5].

To illustrate the beauty of the nonparametric methods in these situations, let us consider the two-sample problem. Suppose that we have two independent continuous-type distributions, F and G. The null hypothesis H_0 is the equality of the two corresponding functions. Say that the sample $X_1, X_2, \ldots, X_m$ arises from F, and the sample $Y_1, Y_2, \ldots, Y_n$ arises from G. We suggest three nonparametric statistics that can be used to test this null hypothesis [2]. The first, Tukey's quick test*, is used when the underlying distributions have short tails, like those of the uniform distribution*. Tukey's statistic is

$$\begin{aligned} T_1 &= (\#Y\text{'s} > \text{ largest } X) \\ &\quad + (\#X\text{'s} < \text{ smallest } Y). \end{aligned}$$

A large T_1 would suggest the alternative hypothesis H_1 that the Y's tend to be larger than the X's. Thus we reject H_0 and accept H_1 if T_1 is greater than or equal to c_1, where

$$\Pr[T_1 \geqslant c_1 | H_0] = \alpha.$$

The second statistic T_2 is that of Mann, Whitney, and Wilcoxon. This statistic is a good one in case the underlying distributions have middle tails, like those of the normal* or logistic* distributions. After combining the two samples, we determine the ranks of the Y's in the combined sample; say those ranks are $R_1, R_2, \ldots, R_n$. One form of the Mann-Whitney-Wilcoxon statistic* is

$$T_2 = \sum_{i=1}^{n} R_i.$$

Now we reject H_0 and accept H_1 if T_2 is greater than or equal to c_2, where

$$\Pr[T_2 \geqslant c_2 | H_0] = \alpha.$$

The third statistic is that associated with the median test. It is $T_3 = \#Y$'s greater than the combined sample median. We reject H_0 if that statistic, T_3, is greater than or equal to c_3, where

$$\Pr[T_3 \geqslant c_3 | H_0] = \alpha.$$

Each of the probabilities denoted by α in these three tests does not depend on the form of the underlying continuous distribution, and sometimes these tests are called distribution-free tests.

For an example of each of these statistics, refer to data on the three types of concrete, and let the samples from A and B represent, respectively, the X and Y values with $m = n = 7$. The computed statistics are $T_1 = 3$, $T_2 = 59$, with $T_3 = 4$. Let us now consider an adaptive procedure that selects one of these three statistics. Considering the combined sample (i.e., the X's and Y's together) use a selector statistic, say Q, and decide whether we have short-tailed distributions, in which case we use the T_1 test; middle-tailed distributions, in which case we use the T_2 test; or long-tailed distributions, in which case we use the T_3 test. It turns out that the overall (selecting and testing) significance level will also equal α because each of T_1, T_2, and T_3 is independent of Q. The reason we have this independence under H_0 is that the order statistics of the combined sample are complete, sufficient statistics for the underlying "parameter," the common distribution $F = G$. Moreover, it is well known that the complete, sufficient statistics for $F = G$ are then independent of statistics whose distributions do not depend upon $F = G$, such as T_1, T_2, and T_3. However, the selector statistic Q is a function of the complete, sufficient statistics, and thus it is also independent of each of the statistics T_1, T_2, and T_3, under H_0. Incidentally, in our example, using the Q and $\overline{Q}$ associated with the illustration about concrete, the statistics T_2 for middle-tailed distributions would be selected and $T_2 = 59$ has a p-value of 0.288. Thus the null hypothesis would not be rejected at the significance level of $\alpha = 0.05$.

Although these nonparametric methods can be generalized to multivariate situations such as regression*, many statisticians do not find them extremely satisfactory in data analysis. Possibly the newer robust statistics show more promise in adaptation; some of them are "almost distribution-free" and lend themselves better to data analysis. Although it is impossible to give many details on robustness in this short article, the idea is illustrated with the trimmed mean.

Suppose that we are attempting to make an inference about the center θ of a symmetric distribution. Let $X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n$ represent the items of a random sample, ordered according to magnitude. The β-trimmed mean is

$$\overline{X}_\beta = \frac{1}{h}\sum_{i=g+1}^{n-g} X_i,$$

where β is usually selected so that $g = \eta\beta$ is an integer (otherwise, $g = [\eta\beta]$, the greatest integer in $\eta\beta$) and where $h = n - 2g$. Of course, $\overline{X}_{\beta=0} = \overline{X}$.

It is well known that

$$Z = \frac{\overline{X} - 0}{S/\sqrt{n-1}},$$

where $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/n$, has a t-distribution* with $n - 1$ degrees of freedom provided that the sample arises from a normal distribution. However, even though the underlying distribution is nonnormal (without really long tails), Z still has a distribution fairly close to this t-distribution. This is what we mean by "almost distribution-free." Now it is not so well known, but true [11], that

$$Z_\beta = \frac{\overline{X}_\beta - 0}{\sqrt{SS(\beta)/h(h-1)}},$$

where

$$\begin{aligned} SS(\beta) = {} & (g+1)(X_{g+1} - \overline{X}_\beta)^2 \\ & + (X_{g+2} - \overline{X})^2 + \cdots \\ & + (X_{n-g-1} - \overline{X})^2 \\ & + (g+1)(X_{n-g} - \overline{X}_\beta)^2, \end{aligned}$$

has an approximate t-distribution with $h - 1$ degrees of freedom for many underlying distributions, so that Z_β is almost distribution-free. Of course, $Z_{\beta=0} = Z$.

In an adaptive procedure using some Z_β to make an inference about θ, a selector statistic, such as the kurtosis K or Q, can be used to choose an appropriate β. This β will be larger for larger values of K and Q. In making inferences about θ based upon a selected Z_β, the overall confidence coefficient or the overall significance level will deviate somewhat from the nominal one. However, these deviations are not great; in many instances we have found that α equals something like 0.06 rather than the nominal $\alpha = 0.05$. Thus we can place great reliability on the level of the resulting inferences.

These adaptive and robust methods have been extended to multivariate situations and the interested reader is referred to some of the following articles and their references for further study. The future seems bright for adaptive methods, and these will bring applications and theory closer together.

REFERENCES

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location*. Princeton University Press, Princeton, N. J.
2. Conover, W. F. (1971). *Practical Nonparametric Statistics*. Wiley, New York.
3. Hogg, R. V. (1967). *J. Amer. Statist. Ass.*, **62**, 1179–1186.
4. Hogg, R. V. (1974). *J. Amer. Statist. Ass.*, **69**, 909–927.
5. Hogg, R. V., Fisher, D. M., and Randles, R. H. (1975). *J. Amer. Statist. Ass.*, **70**, 656–661.
6. Hogg, R. V. (1979). *Amer. Statist.*, **33**, 108–115.
7. Huber, P. J. (1973). *Ann. Math. Statist.*, **43**, 1041–1067.
8. Huber, P. J. (1973). *Ann. Statist.*, **1**, 799–821.
9. Jaeckel, L. A. (1971). *Ann. Math. Statist.*, **42**, 1540–1552.
10. Randles, R. H., Ramberg, J. S., and Hogg, R. V. (1973). *Technometrics*, **15**, 769–778.
11. Tukey, J. W. and McLaughlin, D. H. (1963). *Sankhyā A*, **25**, 331–352.

See also DISTRIBUTION-FREE METHODS; EXPLORATORY DATA ANALYSIS; and ROBUST ESTIMATION.

ROBERT V. HOGG

ADAPTIVE SAMPLING

Adaptive sampling is a method of unequal probability sampling whereby the selection of sampling units at any stage of the sampling process depends on information from the units already selected. In general terms it means that if you find what you are looking for at a particular location, you sample in the vicinity of that location with the hope of obtaining even more information.

Methods of estimation were initially developed in the three pioneering papers of Thompson [23–25] and the sampling book by Thompson [26]. The material considered in this review is described briefly by Seber and Thompson [20], while full details are given in the book by Thompson and Seber [31].

ADAPTIVE CLUSTER SAMPLING

Suppose we have a population spread over a large area which is highly clumped but is generally sparse or empty between clumps. If one selects a simple random sample (without replacement) of units, then most of the units selected will be empty. Density estimation based on this meager information will then have poor precision. Further, if the population species is rare, we will get little physiological information about individuals. It would be better to begin with an initial sample and, if individuals are detected on one of the selected units, then sample the neighboring units of that unit as well. If further individuals are encountered on a unit in the neighborhood, then the neighborhood of that unit is also added to the sample, and so on, thus building up a cluster of units. We call this *adaptive cluster sampling**. If the initial sample includes a unit from a clump, then the rest of the clump will generally be sampled. Such an approach will give us a greater number of individuals.

As well as counting individuals, we may wish to measure some other characteristic of the unit, for example plant biomass or pollution level, or even just note the presence or absence of some characteristic using an indicator variable. In addition to rare-species and pollution studies, we can envisage a wide range of populations which would benefit from adaptive sampling, for example populations which form large aggregations such as fish, marine mammals, and shrimp. We can also add mineral deposits and rare infectious diseases in human populations (e.g., AIDS) to our list. Recently the method has been used in sampling houses for a rare characteristic [5] and in sampling animal habitats [15].

To set out the steps involved in adaptive cluster sampling we begin with a finite population of N units indexed by their "labels" $(1, 2, \ldots, N)$. With unit i is associated a variable of interest y_i for $i = 1, 2, \ldots, N$. The object is to select a sample, observe the y-values for the units in the sample, and then estimate some function of the population y-values such as the population total $\sum_{i=1}^{N} y_i = \tau$ or the population mean $\mu = \tau/N$.

The first step is to define, for each unit i, a neighborhood consisting of that unit and a set of "neighboring" units. For example, we could choose all the adjacent units with a common boundary, which, together with unit i, form a cross. Neighborhoods can be defined to have a variety of patterns; the units (plots) in a neighborhood do not have to be contiguous. However, they must have a *symmetry* property, that is, if unit j is in the neighborhood of unit i, then unit i is in the neighborhood of unit j. We assume, for the moment, that these neighborhoods do not depend on y_i.

The next step is to specify a condition C (for instance, $y > c$, where c is a specified constant). We now take an initial random sample of n_1 units selected with or without replacement from the N units in the population. Whenever the y-value of a unit i in the initial sample satisfies C, all units in the neighborhood of unit i are added to the sample. If in turn any of the added units satisfies the condition, still more units are added. The process is continued until a cluster of units is obtained which contains a "boundary" of units called *edge* units that do not satisfy C. If a unit selected in the initial sample does not satisfy C, then there is no augmentation and we have a cluster of size one. The process is demonstrated in Fig. 1, where the units are plots and the neighborhoods form a cross. Here y_i is the number of animals on plot i, and $c = 0$, so that a neighborhood is added every time animals are found. In Fig. 1a we see one of the initial plots which happens to contain one animal. As it is on the edge of

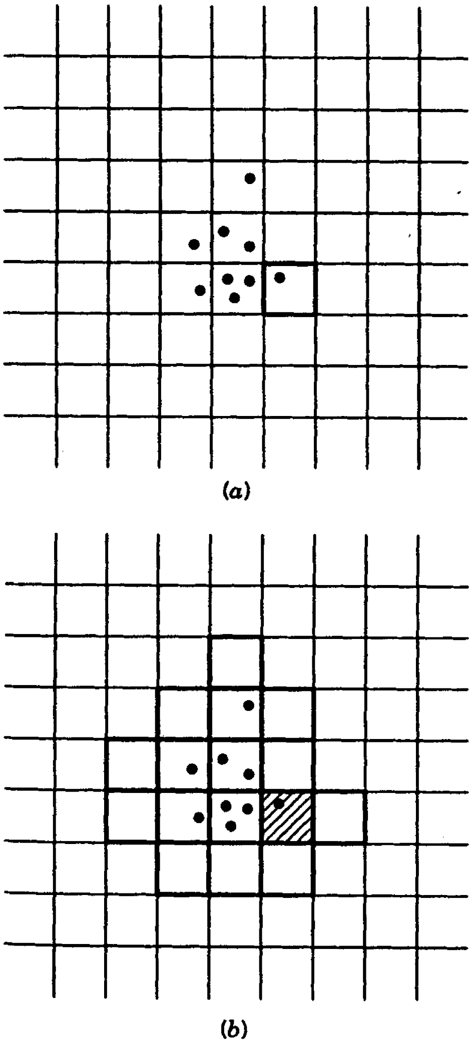


Figure 1. (*a*) Initial sample plot. (*b*) Cluster obtained by adding adaptively.

a "clump," we see that the adaptive process leads to the cluster of plots in Fig. 1*b*.

We note that even if the units in the initial sample are distinct, as in sampling without replacement, repeats can occur in the final sample, as clusters may overlap on their edge units or even coincide. For example, if two non-edge units in the same cluster are selected in the initial sample, then that whole cluster occurs twice in the final sample. The final sample then consists of n_1 (not necessarily distinct) clusters, one for each unit selected in the initial sample.

APPLICATIONS AND EXTENSIONS

In applications, other methods are sometimes used for obtaining the initial sample. For instance, in forestry the units are trees and these are usually selected by a method of unequal probability sampling*, where the probability of selecting a tree is proportional to the basal area of a tree (the cross-sectional area of a tree at the basal height—usually 4.5 feet in the USA). Roesch [16] described a number of estimators for this situation.

In ecology, larger sample units other than single plots are often used. For example, a common sampling unit is the strip transect, which we might call the primary unit. In its adaptive modification, the strip would be divided up into smaller secondary units, and if we found animals in a secondary unit, we would sample units on either side of that unit, with still further searching if additional animals are sighted while on this search. Strips are widely used in both aerial and ship surveys of animals and marine mammals. Here the aircraft or vessel travels down a line (called a *line transect**), and the area is surveyed on either side out to a given distance. Thompson [24] showed how the above theory can be applied to this sampling situation. He pointed out that a primary unit need not be a contiguous set of secondary units. For example, in some wildlife surveys the selection of sites chosen for observation is done systematically (with a random starting point), and a single systematic selection then forms the primary unit. We can then select several such primary units without replacement and add adaptively as before. Such a selection of secondary units will tend to give better coverage of the population then a simple random sample.

Clearly other ways of choosing a primary unit to give better coverage are possible. Munholland and Borkowski [13,14] suggest using a Latin square* + 1 design selected from a square grid of secondary units (plots). The Latin square gives a secondary unit in every row and column of the grid, and the extra (i.e. +1) unit ensures that any pair of units has a positive probability of being included in the initial sample. The latter requirement is needed for unbiased variance estimation.

In some situations it is hard to know what c should be for the condition $y > c$. If we choose c too low or too high, we end up with a feast or famine of extra plots. Thompson [28] suggested using the data themselves, in fact the order statistics*. For example, c could be the rth largest y-value in the initial sample statistic, so that the neighborhoods are now determined by the y-values. This method would be particularly useful in pollution studies, where the location of "hot spots" is important.

Another problem, regularly encountered with animal population studies, is that not all animals are detected. Thompson and Seber [30] developed tools for handling incomplete detectability for a wide variety of designs, including adaptive designs, thus extending the work of Steinhorst and Samuel [22].

Often we are in a multivariate situation where one needs to record several characteristics or measurements on each unit, e.g. the numbers of different species. Thompson [27] pointed out that any function of the variables can be used to define the criterion C, and obtained unbiased estimates of the mean vector and covariance matrix for these variables.

We can use any of the above methods in conjunction with stratification. If we don't allow the clusters to cross stratum boundaries, then individual stratum estimates are independent and can be combined in the usual fashion. Thompson [25] extended this theory to allow for the case where clusters do overlap. Such an approach makes more efficient use of sample information.

Finally, there are two further developments relating to design, namely, selecting networks without replacement and a two-stage sampling procedure [17,18].

UNBIASED ESTIMATION

Although the cluster is the natural sample group, it is not a convenient entity to use for theoretical developments, because of the double role that edge units can play. If an edge unit is selected in the initial sample, then it forms a cluster of size 1. If it is not selected in the initial sample, then it can still be selected by being a member of any cluster for which it is an edge unit. We therefore introduce the idea of the network A_i for unit i, defined to be the cluster generated by unit i but with its edge units removed. In Fig. 1(b) we get the sampled network by omitting the empty units from the sampled cluster. Here the selection of *any* unit in the network leads to the selection of *all* of the network. If unit i is the only unit in a cluster satisfying C, then A_i consists of just unit i and forms a network of size 1. We also define any unit which does not satisfy C to be a network of size 1, as its selection does not lead to the inclusion of any other units. This means that all clusters of size 1 are also networks of size 1. Thus any cluster consisting of more than one unit can be split into a network and further networks of size 1 (one for each edge unit). In contrast to having clusters which may overlap on their edge units, the distinct networks are *disjoint* and form a *partition* of the N units.

Since the probability of selecting a unit will depend on the size of the network it is in, we are in the situation of unequal-probability sampling and the usual estimates based on equal-probability sampling will be biased. However, we have the well-known Horvitz—Thompson* (HT) and Hansen–Hurwitz (HH) estimators (cf. refs. [8] and [9]) for this situation, the latter being used in sampling with replacement. These estimators, however, require knowing the probability of selection of each unit in the final sample. Unfortunately these probabilities are only known for units in networks selected by the initial sample and not for the edge units attached to these networks. Therefore, in what follows we ignore all edge units which are not in the initial sample and use only network information when it comes to computing the final estimators.

Motivated by the HT estimator for the population mean μ, we consider

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i \frac{I_i}{E[I_i]},$$

where I_i takes the value 1 if the initial sample intersects network A_i, and 0 otherwise; $\hat{\mu}$ is an unbiased estimator for sampling with or without replacement.

Another possible estimator (motivated by the HH estimator) which is also obviously unbiased for sampling with or without replacement, is

$$\tilde{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i \frac{f_i}{E[f_i]},$$

where f_i is the number of times that the ith unit in the final sample appears in the estimator, that is, the number of units in the initial sample which fall in (intersect) A_i determined by unit i; $f_i = 0$ if no units in the initial sample intersect A_i. It can be shown that

$$\tilde{\mu} = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \overline{w}, \quad \text{say},$$

where w_i is the mean of the observations in A_i, i.e., $\overline{w}$ is the mean of the n_1 (not necessarily distinct) network means. *See also* NETWORK ANALYSIS.

ADAPTIVE ALLOCATION

There are other ways of adaptively adding to an initial sample. For instance, suppose the population is divided up into strata or primary units each consisting of secondary units. An initial sample of secondary units is taken in each primary unit. If some criterion is satisfied such as $\overline{y} > c$, then a further sample of units is taken from the *same* primary unit. Kremers [12] developed an unbiased estimator for this situation.

If the clumps tend to be big enough so that they are spread over several primary units, we could use what is found in a particular primary unit to determine the level of the sampling in the next. This is the basis for the theory developed by Thompson et al. [29]. Other forms of augmenting the initial sample which give biased estimates are described by Francis [6,7] and Jolly and Hampton [10,11]. This kind of adaptive sampling based on allocating more units rather than adding more neighborhoods is called *adaptive allocation*.

RAO—BLACKWELL MODIFICATION

An adaptive sample can be defined as one for which the probability of obtaining the sample depends only on the distinct unordered y-observations in the sample, and not on the y-values outside the sample. In this case d, the set of distinct unordered labels in the sample together with their associated y-values, is minimal sufficient for μ. This is proved for "conventional designs" by Cassel et al. [3] and Chaudhuri and Stenger [4], and their proofs readily extend to the case of adaptive designs. (This extension is implicit in Basu [1].) This means that an unbiased estimator which is not a function of d can be "improved" by taking the expectation of the estimator conditional on d to give an estimator with smaller variance. For example, consider three unbiased estimators of μ, namely $\overline{y}_1$ (the mean of the initial sample of n_1 units), $\hat{\mu}$, and $\tilde{\mu}$. Each of these depends on the order of selection, as they depend on which n_1 units are in the initial sample; $\tilde{\mu}$ also depends on repeat selections; and when the initial sample is selected with replacement, all three estimators depend on repeat selections. Since none of the three estimators is a function of the minimal sufficient statistic d, we can apply the Rao—Blackwell theorem*. If T is any one of the three estimators, then $E[T|d]$ will give a better unbiased estimate, i.e. one with smaller variance. We find that this estimator now uses all the units including the edge units.

Finally we mention the "model-based" or "superpopulation" approach (cf. Särndal et al. [19], for example). Here the population vector $\mathbf{y}$ of y-values is considered to be a realization of a random vector $\mathbf{Y}$ with some joint distribution F, which may depend on an unknown parameter ϕ. In a Bayesian framework ϕ will have a known prior distribution. For this model-based approach, Thompson and Seber [31] indicate which of the results for conventional designs carry over to adaptive designs and which do not. They also show in their Chapter 10 that optimal designs tend to be adaptive.

RELATIVE EFFICIENCY

An important question one might ask about adaptive sampling is "How does it compare with, say, simple random sampling?" This question is discussed by Thompson and Seber [31, Chapter 5], and some guidelines are given. Cost considerations are also important. Simple examples given by them throughout their book suggest that there are large gains in efficiency to be had with clustered populations. Two simulation studies which shed light on this are by Brown [2] and Smith et al. [21].

REFERENCES

1. Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā A*, **31**, 441–454.
2. Brown, J. A. (1994). The application of adaptive cluster sampling to ecological studies. In *Statistics in Ecology and Environmental Monitoring*, D. J., Fletcher and B. F. J., Manly. eds., University of Otago Press, Dunedin, New Zealand, pp. 86–97.
3. Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*, Wiley, New York.
4. Chaudhuri, A. and Stenger, H. (1992). *Survey Sampling: Theory and Methods*. Marcel Dekker, New York.
5. Danaher, P. J. and King, M. (1994). Estimating rare household characteristics using adaptive sampling. *New Zealand Statist.*, **29**, 14–23.
6. Francis, R. I. C. C. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand J. Marine and Freshwater Res.*, **18**, 59–71.
7. Francis, R. I. C. C. (1991). Statistical properties of two-phase surveys: comment. *Can. J. Fish. Aquat. Sci.*, **48**, 1128.
8. Hansen, M. M. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, **14**, 333–362.
9. Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.
10. Jolly, G. M. and Hampton, I. (1990). A stratified random transect design for acoustic surveys of fish stocks. *Can. J. Fish. Aquat. Sci.*, **47**, 1282–1291.
11. Jolly, G. M. and Hampton, I. (1991). Reply of comment by R. I. C. C. Francis. *Can. J. Fish. Aquat. Sci.*, **48**, 1128–1129.
12. Kremers, W. K. (1987). Adaptive Sampling to Account for Unknown Variability Among Strata. *Preprint No. 128*, Institut für Mathematik, Universität Augsburg, Germany.
13. Munholland, P. L. and Borkowski, J. J. (1993). Adaptive Latin Square Sampling +1 Designs. *Technical Report No. 3-23-93*, Department of Mathematical Sciences, Montana State University, Bozeman.
14. Munholland, P. L. and Borkowski, J. J. (1996). Latin square sampling +1 designs. *Biometrics*, **52**, 125–132.
15. Ramsey, F. L. and Sjamsoe'oed, R. (1994). Habitat association studies in conjunction with adaptive cluster samples. *J. Environmental Ecol. Statist.*, **1**, 121–132.
16. Roesch, F. A., Jr. (1993). Adaptive cluster sampling for forest inventories. *Forest Sci.*, **39**, 655–669.
17. Salehi, M. M. and Seber, G. A. F. (1997). Adaptive cluster sampling with networks selected without replacement. *Biometrika*, **84**, 209–219.
18. Salehi, M. M. and Seber, G. A. F. (1997). Two-stage adaptive cluster sampling. *Biometrics*, **53**, 959–970.
19. Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
20. Seber, G. A. F. and Thompson, S. K. (1994). Environmental adaptive sampling. In *Handbook of Statistics, Vol. 12 (Environmental Sampling)*, G. P. Patil and C. R. Rao, eds., New York, North Holland/Elsevier Science, pp. 201–220.
21. Smith, D. R., Conroy, M. J., and Brakhage, D. H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics*, **51**, 777–788.
22. Steinhorst, R. K. and Samuel, M. D. (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics*, **45**, 415–425.
23. Thompson, S. K. (1990). Adaptive cluster sampling. *J. Amer. Statist. Ass.*, **85**, 1050–1059.
24. Thompson, S. K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, **47**, 1103–1115.
25. Thompson, S. K. (1991). Stratified adaptive cluster sampling. *Biometrika*, **78**, 389–397.
26. Thompson, S. K. (1992). *Sampling*. Wiley, New York.
27. Thompson, S. K. (1993). Multivariate aspects of adaptive cluster sampling. In *Multivariate Environmental Statistics*, G. P. Patil and C. R. Rao, eds., New York, North Holland/Elsevier Science, pp. 561–572.
28. Thompson, S. K. (1995). Adaptive cluster sampling based on order statistics. *Environmetrics*, **7**, 123–133.
29. Thompson, S. K., Ramsey, F. L., and Seber, G. A. F. (1992). An adaptive procedure for

sampling animal populations. *Biometrics*, **48**, 1195–1199.

30. Thompson, S. K. and Seber, G. A. F. (1994). Detectability in conventional and adaptive sampling. *Biometrics*, **50**, 712–724.
31. Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive Sampling*. Wiley, New York.

See also ADAPTIVE METHODS; CLUSTER SAMPLING; LINE TRANSECT SAMPLING; POPULATION SIZE, HORVITZ-THOMPSON ESTIMATOR FOR; RAO–BLACKWELL THEOREM; TRANSECT METHODS; and UNEQUAL PROBABILITY SAMPLING.

GEORGE A. F. SEBER

ADDITION THEOREM

Let A_i and A_j be two events defined on a sample space. Then

$$\Pr[A_i \cup A_j] = \Pr[A_i] + \Pr[A_j] - \Pr[A_i \cap A_j],$$

where $\Pr[A_i \cup A_j]$ denotes the probability of A_i or A_j or both occurring, $\Pr[A_i]$ and $\Pr[A_j]$ denote respectively the probability of A_i and the probability of A_j, and $\Pr[A_i \cap A_j]$ denotes the probability of both A_i and A_j occurring.

The theorem is extended for the general case of n events as follows:

$$\begin{aligned}\Pr[A_1 \cup \cdots \cup A_n] &= \sum_{i=1}^{n} \Pr[A_i] \\ &- \sum_{i_1}^{n-1} \sum_{<i_2}^{n} \Pr[A_{i_1} \cap A_{i_2}] \\ &+ \sum_{i_1}^{n-2} \sum_{<i_2}^{n-1} \sum_{<i_3}^{n} \Pr[A_{i_1} \cap A_{i_2} \cap A_{i_3}] \\ &- \cdots + (-1)^{n+1} \Pr[\cap_{i=1}^{n} A_i].\end{aligned}$$

It is also called *Waring's theorem*.

See also BONFERRONI INEQUALITIES AND INTERVALS; BOOLE'S INEQUALITY; and INCLUSION-EXCLUSION METHOD.

ADDITIVE RISK MODEL, AALEN'S

THE MODEL

In medical statistics and survival analysis*, it is important to assess the association between risk factors and disease occurrence or mortality. Underlying disease mechanisms are invariably complex, so the idea is to simplify the relationship between survival patterns and covariates in such a way that only essential features are brought out. Aalen's (1980) additive risk model [1] is one of three well-developed approaches to this problem, the others being the popular proportional hazards model introduced by D. R. Cox in 1972 (*see* PROPORTIONAL HAZARDS MODEL, COX'S), and the accelerated failure-time model, which is a linear regression model with unknown error distribution, introduced in the context of right-censored survival data by R. G. Miller in 1976.

Aalen's model expresses the conditional hazard function $\lambda(t|\mathbf{z})$ of a survival time T as a linear function of a p-dimensional covariate vector $\mathbf{z}$:

$$\lambda(t|\mathbf{z}) = \boldsymbol{\alpha}(t)'\mathbf{z} = \sum_{j=1}^{p} \alpha_j(t) z_j, \qquad (1)$$

where $\boldsymbol{\alpha}(t)$ is a nonparametric p-vector of regression functions [constrained by $\lambda(t|\mathbf{z}) \geqslant 0$] and $\mathbf{z} = (z_1, \ldots, z_p)'$. Some authors refer to (1) as the linear hazard model.

As a function of the covariates $z_1, \ldots, z_p$, the *additive* form of Aalen's model contrasts with the *multiplicative* form of Cox's model:

$$\begin{aligned}\lambda(t|\mathbf{z}) &= \lambda_0(t) \exp\{\boldsymbol{\beta}'\mathbf{z}\} \\ &= \lambda_0(t) \prod_{j=1}^{p} \exp\{\beta_j z_j\},\end{aligned}$$

where $\lambda_0(t)$ is a nonparametric baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression parameters. Aalen's model has the feature that the influence of each covariate can vary separately and nonparametrically through time, unlike Cox's model or the accelerated failure-time model. This feature can be desirable in some applications, especially when there are a small number of covariates.

Consider the following simple example with three covariates: T is the age at which an individual contracts melanoma (if at all), $z_1 =$ indicator male, $z_2 =$ indicator female, and $z_3 =$ number of serious sunburns as a child. Then the corresponding regression functions, α_1, α_2, and α_3, can be interpreted as the (age-specific) background rates of melanoma for males and females and as the excess rate of melanoma due to serious sunburns is childhood, respectively.

Aalen's model is expected to provide a reasonable fit to data, since the first step of a Taylor series expansion of a general conditional hazard function about the zero of the covariate vector can be expressed in the form (1). It is somewhat more flexible than Cox's model and can be especially helpful for exploratory data analysis*. A rough justification for the additive form can be given in terms of p independent competing risks*, since the hazard function of the minimum of p independent random variables is the sum of their individual hazard functions.

It is generally sensible to include a nonparametric baseline function in the model, by augmenting $\mathbf{z}$ with a component that is set to 1. Also, it is often natural to center the covariates in some fashion, so the baseline can be interpreted as the "hazard" function for an "average" individual. In some cases, however, a baseline hazard is already implicit in the model and it is not necessary to center the covariates, as in the melanoma example above.

Aalen originally proposed his model in a counting process* setting, which allows time-dependent covariates and general patterns of censorship, and which can be studied using powerful continuous-time martingale* techniques. In a typical application the observed survival times are subject to right censorship, and it is customary to assume that the censoring time, C say, is conditionally independent of T given $\mathbf{z}$. One observes $(X, \delta, \mathbf{z})$, where $X = T \wedge C$ and $\delta = I\{X = T\}$. Aalen's model (1) is now equivalent to specifying that the counting process $N(t) = I(X \leqslant t, \delta = 1)$, which indicates an uncensored failure by time t, has intensity process

$$\lambda(t) = \boldsymbol{\alpha}(t)'\mathbf{y}(t),$$

where $\mathbf{y}(t) = \mathbf{z}I\{X \geqslant t\}$ is a covariate process.

MODEL FITTING

To fit Aalen's model one first estimates the p-vector of integrated regression functions $\mathbf{A}(t) = \int_0^t \boldsymbol{\alpha}(s)\,ds$. Denote by $(t_i, \delta_i, \mathbf{z}_i)$ the possibly right-censored failure time t_i, indicator of noncensorship δ_i, and covariate vector $\mathbf{z}_i$ for n individuals. Let $\mathbf{N} = (N_1, \ldots, N_n)'$ and $\mathbf{Z} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$, where N_i is the counting process and $\mathbf{y}_i$ is the associated covariate process for individual i.

Aalen [1] introduced an ordinary least squares (OLS) type estimator of $\mathbf{A}(t)$ given by

$$\hat{\mathbf{A}}(t) = \int_0^t (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\,d\mathbf{N},$$

where the matrix inverse is assumed to exist; $\hat{\mathbf{A}}$ is a step function, constant between uncensored failures, and with jump

$$\boldsymbol{\Delta}_i = \left(\sum_{t_k \geqslant t_i} \mathbf{z}_k\mathbf{z}_k'\right)^{-1} \mathbf{z}_i \qquad (2)$$

at an uncensored failure time t_i. The matrix inverse exists unless there is collinearity* between the covariates or there are insufficiently many individuals at risk at time t_i. A heuristic motivation for $\hat{\mathbf{A}}$ comes from applying the method of least squares to increments of the multivariate counting process $\mathbf{N}$. The estimator is consistent and asymptotically normal [14,9]. The covariance matrix of $\hat{\mathbf{A}}(t)$ can be estimated [1,2] by $\hat{\mathbf{V}}(t) = \sum_{t_i \leqslant t} \delta_i \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i'$.

Plots of the components of $\hat{\mathbf{A}}(t)$ against t, known as *Aalen plots*, are a useful graphical diagnostic tool for studying time-varying covariate effects [2,3,4,7,9,12,13]. Mau [12] coined the term Aalen plots and made a strong case for their importance in survival analysis*. Roughly constant slopes in the plots indicate periods when a covariate has a non-time-dependent regression coefficient; plateaus indicate times at which a covariate has no effect on the hazard. Interpretation of the plots is helped by the inclusion of pointwise or simultaneous confidence limits. An approximate pointwise $100(1-\alpha)\%$ confidence interval for the jth component of $\mathbf{A}(t)$ is given by

$$\hat{\mathbf{A}}_j(t) \pm z_{\alpha/2}\hat{\mathbf{V}}_{jj}(t)^{1/2},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution and $\hat{\mathbf{V}}_{jj}(t)$ is the jth entry on the diagonal of $\hat{\mathbf{V}}(t)$. To avoid wild fluctuations in the plots (which occur when the size of the risk set is small), estimation should be restricted to time intervals over which the matrix inverse in (2) is numerically stable.

Figure 1 shows an Aalen plot based on survival data for 495 myelamatosis patients [17]. The plot gives the estimated integrated regression function for one particular covariate, serum β_2-microglobulin, which was log-transformed to adjust for skewness. Pointwise 95% confidence limits are also shown. Serum β_2-microglobulin is seen to have a strong effect on survival during the first two years of follow-up.

The vector of regression functions $\boldsymbol{\alpha}$ can be estimated by smoothing the increments of $\hat{\mathbf{A}}$. One approach is to extend Ramlau-Hansen's kernel estimator [19] to the additive-risk-model setting [3,9,14]. For a kernel function K that integrates to 1 and some bandwidth $b > 0$,

$$\hat{\boldsymbol{\alpha}}(t) = b^{-1}\sum_{i=1}^{n} K\left(\frac{t-t_i}{b}\right)\boldsymbol{\Delta}_i$$

consistently estimates $\boldsymbol{\alpha}$ provided the bandwidth tends to zero at a suitable rate with increasing sample size. Plots of the regression function estimates in some real and simulated data examples have been given by Aalen [3].

Huffer and McKeague [9] introduced a weighted least squares* (WLS) estimator of $\mathbf{A}$; see Fig. 2 for a comparison with the OLS estimator. The weights consistently estimate $[\lambda(t|\mathbf{z}_i)]^{-1}$ and are obtained by plugging $\hat{\alpha}(t)$ and $\mathbf{z} = \mathbf{z}_i$ into (1). The WLS estimator is an approximate maximum-likelihood estimator and an approximate solution to the score equations [20]. It is consistent and asymptotically normal provided $\lambda(t|\mathbf{z})$ is bounded away from zero [9,14]. Furthermore, the WLS

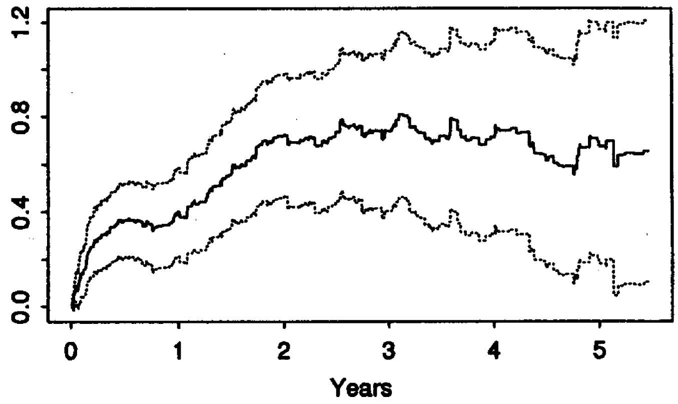


Figure 1. An Aalen plot with 95% confidence limits for the myelamatosis data.

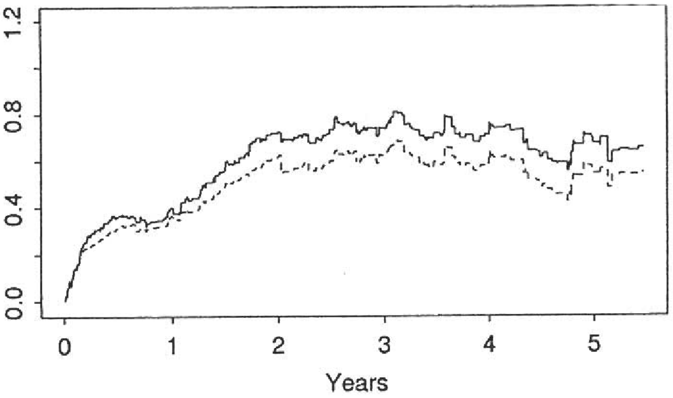


Figure 2. Comparison of Aalen plots of the WLS estimates (dashed line) and OLS (solid line) estimates for the myelamatosis data.

estimator is asymptotically efficient in the sense of having minimal asymptotic variance [6,20,4]. Simulation studies [9] show that significant variance reductions are possible using WLS compared with OLS estimators, especially in large samples where the weights are more stable. When there are no covariates ($p = 1$ and $\mathbf{z}_i = 1$), the OLS and WLS estimators reduce to the Nelson—Aalen estimator of the baseline cumulative hazard function. Simultaneous confidence bands for $\mathbf{A}$ based on OLS and WLS estimators, for continuous or grouped data, can be found in refs. 9, 15.

Tests of whether a specific covariate (say the jth component of $\mathbf{z}$) has any effect on survival can be carried out within the Aalen-model setting. The idea is to test the null hypothesis $H_0 : \mathbf{A}_j(t) = 0$ over the follow-up period. This can be done [2] using a test statistic of the form $\sum_{i=1}^{n} w(t_i)\boldsymbol{\Delta}_{ij}$ for a suitable weight function w. Kolmogorov—Smirnov-type tests* are also available [9]; such tests are equivalent to checking whether the confidence band for the jth component of $\mathbf{A}$ contains the zero function.

To predict survival under Aalen's model, one estimates the conditional survival probability $P(T > t|\mathbf{z}) = \exp\{-\mathbf{A}(t)'\mathbf{z}\}$. This can be done using the product-limit estimator

$$\hat{P}(T > t|\mathbf{z}) = \prod_{t_i \leqslant t}(1 - \boldsymbol{\Delta}_i'\mathbf{z}),$$

or by plugging $\hat{\mathbf{A}}(t)$ into $P(T > t|\mathbf{z})$ in place of the unknown $\mathbf{A}(\mathbf{t})$. When there are no covariates, $\hat{P}(T > t|\mathbf{z})$ reduces to the Kaplan–Meier estimator* of the survival function corresponding to the baseline hazard.

MODEL DIAGNOSTICS

Some goodness-of-fit checking procedures are available for additive risk models. Aalen [2,3] suggested making plots against t of sums of the martingale residual processes $\hat{M}_i(t) = \delta_i I(t_i \leqslant t) - \hat{\mathbf{A}}(t_i \wedge t)'\mathbf{z}_i$ over groups of individuals. If the model fits well, then the plots would be expected to fluctuate around the zero line. McKeague and Utikal [16] suggested the use of a standardized residual process plotted against t and $\mathbf{z}$, and developed a formal goodness-of-fit test for Aalen's model.

Outlier detection has been studied by Henderson and Oman [8], who considered the effects on $\hat{\mathbf{A}}(t)$ of deletion of an observation from a data set. They show that unusual or influential observations* can be detected quickly and easily. They note that Aalen's model has an advantage over Cox's model in this regard, because closed-form expressions for the estimators are available, leading to exact measures of the effects of case deletion.

Mau [12] noticed that Aalen plots are useful for diagnosing time-dependent covariate effects in the Cox model. To aid interpretation of the plots in that case, Henderson and Milner [7] suggested that an estimate of the shape of the curve expected under proportional hazards be included.

RELATED MODELS

In recent years a number of variations on the additive structure of Aalen's model have been introduced. McKeague and Sasieni [17] considered a partly parametric additive risk model in which the influence of only a subset of the covariates varies nonparametrically over time, and that of the remaining covariates is constant:

$$\lambda(t|\mathbf{x}, \mathbf{z}) = \boldsymbol{\alpha}(t)'\mathbf{x} + \boldsymbol{\beta}'\mathbf{z}, \qquad (3)$$

where $\mathbf{x}$ and $\mathbf{z}$ are covariate vectors and $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ are unknown. This model may be more appropriate than (1) when there are a large number of covariates and it is known that the influence of only a few of the covariates is time-dependent. Lin and Ying [10] studied an additive analogue of Cox's proportional hazards model that arises as a special case of (3):

$$\lambda(t|\mathbf{z}) = \alpha_0(t) + \boldsymbol{\beta}'\mathbf{z}. \qquad (4)$$

Efficient WLS-type estimators for fitting (3) and (4) have been developed.

A variation in the direction of Cox's proportional hazards model [5] has been studied by Sasieni [21,22]: the *proportional excess hazards* model

$$\lambda(t|\mathbf{x}, \mathbf{z}) = \alpha_0(t|\mathbf{x}) + \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{z}\}, \qquad (5)$$

where $\alpha_0(t|\mathbf{x})$ is a known background hazard (available from national mortality statistics say) and $\lambda_0(t)$ and $\boldsymbol{\beta}$ are unknown. A further variation in this direction is due to Lin and Ying [11], who considered an *additive–multiplicative hazards* model that includes

$$\lambda(t|\mathbf{x},\mathbf{z}) = \boldsymbol{\gamma}'\mathbf{x} + \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{z}\}, \qquad (6)$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\lambda_0(t)$ are unknown. Finding efficient procedures for fitting the models (5) and (6) involves a combination of Cox partial likelihood* techniques and the estimation of efficient weights similar to those needed for the standard additive risk model (1).

CONCLUSION

Despite the attractive features of Aalen's model as an alternative to Cox's model in many application, it has received relatively little attention from practitioners or researchers. Cox's model has been perceived to be adequate for most applications, but it can lead to serious bias when the influence of covariates is time-dependent. Fitting separate Cox models over disjoint time intervals (years, say) is an ad hoc way around this problem. Aalen's model, however, provides a more effective approach. Interest in it, and especially in Aalen plots, is expected to increase in the future.

Acknowledgment
This research was partially supported by NSF Grant ATM-9417528.

REFERENCES

1. Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Lecture Notes Statist.*, **2**, 1–25. Springer-Verlag, New York. (Aalen originally proposed his model at a conference in Poland; this paper appeared in the conference proceedings.)
2. Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statist. Med.*, **8**, 907–925. (A readable introduction to additive risk models with an emphasis on graphical techniques. Results based on the Aalen and Cox models are compared.)
3. Aalen, O. O. (1993). Further results on the nonparametric linear regression model in survival analysis. *Statist. Med.*, **12**, 1569–1588. (Studies the use of martingale residuals for assessing goodness-of-fit of additive risk models.)
4. Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. (A comprehensive and in-depth survey of the counting process approach to survival analysis, including the additive risk model and its associated estimators.)
5. Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, 187–220.
6. Greenwood, P. E. and Wefelmeyer, W. (1991). Efficient estimating equations for nonparametric filtered models. In *Statistical Inference in Stochastic Processes*, N. U. Prabhu and I. V. Basawa, eds., pp. 107–141. Marcel Dekker, New York. (Shows that the weighted least-squares estimator is an approximate maximum-likelihood estimator and that this implies asymptotic efficiency.)
7. Henderson, R. and Milner, A. (1991). Aalen plots under proportional hazards. *Appl. Statist.*, **40**, 401–410. (Introduces a modification to Aalen plots designed to detect time-dependent covariate effects in Cox's proportional-hazards model.)
8. Henderson, R. and Oman, P. (1993). Influence in linear hazard models. *Scand. J. Statist.*, **20**, 195–212. (Shows how to detect unusual or influential observations under Aalen's model.)
9. Huffer, F. W. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Ass.*, **86**, 114–129.
10. Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.
11. Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive–multiplicative hazard models for counting processes. *Ann. Statist.*, **23**, 1712–1734.
12. Mau, J. (1986). On a graphical method for the detection of time-dependent effects of covariates in survival data. *Appl. Statist.*, **35**, 245–255. (Makes a strong case that Aalen plots can provide important information that might be missed when only Cox's proportional-hazards model is applied.)
13. Mau, J. (1988). A comparison of counting process models for complicated life histories. *Appl. Stoch. Models Data Anal.*, **4**, 283–298.

(Aalen's model is applied in the context of intermittent exposure in which individuals alternate between being at risk and not.)

14. McKeague, I. W. (1988). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model. In *Statistical Inference from Stochastic Processes*, N. U. Prabhu, ed., *Contemp. Math.*, 80, 139–152. American Mathematical Society, Providence. (Studies the asymptotic properties of estimators in the counting process version of Aalen's model.)
15. McKeague, I. W. (1988). A counting process approach to the regression analysis of grouped survival data. *Stoch. Process. Appl.*, **28**, 221–239. (Studies the asymptotic properties of grouped data based estimators for Aalen's model.)
16. McKeague, I. W. and Utikal, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.*, **18**, 177–195.
17. McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, **81**, 501–514.
18. Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**, 449–464.
19. Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453–466.
20. Sasieni, P. D. (1992). Information bounds for the additive and multiplicative intensity models. In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Goel, eds., pp. 249–265. Kluwer, Dordrecht. (Shows asymptotic efficiency of the weighted least-squares estimators in the survival analysis setting.)
21. Sasieni, P. D. (1995). Efficiently weighted estimating equations with application to proportional excess hazards. *Lifetime Data Analysis*, in press.
22. Sasieni, P. D. (1996). Proportional excess hazards. *Biometrika*, **83**, 127–141.

See also COUNTING PROCESSES; PROPORTIONAL HAZARDS MODEL, COX'S; SEMIPARAMETRICS; and SURVIVAL ANALYSIS.

IAN W. MCKEAGUE

ADMISSIBILITY

DEFINITION

Admissibility is a very general concept that is applicable to any procedure of statistical inference*. The statistical literature contains discussions of admissibility of estimators*, confidence intervals*, confidence sets, tests of hypotheses, sampling designs in survey sampling*, and so on. For each class of procedures, there is formulated a definition of admissibility which is appropriate for that class only. But all such definitions are based on a common underlying notion—that a procedure is admissible if and only if there does not exist within that class of procedures another one which performs uniformly at least as well as the procedure in question and performs better than it in at least one case. Here "uniformly" always means for all values of the parameter* (or parameters) that determines the (joint) probability distribution* of the random variables under investigation. It thus remains only to define how the condition of "performing as well as" is interpreted in each case. All such definitions are based closely on that of the admissibility of a decision rule* formulated in Abraham Wald's theory of statistical decision functions or decision theory*, as it is briefly called. In fact, the importance of the notion of admissibility in statistical theory rests on the adoption of the decision-theoretic approach to statistical problems formulated in Wald's theory.

DECISION THEORY

Wald's theory formulates the following general model for statistical decision making. Let $\mathscr{S}$ denote the sample space of the random variables under investigation, of which the true (joint) probability distribution is unknown, it being known only to belong to a family $\mathscr{P} = (P_\theta, \theta \in \Omega)$. Depending upon the object of the investigation (e.g., point or interval estimation or hypothesis testing*, etc.), there is a specific set $\mathscr{A}$ of all possible decisions a which the statistician may make. A decision rule δ is a function which prescribes for each sample point z how the decision would be made, i.e., some specific a chosen from $\mathscr{A}$. (δ may either assign a unique a to each z—such a rule is called a nonrandomized rule—or it may assign to each z a probability distribution δ_z on $\mathscr{A}$, the choice of a specific a being made according to that probability distribution by an independent random experiment.) Consequences of wrong decisions are allowed for by assuming a suitable nonnegative loss function* $\mathscr{L}(a, \theta)$. When the experiment, i.e., the taking of observations on the random variables, is repeated a large number of times under identical conditions, the average long-run loss converges to its "expectation" or "mean value" $\mathscr{R}(\theta, \delta)$ which is called the risk function*. Admissibility of decision rules is defined in terms of the risk function as follows:

A decision rule δ_2 is better than a rule δ_1 if

$$\mathscr{R}(\theta, \delta_2) \leqslant \mathscr{R}(\theta, \delta_1) \quad \text{for all } \theta \in \Omega \qquad (1)$$

and the strict inequality in (1) holds for at least one $\theta \in \Omega$.

A decision rule δ is admissible if there exists no rule δ_1 which is better than δ. *See* DECISION THEORY for further details.

A LIMITATION

A limitation of the admissibility principle may be noted here. The central problem of decision theory is: Under the conditions of a given decision problem, how should the choice of a decision rule be effected from the class of all decision rules? The admissibility criterion requires that inadmissible rules be left out of consideration. This leaves the class of admissible rules, which, however, is generally very large, and admissibility says nothing as to how a choice should be made from this large class. The choice is therefore made in practice by applying other statistical principles, such as unbiasedness*, minimaxity, and invariance*, or by taking into consideration the statistician's prior beliefs* regarding the weight to be attached to the different possible values of the parameter. In the last-mentioned case, if the prior beliefs are expressed as a probability distribution τ on Ω, the risk function $\mathscr{R}(\theta, \delta)$ integrated with respect to τ gives the Bayes risk $r(\tau, \delta)$. The appropriate decision rule, called the Bayes rule, is then the one that minimizes $r(\tau, \delta)$ for given τ. This is the Bayesian mode of inference. In Bayesian inference*, there is thus an optimum decision rule which often is unique, determined by the prior beliefs, and hence the concept of admissibility has less importance. In Wald's theory, however, the approach is non-Bayesian and based on the long-run frequency. Thus the risk function $\mathscr{R}(\theta, \delta)$ represents the average loss in the long run when the experiment is repeated a large number of times under identical conditions. There was the same approach in the earlier Neyman–Pearson* theories of interval estimation and hypothesis testing, which are included in Wald's theory as particular cases. This approach is often referred to as the N-P-W approach. The importance of the criterion of admissibility is thus related to the N-P-W approach.

Another point to be noted is that if a decision rule is inadmissible, there exists another that should be used in preference to it. But this does not mean that every admissible rule is to be preferred over any inadmissible rule. It is easy to construct examples of rules that are admissible but which it would be absurd to use. An example in point estimation is an estimator that is equal to some constant k whatever be the observations. The risk function then vanishes for $\theta = k$ and the estimator is an admissible one. (See ref. 11 for a more sophisticated example.)

PARTICULAR PROCEDURES

Any statistical inference procedure such as point or interval estimation* corresponds simply to a class of decision rules. But the definition of admissibility considered appropriate for a particular procedure may not always be equivalent to the decision theory definition in terms of a risk function. For example, consider interval estimation of the parameter θ in a probability density function $f(x, \theta)$ on the basis of independent observations. Let x denote collectively the observations and T_1, T_2 the statistics (functions of x) defining the confidence interval. [The pair (T_1, T_2) constitutes in this case the decision rule.] Admissibility for confidence intervals is defined as follows. "A set of confidence intervals $\{T_1, T_2\}$ is admissible if there exists no other set $\{T_1^*, T_2^*\}$ such that

(a) $$T_2^*(\mathbf{x}) - T_1^*(\mathbf{x}) \leqslant T_2(\mathbf{x}) - T_1(\mathbf{x}) \quad \text{for all } \mathbf{x},$$

$$\text{(b)}\quad P_\theta\{T_1^*(\mathbf{x}) \leqslant \theta \leqslant T_2^*(\mathbf{x})\} \geqslant P_\theta\{T_1(\mathbf{x}) \leqslant \theta \leqslant T_2(\mathbf{x})\} \quad \text{for all } \theta,$$

the strict inequality in (b) holding for at least one θ. [The probabilities in (b) are the inclusion probabilities.] This definition is obviously not reducible to one based on the risk function, as there are two inequalities; moreover, the first of these is required to hold at each sample point $\mathbf{x}$, there being no averaging over the sample space, which is an essential ingredient in the definition of a risk function.

In the case of point estimation*, on the other hand, the definition of admissibility is identical with that for decision rules. A loss function that is found reasonable and mathematically convenient, particularly in point estimation problems, is the squared error, i.e., $L(t, \theta) = c(t - \theta)^2$, where t is the estimate and c is any positive constant. Further if, as often is the case, the estimators are restricted to the class of unbiased* estimators, then the admissibility criterion reduces to one based on the variance or equivalently the efficiency* of the estimator.

In the case of hypothesis testing* of the null hypothesis $\theta = \theta_0$, the decision-theoretic definition of admissibility reduces to one based on the power function* in the Neyman–Pearson theory by a suitable choice of the loss function, namely, by putting $\mathscr{L}(\theta, a_1) = 0$ and $\mathscr{L}(\theta, a_2) = 1$, where for the value θ, a_1 is the correct decision and a_2 the incorrect one. The set $\mathscr{A}$ consists in this case of only two points, corresponding to the rejection or nonrejection of the null hypothesis. (See ref. 2 for a paper dealing with admissibility of tests.)

SPECIAL TYPES OF ADMISSIBILITY

These are extensions of the basic notion of admissibility.

Strong and Weak Admissibility

In some cases it is found necessary to introduce weak and strong versions of admissibility, strong admissibility being based on a more stringent criterion. For example, in the problem of interval estimation (see the section “Particular Procedures”), if condition (a) is replaced by

$$\text{(a}^*\text{)}\quad E_\theta\{T_2^* - T_1^*\} \leqslant E_\theta\{T_2 - T_1\} \quad \text{for all } \theta,$$

and the sign of strict inequality required to hold for at least one θ in either (a*) or in (b), we obtain a more stringent criterion as the set of alternatives is enlarged. (See ref. 8 for weak and strong admissibility of confidence sets.)

ϵ-Admissibility

A decision rule δ_0 is said to be ϵ-admissible if there exists no other decision rule δ_1 such that

$$\mathscr{R}(\theta, \delta_1) < \mathscr{R}(\theta, \delta_0) - \epsilon \quad \text{for all } \theta \in \Omega.$$

(See “Decision Theory” section of this entry for definitions of these terms.)

ϵ-admissibility provides a measure of the extent by which an inadmissible rule falls short of being admissible. (See ref. 10 for an application.)

Uniform Admissibility

This term is special to survey sampling* theory. Earlier investigations had related mostly to the admissibility of a particular estimator e_1 under a given sampling design p_1. But in survey sampling, the choice of the sampling design is generally, subject to certain limitations of cost and time, within the statistician’s control. This leads to the notion of the joint admissibility of the pair (e_1, p_1) within a class of pairs (e, p). (Such a pair is now called a *sampling strategy* or, more simply, a *strategy*.) It would be pointless to consider the joint admissibility within the class of all possible pairs (e, p), as then the only admissible sampling design would be that in which the whole population is observed with probability 1. It is therefore necessary to place a restriction on the class $\mathscr{P}$ of designs. The restrictions usually assumed are that the expected sample size or the expected sampling cost under p should not exceed certain limits, as these are the restraints that generally apply in practice. Of course, the particular sampling design p_1

must also satisfy the restriction. The term "uniform admissibility" denotes just this joint admissibility of an estimator e_1 and a sampling design p_1, defined as usual, within a class of pairs (e, p) such that p belongs to a specified class $\mathscr{C}$ of designs. Note that the uniform admissibility of (e_1, p_1) is a stronger property than the admissibility of e_1 under the given design p_1, as the former implies the latter. (See refs. 14 and 15 for some recent related results.)

Hyperadmissibility

This notion is special to survey sampling theory* and denotes broadly that an estimator is admissible for the population as a whole and also for every possible subpopulation of that population (see ref. 6 for more details).

Admissibility within a Restricted Class

It is often necessary to consider the admissibility of a procedure within a restricted class of procedures. For example, in the case of point estimation, an unbiased estimator T_0 of a parameter θ is said to be admissible within the unbiased class if there exists no other unbiased estimator T_1 of θ that is "better" than T_0.

RELATIONS WITH COMPLETENESS* AND EFFICIENCY*

In Wald's theory, the notion of admissibility is intimately related to that of completeness. The theory requires that the statistician should restrict the choice of a decision rule to the class of all admissible rules. But in general, there is no simple characteristic that distinguishes the class of all admissible rules from that of all inadmissible ones. This leads to the notion of a complete class that contains all the admissible rules: "A class $\mathscr{C}$ of decision rules is said to be *complete* if given any rule δ_1 not in $\mathscr{C}$, there exists at least one rule δ_0 in $\mathscr{C}$ that is better than δ_1." (See the "Decision Theory" section for definition of betterness.) Hence if a class of rules is known to be complete, the statistician may validly restrict the choice of decision rule to such class as all the excluded rules are necessarily inadmissible. Of course, after choosing a particular rule from a complete class, it would have to be tested for its admissibility. It is further shown in the theory that there exists a simply characterized class of decision rules (the class of generalized Bayes rules) which under very general conditions forms a complete class.

Essential completeness is a sharpening of the notion of completeness "A class $\mathscr{C}$ of decision rules is said to be *essentially complete* if given any rule δ_1 not in $\mathscr{C}$, there exists at least one rule δ in $\mathscr{C}$ which is as good as δ_1, i.e., such that $\mathscr{R}(\theta, \delta_0) \leqslant \mathscr{R}(\theta, \delta_1)$ for all $\theta \in \Omega$." Clearly, a statistician may validly restrict the choice of a decision rule to an essentially complete class if it exists. It is shown in the theory that "the class of decision rules based on a sufficient statistic is always essentially complete." This proposition provides the decision-theoretic justification for the sufficiency principle* in statistics. For some related propositions such as the Rao–Blackwell theorem*, see ref. 4.

STEIN'S RESULT

A notable result relating to admissibility is that of Stein [16]: For k independent normal variables, the sample means are jointly inadmissible for the population means with the squared errors as loss function if $k \geqslant 3$. The theoretical and practical implications of Stein's results have been a matter of debate [1,1]; *see* JAMES–STEIN ESTIMATORS and SHRINKAGE ESTIMATORS.

SURVEY SAMPLING*

Survey sampling essentially involves no new point of principle. The commonly considered estimation problem is to estimate the population total. If the squared error is taken as the loss function, as is often the case, the admissibility of an estimator is defined as follows: An estimator $e(s, \mathbf{x})$ is admissible if there does not exist any other estimator $e'(s, \mathbf{x})$ such that

$$\sum_s p(s)[e'(s, \mathbf{x}) - T(\mathbf{x})]^2$$
$$\leqslant \sum_s p(s)[e(s, \mathbf{x}) - T(\mathbf{x})]^2 \quad \text{for all } \mathbf{x}$$

and the strict inequality holds for at least one $\mathbf{x}$. Here $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ denotes the population vector and is the parameter, $T(\mathbf{x}) = \sum_{i=1}^{N} x_i$, $p(s)$ is the probability of the sample s under the chosen sampling design, and N denotes the number of units in the population. The estimator $e(s, \mathbf{x})$ must, of course, depend only on the x_i observed in the sample. See SURVEY SAMPLING for further details.

The following important general results have been proved recently. (The lateness of the results is a consequence of the fact the correct model for survey sampling was developed only after 1950; see SURVEY SAMPLING.)

1. The sample mean is admissible as the estimator of the population mean in the entire class of all estimators whatever the sampling design, and for a very wide class of loss functions [7].

2. The Horwitz–Thompson estimator* is always admissible in the restricted class of unbiased estimators [5].

Suppose that samples are taken independently from k different finite populations. Are the sample means together jointly admissible for the population means with squared error as loss function? It is found that they are. Thus in the case of finite populations, an effect corresponding to Stein's result for the multivariate normal population does not occur. This is a very recent result [9].

REFERENCES

1. Alam, K. (1975). *J. Multivariate Anal.*, **5**, 83–95.
2. Brown, L. D., Cohen, A., and Strawderman, W. E. (1979). *Ann. Statist.*, **3**, 569–578.
3. Efron, B. and Morris, C. (1973). *J. R. Statist. Soc. B*, **35**, 379–421.
4. Ferguson, T. S. (1967). *Mathematical Statistics: A Decision-Theoretical Approach*. Academic Press, New York.
5. Godambe, V. P. and Joshi, V. M. (1965). *Ann. Math. Statist.*, **36**, 1707–1722.
6. Hanurav, T. V. (1968). *Ann. Math. Statist.*, **39**, 621–641.
7. Joshi, V. M. (1968). *Ann. Math. Statist.*, **39**, 606–620.
8. Joshi, V. M. (1969). *Ann. Math. Statist.*, **40**, 1042–1067.
9. Joshi, V. M. (1979). *Ann. Statist.*, **7**, 995–1002.
10. Kagan, A. M. (1970). *Sankhyā A*, **32**, 37–40.
11. Makani, S. M. (1977). *Ann. Statist.*, **5**, 544–546.
12. Neyman, J. (1937). *Philos. Trans. R. Soc. Lond. A*, **236**, 333–380.
13. Neyman, J. and Pearson, E. S. (1933). *Philos. Trans. R. Soc. Lond. A*, **231**, 289–337.
14. Scott, A. J. (1975). *Ann. Statist.*, **3**, 489–491.
15. Sekkappan, R. M. and Thompson, M. E. (1975). *Ann. Statist.*, **3**, 492–499.
16. Stein, C. (1956). *Proc. 3rd Berkeley Symp. Math. Stat. Prob.*, Vol. 1. University of California Press, Berkeley, Calif., pp. 197–206.
17. Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

See also BAYESIAN INFERENCE; DECISION THEORY; ESTIMATION, CLASSICAL; HYPOTHESIS TESTING; INFERENCE, STATISTICAL; JAMES–STEIN ESTIMATORS; SHRINKAGE ESTIMATORS; and SURVEY SAMPLING.

V. M. JOSHI

ADVANCES IN APPLIED PROBABILITY.

See APPLIED PROBABILITY JOURNALS

AFFLUENCE AND POVERTY INDEXES.

See INDEXES, AFFLUENCE AND POVERTY

AGGREGATE

The word "aggregate" has several meanings. As a verb, it means putting together, or combining, elements that usually differ in some notable respect. As a noun, it is used to describe the result of this process. The word is also sometimes used as a synonym for "total," as in "aggregate production" and "aggregate debt."

In geology, and especially in mining engineering, the word is specifically applied to collections of samples of ore.

See also ARITHMETIC MEAN and GEOLOGY, STATISTICS IN.

AGGREGATE INDEX NUMBERS.

See INDEX NUMBERS

AGGREGATION

Aggregation may be a phenomenon of direct interest, as in the study of biological populations, or it may reflect the necessary reduction of primary data to produce a usable statistical summary, as in the construction of index numbers*. Since the two topics are quite distinct, we consider them separately.

AGGREGATION AS AN OBSERVABLE PHENOMENON

It is often true that events (individuals) cluster in time or space or both (e.g., larvae hatching from eggs laid in a mass, aftershocks of an earthquake). Thus if the random variable of interest is the number of events occurring in an interval of time (or in a selected area), the clustering is manifested in a greater probability of extreme events (large groups) than would be expected otherwise. Alternatively, individual members of a population may be in close proximity because of environmental conditions. In either case, the population is said to be aggregated.

A standard initial assumption (corresponding to the absence of aggregation) is that the random variable follows a Poisson distribution*, and various indices have been proposed to detect departures from the Poisson process*. These methods are based upon data collected either as quadrat counts or as measurements of distance (or time) from randomly selected individuals (or points) to the nearest individual, known as nearest-neighbor distances. For example, the index of dispersion is defined as $I = s^2/m$, where m and s^2 denote the sample mean and variance. For the Poisson process, $E(I|m > 0) \doteq 1$; values of I significantly greater than 1 suggest aggregation; $I < 1$ is indicative of regular spacing of the individuals [5, Chap. 4]. Other measures, based upon both quadrat counts and distances, are summarized in Pielou [15, Chaps. 8 and 10] and Cormack [6]. When different kinds of individual (e.g., species) have different aggregation patterns, this make inferences about population characteristics such as diversity* much more difficult.

If the Poisson process is used to describe parents (or centers), each parent may give rise to offspring (or satellites). If these clusters are independent but have identical size distributions, the resulting distribution for the total count is a (Poisson) randomly stopped sum distribution*. If environmental heterogeneity is postulated, a compound distribution*, usually based on the Poisson, is appropriate. For both classes of Poisson-based distributions, $I > 1$. These standard distributions lack an explicit spatial or temporal dimension, for which a dispersal mechanism must be incorporated. The resulting model, known as a center-satellite process, has three components: a Poisson process for locating the cluster center, a distribution to generate the number of satellites, and a dispersal distribution to describe displacements from the center. This class of processes was introduced by Neyman and Scott [14] and is mathematically equivalent to the class of doubly stochastic Poisson processes defined for heterogeneity (See Bartlett [3, Chap. 1]).

A more empirical approach to aggregation is that of Taylor [18], who suggests that the population mean, μ, and variance, σ^2, are related by the power law:

$$\sigma^2 = A\mu^b, \qquad A > 0, \quad b > 0.$$

It is argued that values of b greater than 1 (the Poisson value) reflect density dependence in the spatial pattern of individuals. Although this view has been contested (see the discussion in Taylor [18]), a substantial body of empirical evidence has been presented in its support [19].

Knox [12] developed a test to detect the clustering of individuals in space and time, which may be formulated as follows. Suppose that n individuals (e.g., cases of a disease) are observed in an area during a time period. If cases i and j are less than a specified critical distance from one another, set the indicator variable $w_{ij} = 1$; otherwise, set $w_{ij} = 0$. Similarly, if i and j occur within a specified time of one another, set $y_{ij} = 1$; otherwise, set $y_{ij} = 0$. Then the space-time interaction coefficient is

$$\mathrm{STI} = \sum_{i \neq j}\sum w_{ij} y_{ij}$$

For example, for a disease such as measles, we might consider cases within 1 mile of each other occurring 10 days or less apart (the length of the latent period). If n_S and n_T denote the number of adjacent pairs in space and time, respectively, and both are small relative to n, then the conditional distribution of STI given n_S, n_T, and n is approximately Poisson with expected value $n_S n_T/n$. The test has been extended to several spatial and temporal scales by Mantel [13]. For further details, see Cliff and Ord [5, Chaps. 1 and 2].

[*Editor's Addendum*. A population that has its individuals evenly distributed is frequently called *overdispersed*, while an aggregated population with its individuals clustered in groups can be described as *underdispersed*; *see* OVERDISPERSION.]

AGGREGATION AS A STATISTICAL METHOD

Aggregation in this sense involves the compounding of primary data in order to express them in summary form. Also, such an exercise is necessary when a model is specified at the micro (or individual) level but the usable data refer to aggregates. Then the question that arises is whether the equations of the micro model can be combined in such a way as to be consistent with the macro (or aggregate) model to which the data refer.

We may wish to compound individual data records, such as consumers' expenditures, or to combine results over time and/or space. The different cases are described in turn.

Combining Individual Records

Consider a population of N individuals in which the ith individual $(i = 1, \ldots, N)$ has response Y_i to input x_i of the form

$$Y_i = f(x_i, \beta_i) + \epsilon_i$$

where β_i denotes a (vector of) parameter(s) specific to the ith individual and ϵ_i denotes a random-error term. For example, the equation may represent the consumer's level of expenditure on a commodity given its price. Then the total expenditure is $Y = \sum Y_i$ (summed over $i = 1, \ldots, N$), and the average input is $x = \sum x_i/N$.

In general, it is not possible to infer an exact relationship between Y and x from the micro relations. The few results available refer to the linear aggregation of linear equations. Theil [20] showed that when f denotes a linear function so that

$$Y_i = \alpha_i + \beta_i x_i + \epsilon_i,$$

perfect aggregation is possible, in that we may consider a macro relation of the form

$$Y = \alpha + \beta x^* + \epsilon,$$

where $\alpha = \sum \alpha_i$, $\beta = \sum \beta_i$, $x^* = \sum \beta_i x_i/\beta$, and $\epsilon = \sum \epsilon_i$. That is, we must use the weighted average x^* rather than the natural average x. Further, a different aggregation procedure is required for each regressor variable and for the same regressor variable with different response variables [2, Chap. 20; 20, Chap. 2]. If we use the natural average, the macro relation is

$$Y = \alpha + \beta x + N \operatorname{cov}(x_i, \beta_i) + \epsilon,$$

where the covariance is evaluated over the N members of the population and represents the aggregation bias. This bias is small, for example, when x is the price variable in a consumer demand equation, but may be much more substantial when x denotes consumers' income in such a relationship. When the micro relationship is a nonlinear function of x_i, the nonlinearity will generate a further source of aggregation bias. It must be concluded that exact aggregation is rarely possible, although the bias may be small in many cases. For further discussion and recent developments of the theory, see Ijiri [11].

Aggregation of Groups

Instead of forming a macro relation from a known group of individuals, we may wish to identify suitable groups from a finer classification of individuals. This is a necessary step in the construction of broad industrial classifications for use in input-output systems. See ECONOMETRICS. Blin and Cohen [4] propose a method of cluster analysis* for solving this problem.

Temporal Aggregation

Variates may be continuous or summed over a unit time interval, although the variate is recorded only as an aggregate over periods of r units duration. For a model that is linear in the regressor variables and has time-invariant parameters, aggregation is straightforward provided that there are no lagged variables. However, if

$$Y_t = \alpha + \beta x_{t-k} + \epsilon_t$$

for some $k > 0$ and k not a multiple of r, exact aggregation is not possible; any aggregated model will involve x-values for two or more time periods [20, Chap. 4]. Such models are often formulated using distributed lags*. Also, Granger and Morris [9] show that the autoregressive–moving average (ARMA) models* are often appropriate in this case.

The aggregation of time series* exhibiting positive autocorrelation tends to increase the values of the various test statistics and thereby give rise to overoptimistic assessments of the model [17]. However, Tiao and Wei [21] have shown that whereas aggregation can considerably reduce the efficiency of parameter estimators*, it has much less effect upon prediction efficiency. Indeed, it has been shown that there are circumstances where forecasts from aggregate relations may be more accurate than an aggregate of forecasts from micro relations [1,10].

The discussion so far has assumed that β does not vary over time. For a discussion of aggregation when there is a change of regime (time-varying parameters), see Goldfeld and Quandt [8, Chap. 4].

Spatial Aggregation

Many problems in spatial aggregation are similar to those of time series, but they are further compounded by the (sometimes necessary) use of areas of irregular shape and different size. Yule and Kendall [22] first showed how different aggregations of spatial units affect estimates of correlation*. Cliff and Ord [5, Chap. 5] give a general review of methods for estimating the autocovariance and cross-correlation functions using a nested hierarchy of areal sampling units. The estimation of these functions for irregular areas depends upon making rather restrictive assumptions about the nature of interaction between areas.

Cliff and Ord considered data from the *London Atlas*; a 24 × 24 lattice of squares of side 500 meters was laid over the Greater London area and the percentage of land used for commercial (X), industrial (Y), office (Z), and other purposes was recorded for each square. The correlation between each X and Y for different combinations of grid squares were as follows:

	Size of Spatial Unit			
	1 × 1	2 × 2	4 × 4	8 × 8
corr(X, Z)	0.19	0.36	0.67	0.71
corr(Y, Z)	0.09	0.16	0.33	0.34

The correlation functions exhibit an element of mutual exclusion at the smallest levels, and the positive correlation for larger spatial units indicates the general effects of areas zoned for housing and nonhousing purposes. Here, as in time series, we must think in terms of a distance or time-dependent correlation function and not a unique “correlation” between variables.

A PERSPECTIVE

Aggregation appears both as a phenomenon of interest in its own right and as a necessary evil in modeling complex processes. The center-satellite models have proved useful in astronomy [14], in ecology [3,15], in geography [5], and several other disciplines. At the present time, such processes offer a flexible tool for simulation work, although further work on the theory and analysis of such processes is desirable; the data analytic distance methods of Ripley [16] represent a useful step in the right direction. In epidemiology [12,13] and hydrology (models of storms, etc.) the development of clusters in both space and time is important, although relatively little work exists to date. The work of Taylor [18] represents a challenge to the theoretician, as useful models generating the empirical regularities observed by Taylor are still lacking.

Econometricians seem to be turning away from the view that an aggregate model is the sum of its parts and placing more emphasis upon aggregated models per se. The nonlinearities of the aggregation procedure, combined with the complexities of the underlying processes [8], suggest that aggregated models with time-dependent parameters are likely to play an increasing role in economics and other social sciences.

Where the level of aggregation is open to choice [4], further work is needed to identify suitable procedures for combining finer units into coarser ones. Similar problems arise in quadrat sampling* [15, p. 222].

The use of a sample to estimate the mean over an area or volume is of interest in the geosciences (e.g., drillings in an oil field). Estimators for such aggregates are based upon a variant of generalized least squares* known as "kriging"*; see Delfiner and Delhomme [7] and the papers of Matheron cited therein for further details.

In all these areas, much remains to be discussed about the statistical properties of the estimators currently used, and there is still plenty of scope for the development of improved methods.

REFERENCES

1. Aigner, D. J., and Goldfeld, S. M. (1974). *Econometrica*, **42**, 113–134.
2. Allen, R. G. D. (1959). *Mathematical Economics*. Macmillan, London. (Outlines the aggregation problem in econometric modeling and sets it in the context of other aspects of mathematical economics; written at an intermediate mathematical level.)
3. Bartlett, M. S. (1975). *The Statistical Analysis of Spatial Pattern*. Chapman & Hall, London. (A concise introduction to the theory of spatial point processes and lattice processes with a variety of applications in ecology.)
4. Blin, J. M. and Cohen, C. (1977). *Rev. Econ. Statist.*, **52**, 82–91. (Provides a review and references on earlier attempts to form viable aggregates as well as new suggestions.)
5. Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models, Inference and Applications*. Pion, London. (Discusses aggregation problems in the context of spatial patterns with examples drawn from ecology and geography; written at an intermediate mathematical level.)
6. Cormack, R. M. (1979). In *Spatial and Temporal Processes in Ecology*, R. M. Cormack and J. K. Ord, eds. International Co-operative Publishing House, Fairland, Md., pp. 151–211. (An up-to-date review of spatial interaction models with an extensive bibliography. Other papers in the volume cover related aspects.)
7. Delfiner, P. and Delhomme, J. P. (1975). In *Display and Analysis of Spatial Data*, J. C. Davis and J. C. McCullagh, eds. Wiley, New York; pp. 96–114. (This volume contains several other papers of general interest on spatial processes.)
8. Goldfeld, M. and Quandt, R. E. (1976). *Studies in Nonlinear Estimation*. Cambridge, Mass.: Ballinger.
9. Granger, C. W. J. and Morris, J. J. (1976). *J. R. Statist. Soc. A*, **139**, 246–257.
10. Grunfeld, Y. and Griliches, Z. (1960). *Rev. Econ. Statist.*, **42**, 1–13.
11. Ijiri, Y. (1971). *J. Amer. Statist. Ass.*, **66**, 766–782. (A broad review of aggregation for economic models and an extensive bibliography.)
12. Knox, E. G. (1964). *Appl. Statist.*, **13**, 25–29.
13. Mantel, N. (1967). *Cancer Res.*, **27**, 209–220.
14. Neyman, J. and Scott, E. L. (1958). *J. R. Statist. Soc. B*, **20**, 1–43. (The seminal paper on clustering processes.)
15. Pielou, E. C. (1977). *Mathematical Ecology*. Wiley, New York. (Describes methods for the measurement of aggregation among individuals; extensive bibliography.)
16. Ripley (1977).
17. Rowe, R. D. (1976). *Int. Econ. Rev.*, **17**, 751–757.
18. Taylor, L. R. (1971). In *Statistical Ecology*, Vol. 1, G. P. Patil et al., eds. Pennsylvania State University Press, University Park, Pa., pp. 357–377.
19. Taylor, L. R. and Taylor, R. A. J. (1977). *Nature, (Lond.)*, **265**, 415–421.
20. Theil, H. (1954). *Linear Aggregates of Economic Relations*. North-Holland, Amsterdam. (The definitive work on aggregation for economic models.)
21. Tiao, G. C. and Wei, W. S. (1976). *Biometrika*, **63**, 513–524.
22. Yule, G. U. and Kendall, M. G. (1965). *An Introduction to the Theory of Statistics*. Charles Griffin, London.

See also DISPERSION THEORY, HISTORICAL DEVELOPMENT OF; DIVERSITY INDICES; ECOLOGICAL STATISTICS; ECONOMETRICS; and OVERDISPERSION.

J. KEITH ORD

AGING FIRST-PASSAGE TIMES

First-passage times of appropriate stochastic processes* have often been used to represent times to failure of devices or systems subjected to shocks and wear, random repair times, and random interruptions during their operations. Therefore the aging properties of such first-passage times have been widely investigated in the reliability* and maintenance literature. In this entry we overview some basic results regarding first-passage times that have aging properties that are related as follows (the definitions of these notions are given later):

$$\begin{tikzcd}
\mathrm{PF}_2 \arrow[r, Rightarrow] & \mathrm{IFR} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{DMRL} \arrow[r, Rightarrow] & \mathrm{NBUE} \\
 & \mathrm{IFRA} \arrow[r, Rightarrow] & \mathrm{NBU} \arrow[r, Rightarrow] & \mathrm{NBUC} \arrow[u, Rightarrow]
\end{tikzcd}$$

Some terminology and notation that are used throughout the entry are described below. Let $\{X_n, n \geqslant 0\}$ be a discrete-time stochastic process with state space $[0, \infty)$. We assume that the process starts at 0, that is, $P\{X_0 = 0\} = 1$. For every $z \geqslant 0$ we denote by T_z the first time that the process crosses the threshold z, that is, $T_z \equiv \inf\{n \geqslant 0 : X_n \geqslant z\}(T_z = \infty \text{ if } X_n < z \text{ for all } n \geqslant 0)$. If, for example, T_z is new better than used (NBU) [increasing failure rate (IFR), increasing failure-rate average (IFRA), etc.] for any $z \geqslant 0$ (*see* HAZARD RATE AND OTHER CLASSIFICATIONS OF DISTRIBUTIONS) then the process $\{X_n, n \geqslant 0\}$ is called an NBU [IFR, IFRA, etc.] process. In a similar manner one defines an NBU [IFR, IFRA, etc.] continuous-time process. In this entry we point out many instances of NBU, IFRA, IFR, and other processes that have first-passage times with similar aging properties. We do not consider here antiaging properties such as NWU (new worse than used), DFR (decreasing failure rate), DFRA (decreasing failure-rate average), etc. Some antiaging properties of first-passage times can be found in the references.

MARKOV PROCESSES

Consider a discrete-time Markov process* $\{X_n, n \geqslant 0\}$ with the discrete state space $\mathbb{N}_+ = \{0, 1, \ldots\}$. Denote by $\boldsymbol{P} = \{p_{ij}\}_{i\in\mathbb{N}_+, j\in\mathbb{N}_+}$ the transition matrix of the process. Keilson, in a pioneering work [16], obtained many distributional properties, such as complete monotonicity, for various first-passage times of such processes. The strongest aging property that we consider here is that of log-concavity. A nonnegative random variable is said to have the *Pólya frequency of order* 2 (PF_2) property if its (discrete or continuous) probability density is log-concave. *See* PÓLYA TYPE 2 FREQUENCY (PF_2) DISTRIBUTIONS. Assaf et al. [4] have shown the following result:

$\mathbf{PF_2}$ Theorem. If the transition matrix $\boldsymbol{P}$ is totally positive of order 2 (TP_2) (that is, $p_{ij}p_{i'j'} \geqslant p_{i'j}p_{ij'}$ whenever $i \leqslant i'$ and $j \leqslant j'$), then T_z has a log-concave density for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is a PF_2 process.

They also extended this result to some continuous-time Markov processes with discrete state space. Shaked and Shanthikumar [35] extended it to continuous-time pure jump Markov processes with continuous state space. *See* also TOTAL POSITIVITY.

An aging notion that is weaker than the PF_2 property is the notion of IFR (*increasing failure rate*). A discrete nonnegative random variable T is said to have this property if $P\{T \geqslant n\}$ is log-concave on $\mathbb{N}_+$, or, equivalently, if its discrete hazard rate* function, defined by $P\{T = n\}/P\{T \geqslant n\}$, is nondecreasing on $\mathbb{N}_+$. If $\boldsymbol{P}$ is the transition matrix of the discrete-time Markov process $\{X_n, n \geqslant 0\}$, then let $\boldsymbol{Q}$ denote the matrix of left partial sums of $\mathbf{P}$. Formally, the i, jth element of $\boldsymbol{Q}$, denoted by q_{ij}, is defined by $q_{ij} = \sum_{k=0}^{j} p_{ik}$. Durham et al. [10] essentially proved the following result.

IFR Theorem. If $\boldsymbol{Q}$ is TP_2, then T_z is IFR for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is an IFR process.

This strengthens previous results of Esary et al. [13] and of Brown and Chaganty [8]. Using ideas of the latter, it can be extended to some continuous-time Markov processes with discrete state space. Shaked and Shanthikumar [35] have extended this result to continuous-time pure jump Markov processes with continuous state space; see also refs. 17, 38.

One application of the IFR result described above is given in Kijima and Nakagawa [18]. They considered a Markov process $\{X_n, n \geqslant 0\}$ defined by $X_n = bX_{n-1} + D_n, n = 1, 2, \ldots (X_0 \equiv 0)$, where $0 \leqslant b \leqslant 1$, and the D_n's are independent nonnegative random variables. Such processes arise in studies of imperfect preventive maintenance policies, where each maintenance action reduces the current damage by $100(1-b)\%$, and D_n is the total damage incurred between the $(n-1)$st and the nth maintenance action. Let G_n denote the distribution function of D_n. They showed that if $G_n(x)$ is TP_2 in n and x, and if $G_n(x)$ is log-concave in x for all n, then $\{X_n, n \geqslant 0\}$ is an IFR process.

Since the IFR property is weaker than the PF_2 property, one would expect that a condition weaker than the assumption that $\boldsymbol{P}$ is TP_2 would suffice to guarantee that $\{X_n, n \geqslant 0\}$ is an IFR process. Indeed, the assumption that $\boldsymbol{Q}$ is TP_2 is weaker than the assumption that P is TP_2.

Sometimes the time that the maximal increment of a Markov process passes a certain critical value is of interest. Thus, Li and Shaked [21] studied first-passage times of the form $T_{z,u} \equiv \inf\{n \geqslant 1 : X_n \geqslant z \text{ or } X_n - X_{n-1} \geqslant u\} [T_{z,u} = \infty \text{ if } X_n < z \text{ and } X_n - X_{n-1} < u \text{ for all } n]$. The process $\{X_n, n \geqslant 0\}$ is said to have a *convex transition kernel* if $P\{X_{n+1} > x + y | X_n = x\}$ is nondecreasing in x for all y. They have shown the following result:

Increment IFR Theorem. If $\{X_n, n \geqslant 0\}$ has monotone sample paths, and a convex transition kernel, and if $\boldsymbol{P}$ is TP_2, then $T_{z,u}$ is IFR for all z and u.

They also obtained a version of this result for some continuous-time Markov processes with discrete state space.

An aging notion that is weaker than the IFR property is the notion of IFRA (*increasing failure-rate average*). A discrete nonnegative random variable T is said to have this property if either $P\{T = 0\} = 1$ or $P\{T = 0\} = 0$ and $[P\{T > n\}]^{1/n}$ is nonincreasing in $n = 1, 2, \ldots$. The process $\{X_n, n \geqslant 0\}$ is said to be *stochastically monotone* if $P\{X_{n+1} > x | X_n = y\}$ is nondecreasing in y for every x. Equivalently, $\{X_n, n \geqslant 0\}$ is said to be stochastically monotone if q_{ij} is nonincreasing in i for all j, where $\boldsymbol{Q}$ is the matrix of left partial sums of $\boldsymbol{P}$. From a general result of Shaked and Shanthikumar [34] we obtain the following result.

IFRA Theorem. If $\{X_n, n \geqslant 0\}$ is stochastically monotone and if it has nondecreasing sample paths, then T_z is IFRA for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is an IFRA process.

This result strengthens previous results of Esary et al. [13] and of Brown and Chaganty [8]. Using ideas of the latter, it can be extended to some continuous-time Markov processes with discrete state space. Drosen [9] and Shaked and Shanthikumar [35] have extended it to continuous-time pure jump Markov processes with continuous state space. Özekici and Günlük [28] have used this result to show that some interesting Markov processes that arise in the study of some maintenance policies are IFRA.

Since the IFRA property is weaker than the IFR property, one would expect that a condition weaker than the assumption that $\boldsymbol{Q}$ is TP_2 would suffice to guarantee that $\{X_n, n \geqslant 0\}$ is an IFRA process. Indeed, the assumption that $\{X_n, n \geqslant 0\}$ is stochastically monotone is weaker than the assumption that $\boldsymbol{Q}$ is TP_2. However, for the IFRA result we need to assume that $\{X_n, n \geqslant 0\}$ has nondecreasing sample paths, whereas there is no such assumption in the IFR result.

When the time that the maximal increment of a Markov process passes a certain critical value is of interest, then one studies $T_{z,u}$. Li and Shaked [21] have shown the following result.

Increment IFRA Theorem. If $\{X_n, n \geqslant 0\}$ has monotone convex sample paths and a convex transition kernel, then $T_{z,u}$ is IFRA for all z and u.

They also obtained a version of this result for some continuous-time Markov processes with discrete state space.

An aging notion that is weaker than the IFRA property is the notion of NBU (*new better than used*). A discrete nonnegative random variable T is said to have this property if $P\{T \geqslant n\} \geqslant P\{T - m \geqslant n | T \geqslant m\}$ for all $n \geqslant 0$ and $m \geqslant 0$. Brown and Chaganty [8] proved the following result.

NBU Theorem. If $\{X_n, n \geqslant 0\}$ is stochastically monotone, then T_z is NBU for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is an NBU process.

This strengthens previous results of Esary et al. [13]. Brown and Chaganty [8] extended it to some continuous-time Markov processes with discrete state space.

Since the NBU property is weaker than the IFRA property, it is not surprising that the condition that suffices to imply that $\{X_n, n \geqslant 0\}$ is NBU is weaker than the conditions that imply that $\{X_n, n \geqslant 0\}$ is IFRA.

Marshall and Shaked [25] have identified a condition that is different than stochastic monotonicity, and that still yields that $\{X_n, n \geqslant 0\}$ is an NBU process. Explicitly, they showed that if the strong Markov process $\{X_n, n \geqslant 0\}$ starts at 0, and is free of positive jumps (that is, it can jump up only at most one unit at a time), then it is an NBU process. They obtained a similar result for continuous-time strong Markov processes with discrete or continuous state space. For example, a Wiener process which starts at 0 is an NBU process. Also, if $\{(X_1(t), X_2(t), \ldots, X_m(t)), t \geqslant 0\}$ is a Brownian motion* in $\mathbb{R}^m$, and $Y(t)$, $\equiv [\sum_{i=1}^m X_i^2(t)]^{1/2}$, then the Bessel process $\{Y(t), t \geqslant 0\}$ is NBU.

Li and Shaked [21] studied the NBU property of $T_{z,u}$, the time that the maximal increment of a Markov process passes a certain critical value, and obtained the following result.

Increment NBU Theorem. If $\{X_n, n \geqslant 0\}$ has a convex transzition kernel, then $T_{z,u}$ is NBU for all z and u.

They also obtained a version of this result for some continuous-time Markov processes with discrete state space.

An aging notion that is weaker than the NBU property is the notion of NBUC (*new better than used in convex ordering*). A nonnegative random variable T is said to have this property if $E[\phi(T)] \geqslant E[\phi(T-m)|T \geqslant m]$ for all $m \geqslant 0$ and for all nondecreasing convex functions ϕ for which the expectations are defined. If $\boldsymbol{P}$ is the transition matrix of $\{X_n, n \geqslant 0\}$, then the potential matrix of $\{X_n, n \geqslant 0\}$, which we denote by $\boldsymbol{R}$, is defined by $\boldsymbol{R} \equiv \sum_{n=0}^{\infty} \boldsymbol{P}^n$. Let $\overline{\boldsymbol{R}}$ denote the matrix of left partial sums of $\boldsymbol{R}$, that is, if r_{ij} denotes the ijth element of $\boldsymbol{R}$ and $\overline{r}_{ij}$ denotes the ijth element of $\overline{\boldsymbol{R}}$, then $r_{ij} \equiv \sum_{k=0}^{j} r_{ik}$. Also, let us define the following matrix: $\boldsymbol{R}_m \equiv \sum_{n=m}^{\infty} \boldsymbol{P}^n$, and let $\overline{\boldsymbol{R}}_m$ denote the matrix of left partial sums of $\boldsymbol{R}_m$, that is, if $r_{m,ij}$ denotes the ijth element of $\boldsymbol{R}_m$, and $\overline{r}_{m,ij}$ denotes the ijth element of $\overline{\boldsymbol{R}}_m$, then $\overline{r}_{m,ij} \equiv \sum_{k=0}^{j} r_{m,ik}$. Ocón and Pérez [27] have shown the following result.

NBUC Theorem. If $\overline{r}_{m,ij}$ is nonincreasing in i for all j and m, then T_z is NBUC for all $z \geqslant 0$; that is, $\{X_n, n \geqslant 0\}$ is an NBUC process.

In addition to the condition given in the NBUC Theorem, they also assumed that $\{X_n, n \geqslant 0\}$ has nondecreasing sample paths, but if, for a fixed z, one modifies $\{X_n, n \geqslant 0\}$ so that z is an absorbing state, then the almost sure monotonicity of the sample paths is not required for the conclusion that $\{X_n, n \geqslant 0\}$ is NBUC. They have also extended this result to some continuous-time Markov processes with discrete state space.

Since the NBUC property is weaker than the NBU property, one would expect that a condition weaker than stochastic monotonicity should suffice to imply that $\{X_n, n \geqslant 0\}$ is NBUC. Indeed, it can be shown that if $\{X_n, n \geqslant 0\}$ is stochastically monotone then $\overline{r}_{m,ij}$ is nonincreasing in i for all j and m.

An aging notion that is weaker than the NBUC property is the notion of NBUE (*new better than used in expectation*). A nonnegative random variable T is said to have this property if $E[T] \geqslant E[T-m|T \geqslant m]$ and all $m \geqslant 0$. Karasu and Özekici [15] obtained the following result.

NBUE Theorem. If $\overline{r}_{ij}$ is nonincreasing in i for all j, then T_z is NBUE for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is an NBUE process.

In addition to the condition given in the NBUE Theorem, they also assumed that $\{X_n, n \geqslant 0\}$ has nondecreasing sample paths, but, again, if for a fixed z one modifies $\{X_n, n \geqslant 0\}$ so that z is an absorbing state, then the almost sure monotonicity of the sample paths is not required for the conclusion that $\{X_n, n \geqslant 0\}$ is NBUE. They have also extended this result to some continuous-time Markov processes with discrete state space.

Since the NBUE property is weaker than the NBUC property, it is not surprising that the condition that suffices to imply that $\{X_n, n \geqslant 0\}$ is NBUE is weaker than the condition that implies that $\{X_n, n \geqslant 0\}$ is NBUC.

We close this section with an aging notion that is weaker than the IFR but stronger than the NBUE notion. A nonnegative random variable T is said to have the DMRL (*decreasing mean residual life*) property if $E[T - m|T \geqslant m]$ is nonincreasing in $m \geqslant 0$. If $\boldsymbol{P}$ is the transition matrix of the discrete-time Markov process $\{X_n, n \geqslant 0\}$, then let $\boldsymbol{Q}_m$ denote the matrix of left partial sums of $\boldsymbol{P}^m$. Denote the ijth element of $\boldsymbol{Q}_m$ by $q_{m,ij}$. Ocón and Pérez [27] have shown the following result.

DMRL Theorem. If $\bar{r}_{m,ij}/q_{m,ij}$ is nonincreasing in i for all j and m, then T_z is DMRL for all $z \geqslant 0$, that is, $\{X_n, n \geqslant 0\}$ is an DMRL process.

In addition to the condition given in the DMRL Theorem, they also assumed that $\{X_n, n \geqslant 0\}$ has nondecreasing sample paths, but, again, if for a fixed z one modifies $\{X_n, n \geqslant 0\}$ so that z is an absorbing state, then the almost sure monotonicity of the sample paths is not required for the conclusion that $\{X_n, n \geqslant 0\}$ is DMRL. They have also extended this result to some continuous-time Markov processes with discrete state space.

CUMULATIVE DAMAGE PROCESSES

Suppose that an item is subjected to shocks occurring randomly in (continuous) time according to a counting process* $\{N(t), t \geqslant 0\}$. Suppose that the ith shock causes a nonnegative random damage X_i, and that damages accumulate additively. Thus, the damage accumulated by the item at time t is $Y(t) = \sum_{i=1}^{N(t)} X_i$. We assume that the damage at time 0 is 0. The process $\{Y(t), t \geqslant 0\}$ is called a *cumulative damage* shock process*. Suppose that the item fails when the cumulative damage exceeds a threshold z.

If $\{N(t), t \geqslant 0\}$ is a Poisson process*, and if the damages X_i are independent and identically distributed and are independent of $\{N(t), t \geqslant 0\}$, then $\{Y(t), t \geqslant 0\}$ is a Markov process. The results described in the preceding section can then be used to derive aging properties of the first-passage time T_z. For example, the process $\{Y(t), t \geqslant 0\}$ clearly has monotone sample paths, and is stochastically monotone. Therefore it is IFRA. Esary et al. [13] noticed that if the X_i's are not necessarily identically distributed, but are merely stochastically increasing (that is, $P\{X_i > x\}$ is nondecreasing in i for all x), then the process $\{Y(t), t \geqslant 0\}$ is still IFRA. In fact, they identified even weaker conditions on the X_i's that ensure that $\{Y(t), t \geqslant 0\}$ is IFRA. As another example of the application of the results of the preceding section to the process $\{Y(t), t \geqslant 0\}$, suppose that the damages X_i are independent and identically distributed with a common log-concave distribution function; then the process $\{Y(t), t \geqslant 0\}$ is IFR. In fact, Esary et al. [13] and Shaked and Shanthikumar [35] have identified even weaker conditions on the X_i's that ensure that $\{Y(t), t \geqslant 0\}$ is IFR.

If $\{N(t), t \geqslant 0\}$ is a nonhomogeneous (rather than homogeneous) Poisson process, then $\{Y(t), t \geqslant 0\}$ is not a Markov process, even if the damages X_i are independent and identically distributed, and are independent of $\{N(t), t \geqslant 0\}$. Let $\Lambda(t), t \geqslant 0$, be the mean function of the process $\{N(t), t \geqslant 0\}$. From results of A-Hameed and Proschan [3] it follows that the IFRA results mentioned in the previous paragraph still hold provided Λ is star-shaped. It also follows that the IFR results mentioned in the previous paragraph still hold provided Λ is convex.

Sumita and Shanthikumar [40] have studied a cumulative damage wear process* in which $\{N(t), t \geqslant 0\}$ is a general renewal* process. Let the interarrivals of $\{N(t), t \geqslant 0\}$ be denoted by $U_i, i = 1, 2, \ldots$. Thus, $(U_i, X_i), i = 1, 2, \ldots$, are independent and identically distributed pairs of nonnegative random variables. It is not assumed that, for each i, U_i and X_i are independent. In fact, the model is of particular interest when, for each i, U_i and X_i are not independent. If we define $Y(t) \equiv \sum_{i=1}^{N(t)} X_i$, then in general $\{Y(t), t \geqslant 0\}$ is not a Markov process. They showed that if the U_i's are NBU, and if the pairs (U_i, X_i) possess some positive dependence properties, then $\{Y(t), t \geqslant 0\}$ is an NBU process. They also showed that if the U_i's are NBUE, and if the pairs (U_i, X_i) possess some other positive dependence properties, then $\{Y(t), t \geqslant 0\}$

is an NBUE process. Furthermore, they also showed that if the U_i's are HNBUE (*harmonic new better than used in expectation*) (that is, $\int_t^\infty P\{U_i > x\}\,dx \leqslant \mu \exp(-t/\mu)$ for $t \geqslant 0$, where $\mu = E[U_i]$), if the X_i's are exponential random variables, and if the pairs (U_i, X_i) possess some positive dependence properties, then $\{Y(t), t \geqslant 0\}$ is an HNBUE process. They obtained similar results for the model in which the nth interarrival depends on the $(n-1)$st jump X_{n-1} (rather than on the nth jump). In ref. 39 they considered a wear process that, at time t, is equal to $\max_{0\leqslant n\leqslant N(t)}\{X_n\}$. Again, they did not assume that, for each i, U_i and X_i are independent. they identified conditions under which this process is NBU, NBUE, or HNBUE.

In the preceding paragraphs it has been assumed that the threshold z is fixed. But in many applications it is reasonable to allow it to be random, in which case denote the threshold by Z, and the first-passage time to Z by T_Z. For this model Esary et al. [13] have shown that if Z is IFRA [respectively, NBU] then T_Z is IFRA [respectively, NBU]. They have also shown that if the identically distributed random damages X_i have the PF_2 property, and if Z has the PF_2 property, then T_Z has the PF_2 property.

A-Hameed and Proschan [3] considered a random threshold cumulative damage model in which the damages X_i are still independent, but are not necessarily identically distributed, and the counting process $\{N(t), t \geqslant 0\}$ is a nonhomogeneous Poisson process with mean function Λ. In fact, they assumed that the ith random damage has the gamma distribution* with shape parameter a_i and rate parameter b. Let $A_k \equiv \sum_{i=1}^k a_i, k = 1, 2, \ldots.$ They showed that if A_k is convex [respectively, star-shaped, superadditive] in k, if Λ is convex [respectively, star-shaped, superadditive], and if Z is IFR [respectively, IFRA, NBU]. For the special case when all the a_i's are equal, they showed that if Λ is convex [star-shaped] and if Z is DMRL [NBUE], then T_Z is DMRL [NBUE]. For this special case, Klefsjö [19] showed that if Λ is star-shaped and if Z is HNBUE, then T_Z is HNBUE. Abdel-Hameed [1,2] and Drosen [9] have extended some of these results to pure jump wear processes.

The reader is referred to Shaked [32] for more details and further references on cumulative damage processes.

NON-MARKOVIAN PROCESSES

In many applications in reliability theory*, the underlying wear process is non-Markovian. Various researchers have tried to obtain aging properties of first-passage times for some such processes. We describe some fruits of their efforts.

IFRA Closure Theorem [29]. If $\{X_i(t), t \geqslant 0\}, i = 1, 2, \ldots, n$ are independent IFRA processes, each with nondecreasing sample paths, then $\{\phi(X_1(t), X_2(t), \ldots, X_n(t)), t \geqslant 0\}$ is also an IFRA process whenever ϕ is continuous and componentwise nondecreasing.

The following result follows from refs. 11, 26.

NBU Closure Theorem. If $\{X_i(t), t \geqslant 0\}$, $i = 1, 2, \ldots, n$, are independent NBU processes, each with nondecreasing sample paths, then $\{\phi(X_1(t), X_2(t), \ldots, X_n(t)), t \geqslant 0\}$ is also an NBU process whenever ϕ is continuous and componentwise nondecreasing.

More general results are described in the next section.

Marshall and Shaked [25] and Shanthikumar [37] have identified a host of non-Markovian processes that are NBU. We will try to describe some of these processes in plain words. One class of NBU wear processes that Marshall and Shaked [25] have identified is the following, with shocks and recovery: Shocks occur according to a renewal process with NBU interarrivals. Each shock causes the wear to experience a random jump, where the jumps are independent and identically distributed and are independent of the underlying renewal process. These jumps may be negative as long as the wear stays nonnegative. Between shocks the wear changes in some deterministic manner which depends on the previous history of the process. This deterministic change may correspond to a partial recovery of the underlying device.

A second class of NBU wear processes that Marshall and Shaked [25] have identified is the following, with random repair times: The process starts at 0, and before the first shock it increases in some deterministic manner. Shocks occur according to a Poisson process. Each shock causes the wear to experience a random jump (usually a negative jump, but the process is set equal to 0 if such a jump would carry it below 0), where the jumps are independent and identically distributed, and are independent of the underlying Poisson process. Between shocks the wear increases in some deterministic manner where the rate of increase depends only on the current height of the process. This deterministic change may correspond to a continuous wear of the underlying device, and the jumps correspond to repairs that reduce the wear.

A third class of NBU wear processes that Marshall and Shaked [25] have identified is that of Gaver–Miller [14] processes. These have continuous sample paths that alternately increase and decrease in a deterministic fashion where the rate of increase or decrease depends only on the current height of the process. The random durations of increase are independent and identically distributed exponential random variables, and the random durations of decrease are independent and identically distributed NBU random variables.

Shanthikumar [37] has generalized the first two kinds of processes mentioned above. In particular, he allowed the times between jumps and the magnitude of jumps to be dependent, and he still was able to prove, under some conditions, that the resulting wear processes are NBU. His results also extend the NBU results of Sumita and Shanthikumar [40]. Lam [20] has gone even further and identified a class of stochastic processes that are even more general than those of Shanthikumar [37]. She showed that the processes in that class are NBUE. Marshall and Shaked [26] extended the NBU results that are described above to processes with state space $[0, \infty)^m$.

Semi-Markov processes* are more general than Markov processes in the sense that the sojourn time of the process in each state has a general distribution rather than being exponential. Using coupling arguments, Shanthikumar [37] was able to formulate a set of conditions under which semi-Markov processes are NBU. Lam [20] obtained conditions under which semi-Markov processes are NBUE.

For some more details on the aging properties of first-passage times of non-Markovian processes, and for further references, see the review by Shaked [32].

PROCESSES WITH STATE SPACE $\mathbb{R}^M$

Let $\{\mathbf{X}(t), t \geqslant 0\} = \{(X_1(t), X_2(t), \ldots, X_m(t)), t \geqslant 0\}$ be a stochastic process on $\mathbb{R}_+^m \equiv [0, \infty)^m$. A set $U \subseteq \mathbb{R}_+^m$ is an *upper set* if $\boldsymbol{x} \in U$ and $\boldsymbol{y} \geqslant \boldsymbol{x}$ implies that $\boldsymbol{y} \in U$. The first-passage time of the process $\{\boldsymbol{X}(t), t \geqslant 0\}$ to an upper set U is defined by $T_U \equiv \inf\{t \geqslant 0 : \boldsymbol{X}(t) \in U\}[T_U = \infty \text{ if } \boldsymbol{X}(t) \notin U \text{ for all } t \geqslant 0]$. The process $\{\boldsymbol{X}(t), t \geqslant 0\}$ is called an IFRA [NBU] process if T_U is IFRA [NBU] for all closed upper sets $U \subseteq \mathbb{R}_+^m$. Clearly, every component $\{X_i(t), t \geqslant 0\}$ of an IFRA [NBU] process $\{\boldsymbol{X}(t), t \geqslant 0\}$ is an IFRA [NBU] process on $\mathbb{R}_+$. In this section we consider only processes that start at $\mathbf{0}$. The following characterizations of IFRA and NBU processes are taken from refs. 26, 36.

IFRA and NBU Characterization Theorem

(i) The process $\{\boldsymbol{X}(t), t \geqslant 0\}$, with nondecreasing sample paths, is IFRA if, and only if, for every choice of closed upper sets $U_1, U_2, \ldots, U_n$, the random variables $T_{U_1}, T_{U_2}, \ldots, T_{U_n}$ satisfy that $\tau(T_{U_1}, T_{U_2}, \ldots, T_{U_n})$ is IFRA for every coherent life function τ. (For a definition of coherent life functions see, e.g., Esary and Marshall [12] or Barlow and Proschan [5].)

(ii) The process $\{\boldsymbol{X}(t), t \geqslant 0\}$, with nondecreasing sample paths, is NBU if, and only if, for every choice of closed upper sets $U_1, U_2, \ldots, U_n$, the random variables $T_{U_1}, T_{U_2}, \ldots, T_{U_n}$ satisfy that $\tau(T_{U_1}, T_{U_2}, \ldots, T_{U_n})$ is NBU for every coherent life function τ.

The following IFRA closure properties can be derived from results in Marshall [22].

General IFRA Closures Theorem

(i) Let $\{\boldsymbol{X}_i(t), t \geqslant 0\}$ be an IFRA process on $\mathbb{R}_+^{m_i}$ with nondecreasing sample paths, $i = 1, 2, \ldots, n$. If these n processes are independent, then $\{(\boldsymbol{X}_1(t), \boldsymbol{X}_2(t), \ldots, \boldsymbol{X}_n(t)), t \geqslant 0\}$ is an IFRA process on $\mathbb{R}_+^{\sum_{i=1}^n m_i}$.

(ii) Let $\{\boldsymbol{X}(t), t \geqslant 0\}$ be an IFRA process on $\mathbb{R}_+^{m_1}$, and let $\boldsymbol{\psi} : \mathbb{R}_+^{m_1} \to \mathbb{R}_+^{m_2}$ be a nondecreasing continuous function such that $\boldsymbol{\psi}(\mathbf{0}) = \mathbf{0}$. Then $\{\boldsymbol{\psi}(\boldsymbol{X}(t)), t \geqslant 0\}$ is an IFRA process on $\mathbb{R}_+^{m_2}$.

The following NBU closure properties can be derived from results in Marshall and Shaked [26].

General NBU Closures Theorem

(i) Let $\{\boldsymbol{X}_i(t), t \geqslant 0\}$ be an NBU process on $\mathbb{R}_+^{m_i}$ with nondecreasing sample paths, $i = 1, 2, \ldots, n$. If these n processes are independent, then $\{(\boldsymbol{X}_1(t), \boldsymbol{X}_2(t), \ldots, \boldsymbol{X}_n(t)), t \geqslant 0\}$ is an NBU process on $\mathbb{R}_+^{\sum_{i=1}^n m_i}$.

(ii) Let $\{\boldsymbol{X}(t), t \geqslant 0\}$ be an NBU process on $\mathbb{R}_+^{m_1}$, and let $\boldsymbol{\psi} : \mathbb{R}_+^{m_1} \to \mathbb{R}_+^{m_2}$ be a nondecreasing continuous function such that $\boldsymbol{\psi}(\mathbf{0}) = \mathbf{0}$. Then $\{\boldsymbol{\psi}(\boldsymbol{X}(t)), t \geqslant 0\}$ is an NBU process on $\mathbb{R}_+^{m_2}$.

A discrete-time Markov process $\{\boldsymbol{X}_n, n \geqslant 0\}$ with state space $\mathbb{R}^m$ is said to be *stochastically monotone* if $P\{\boldsymbol{X}_{n+1} \in U | \boldsymbol{X}_n = \boldsymbol{x}\}$ is nondecreasing in $\boldsymbol{x}$ for all upper sets $U \subseteq \mathbb{R}^m$. Brown and Chaganty [8] have shown that if such a stochastically monotone process, with state space $\mathbb{R}_+^m$, starts at $\mathbf{0}$, then it is an NBU process. They also showed that some continuous-time Markov processes are NBU. Brown and Chaganty [8] and Shaked and Shanthikumar [34] have shown that if such a stochastically monotone process, with state space $\mathbb{R}_+^m$, starts at $\mathbf{0}$ and has nondecreasing sample paths, then it is an IFRA process. They also showed that some continuous-time Markov processes are IFRA. (In fact, they as well as Marshall [22] and Marshall and Shaked [26], have considered processes with state spaces that are much more general than $\mathbb{R}_+^m$.)

To see an application of their IFRA results, consider the following model of Ross [30]. Suppose that shocks hit an item according to a nonhomogeneous Poisson process $\{N(t), t \geqslant 0\}$ with mean function Λ. The ith shock inflicts a nonnegative random damage X_i. The X_i's are assumed to be independent and identically distributed, and are also assumed to be independent of the underlying nonhomogeneous Poisson process. Suppose that there is a function D such that the total damage after n shocks is $D(X_1, X_2, \ldots, X_n, 0, 0, 0\ldots)$, where D is a nonnegative function whose domain is $\{(x_1, x_2, \ldots), x_i \geqslant 0, i = 1, 2, \ldots\}$. Define $Y(t) \equiv D(X_1, X_2, \ldots, X_{N(t)}, 0, 0, 0, \ldots), t \geqslant 0$. If $\Lambda(t)/t$ is nondecreasing in $t > 0$, if D is nondecreasing in each of its arguments, and if $D(x_1, x_2, \ldots, x_n, 0, 0, 0, \ldots)$ is permutation symmetric in $x_1, x_2, \ldots, x_n$, for all n, then $\{Y(t), t \geqslant 0\}$ is an IFRA process.

A function $\phi : \mathbb{R}_+^m \to \mathbb{R}_+$ is said to be *subhomogeneous* if $\alpha\phi(\boldsymbol{x}) \leqslant \phi(\alpha\boldsymbol{x})$ for all $\alpha \in [0, 1]$ and all $\boldsymbol{x}$. Note that every coherent life function τ is a nondecreasing subhomogeneous function. A vector $(S_1, S_2, \ldots, S_n)$ of nonnegative random variables is said to be MIFRA (*multivariate increasing failure-rate average*), in the sense of Block and Savits [7], if $\phi(S_1, S_2, \ldots, S_n)$ is IFRA for any nondecreasing subhomogeneous function ϕ (see Marshall and Shaked [24] for this interpretation of the MIFRA property). In a similar manner Marshall and Shaked [24] have defined the notion of MNBU (*multivariate new better than used*). According to Block and Savits [7], a stochastic process $\{\boldsymbol{X}(t), t \geqslant 0\}$ on $\mathbb{R}_+^m$ is said to be a MIFRA process if, for every finite collection of closed upper sets $U_1, U_2, \ldots, U_n$ in $\mathbb{R}_+^m$, the vector $(T_{U_1}, T_{U_2}, \ldots, T_{U_n})$ is MIFRA. Clearly, every MIFRA process is an IFRA process. Block and Savits [7] showed that there exist IFRA processes that are not MIFRA. In a similar manner one can define MNBU processes. Clearly, every MIFRA process is also an MNBU process. It may be of interest to compare the definition of MIFRA and MNBU processes to the characterizations given in the IFRA and NBU Characterization Theorem above. Some multivariate cumulative damage wear processes that are MIFRA will be described now.

Consider m items that are subjected to shocks that occur according to (one) Poisson process $\{N(t), t \geqslant 0\}$. Let X_{ij} denote the damage inflicted by the ith shock on the jth item, $i = 1, 2, \ldots, j = 1, 2, \ldots, m$. Suppose that the vectors $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{im}), i = 1, 2, \ldots$, are independent. Assume that the damages accumulate additively. Thus, the wear process $\{\boldsymbol{Y}(t), t \geqslant 0\} = \{(Y_1(t), Y_2(t), \ldots, Y_m(t)), t \geqslant 0\}$ here has state space $\mathbb{R}_+^m$, where $Y_j(t) = \sum_{i=1}^{N(t)} X_{ij}, t \geqslant 0, j = 1, 2, \ldots, m$.

IFRA Shock Model Theorem [31]. If $\boldsymbol{X}_i \leqslant_{\text{st}} \boldsymbol{X}_{i+1}$ (that is, $E[\psi(\boldsymbol{X}_i)] \leqslant E[\psi(\boldsymbol{X}_{i+1})]$ for all nondecreasing functions ψ for which the expectations are well defined), $i = 1, 2, \ldots,$ then $\{\boldsymbol{Y}(t), t \geqslant 0\}$ is an IFRA process.

Thus, if the jth item is associated with a fixed threshold z_j (that is, the item fails once the accumulated damage of item j crosses the threshold z_j), $j = 1, 2, \ldots, m$, then, by the IFRA Characterization Theorem, the vector of the lifetimes of the items $(T_{z_1}, T_{z_2}, \ldots, T_{z_m})$ satisfies that $\tau(T_{z_1}, T_{z_2}, \ldots, T_{z_m})$ is IFRA for every coherent life function τ.

MIFRA Shock Model Theorem [31]. If $\boldsymbol{X}_i =_{\text{st}} \boldsymbol{X}_{i+1}$ (that is, the $\boldsymbol{X}_i$'s are identically distributed) then $\{\boldsymbol{Y}(t), t \geqslant 0\}$ is a MIFRA process.

Thus, the vector of the lifetimes of the items $(T_{z_1}, T_{z_2}, \ldots, T_{z_m})$ is such that $\phi(T_{z_1}, T_{z_2}, \ldots, T_{z_m})$ is IFRA for every nondecreasing subhomogeneous function ϕ.

Shaked and Shanthikumar [36] have extended the above results to processes with state spaces more general than $\mathbb{R}^m$. They also showed that these results still hold if the shocks occur according to a birth process with nondecreasing birth rates (rather than a homogeneous Poisson process). Marshall and Shaked [23] obtained some multivariate NBU properties for the model described above.

For additional details regarding multivariate IFRA and NBU processes see refs. 32, 33.

REFERENCES

1. Abdel-Hameed, M. (1984). Life distribution properties of devices subject to a Lévy wear process. *Math. Oper. Res.*, **9**, 606–614.
2. Abdel-Hameed, M. (1984). Life distribution properties of devices subject to a pure jump damage process. *J. Appl. Probab.*, **21**, 816–825.
3. A-Hameed, M. S. and Proschan, F. (1973). Nonstationary shock models. *Stochastic Process. Appl.*, **1**, 383–404.
4. Assaf, D., Shaked, M., and Shanthikumar, J. G. (1985). First-passage times with PF$_r$ densities. *J. Appl. Probab.*, **22**, 185–196. (The PF$_r$ property for Markov processes with discrete state space is established in this paper.)
5. Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston. (This is a comprehensive book that discusses various life distributions and many useful probabilistic models in reliability theory.)
6. Block, H. W. and Savits, T. H. (1980). Multivariate increasing failure rate average distributions. *Ann. Probab.*, **8**, 793–801.
7. Block, H. W. and Savits, T. H. (1981). Multidimensional IFRA processes. *Ann. Probab.*, **9**, 162–166. (Processes with multivariate IFRA first-passage times are considered in this paper.)
8. Brown, M. and Chaganty, N. R. (1983). On the first passage time distribution for a class of Markov chains. *Ann. Probab.*, **11**, 1000–1008. (The IFR, IFRA, and NBU properties for Markov processes with monotone transition matrices are presented in this paper.)
9. Drosen, J. W. (1986). Pure jump models in reliability theory. *Adv. Appl. Probab.*, **18**, 423–440. (The IFR, IFRA, and NBU properties for Markov pure jump processes with continuous state space are discussed in this paper.)
10. Durham, S., Lynch, J., and Padgett, W. J. (1990). TP$_2$-orderings and the IFR property with applications. *Probab. Eng. Inform. Sci.*, **4**, 73–88. (The IFR property for Markov processes with discrete state space is established in this paper.)
11. El-Neweihi, E., Proschan, F., and Sethuraman, J. (1978). Multistate coherent systems. *J. Appl. Probab.*, **15**, 675–688.
12. Esary, J. D. and Marshall. A. W. (1970). Coherent life functions. *SIAM J. Appl. Math.*, **18**, 810–814.
13. Esary, J. D., Marshall, A. W., and Proschan, F. (1973). Shock models and wear processes. *Ann. Probab.*, **1**, 627–649. (This paper is a pioneering work on aging properties of first-passage times of cumulative damage wear processes.)
14. Gaver, D. P. and Miller, R. G. (1962). Limiting distributions for some storage problems. In *Studies in Applied Probability and Management Science*, K. J. Arrow, S. Karlin, and H. Scarf, eds. Stanford University Press, pp. 110–126.
15. Karasu, I. and Özekici, S. (1989). NBUE and NWUE properties of increasing Markov processes. *J. Appl. Probab.*, **26**, 827–834.

16. Keilson, J. (1979). *Markov Chains Models—Rarity and Exponentiality*. Springer-Verlag, New York. (Early results on the complete monotonicity of some first-passage times of Markov processes with discrete state space are described in this book.)
17. Kijima, M. (1989). Uniform monotonicity of Markov processes and its related properties. *J. Oper. Res. Soc. Japan*, **32**, 475–490.
18. Kijima, M. and Nakagawa, T. (1991). A cumulative damage shock model with imperfect preventive maintenance. *Naval Res. Logist.*, **38**, 145–156.
19. Klefsjö, B. (1981). HNBUE survival under some shock models. *Scand. J. Statist.*, **8**, 39–47.
20. Lam, C. Y. T. (1992). New better than used in expectation processes. *J. Appl. Probab.*, **29**, 116–128. (The NBUE property for semi-Markov processes and some other non-Markovian processes is given in this paper.)
21. Li, H. and Shaked, M. (1995). On the first passage times for Markov processes with monotone convex transition kernels. *Stochastic Process. Appl.*, **58**, 205–216. (The aging properties for the increment processes of some Markov processes are considered in this paper.)
22. Marshall, A. W. (1994). A system model for reliability studies. *Statist. Sinica*, **4**, 549–565.
23. Marshall, A. W. and Shaked, M. (1979). Multivariate shock models for distributions with increasing hazard rate average. *Ann. Probab.*, **7**, 343–358.
24. Marshall, A. W. and Shaked, M. (1982). A class of multivariate new better than used distributions. *Ann. Probab.*, **10**, 259–264.
25. Marshall, A. W. and Shaked, M. (1983). New better than used processes. *Adv. Appl. Probab.*, **15**, 601–615. (This paper describes a host of non-Markovian processes with NBU first-passage times.)
26. Marshall, A. W. and Shaked, M. (1986). NBU processes with general state space. *Math. Oper. Res.*, **11**, 95–109. (This paper discusses processes with multivariate NBU first-passage times.)
27. Ocón, R. P. and Pérez, M. L. G. (1994). On First-Passage Times in Increasing Markov Processes. *Technical Report*, Department of Statistics and Operations Research, University of Granada, Spain.
28. Özekici, S. and Günlük, N. O. (1992). Maintenance of a device with age-dependent exponential failures. *Naval Res. Logist.*, **39**, 699–714.
29. Ross, S. M. (1979). Multivalued state component systems. *Ann. Probab.*, **7**, 379–383.
30. Ross, S. M. (1981). Generalized Poisson shock models. *Ann. Probab.*, **9**, 896–898.
31. Savits, T. H. and Shaked, M. (1981). Shock models and the MIFRA property. *Stochastic Process. Appl.*, **11**, 273–283.
32. Shaked, M. (1984). Wear and damage processes from shock models in reliability theory. In *Reliability Theory and Models: Stochastic Failure Models, Optimal Maintenance Policies, Life Testing, and Structures*, M. S. Abdel-Hameed, E. Çinlar, and J. Queen, eds. Academic Press, pp. 43–64. (This is a survey paper on wear processes and shock models with a list of references up to 1983.)
33. Shaked, M. and Shanthikumar, J. G. (1986). IFRA processes. In *Reliability and Quality Control*, A. P. Basu, ed., Elsevier Science, pp. 345–352.
34. Shaked, M. and Shanthikumar, J. G. (1987). IFRA properties of some Markov jump processes with general state space. *Math. Oper. Res.*, **12**, 562–568. (The IFRA property is extended to Markov processes with a general state space in this paper.)
35. Shaked, M. and Shanthikumar, J. G. (1988). On the first-passage times of pure jump processes. *J. Appl. Probab.*, **25**, 501–509.
36. Shaked, M. and Shanthikumar, J. G. (1991). Shock models with MIFRA time to failure distributions. *J. Statist. Plann. Inference*, **29**, 157–169.
37. Shanthikumar, J. G. (1984). Processes with new better than used first passage times. *Adv. Appl. Probab.*, **16**, 667–686. (The NBU property for semi-Markov processes and some other non-Markovian processes is discussed in this paper.)
38. Shanthikumar, J. G. (1988). DFR property of first-passage times and its preservation under geometric compounding. *Ann. Probab.*, **16**, 397–406.
39. Shanthikumar, J. G. and Sumita, U. (1984). Distribution properties of the system failure time in a general shock model. *Adv. Appl. Probab.*, **16**, 363–377.
40. Sumita, U. and Shanthikumar, J. G. (1985). A class of correlated cumulative shock models. *Adv. Appl. Probab.*, **17**, 347–366.

See also CUMULATIVE DAMAGE MODELS; HAZARD RATE AND OTHER CLASSIFICATIONS OF DISTRIBUTIONS; JUMP PROCESSES; MARKOV PROCESSES; RELIABILITY, PROBABILISTIC; SHOCK MODELS; and WEAR PROCESSES.

M. SHAKED
H. LI

AGREEMENT ANALYSIS, BASIC

Basic agreement analysis [4] is a model-based approach to analyzing subjective categorical data*. Two or more raters place objects into K unordered, mutually exclusive, and exhaustive categories. Basic agreement analysis provides a principled measure of the amount of inter-rater agreement as well as estimates of rater bias and the true probabilities $\tau(i)$ of the categories $i, i = 1, \ldots, K$. The approach thereby controls for possible confounding effects of systematic rater bias in the analysis of subjective categorical data.

Each rater's judgment process is modeled as a mixture of two components: an error process that is unique for the rater in question, and an agreement process that operationalizes the true values of the objects to be classified. The probability $\Pr[X_r = i]$ that rater r places a randomly selected object into category i is given by

$$\Pr[X_r = i] = \lambda\tau(i) + (1-\lambda)\epsilon_r(i),$$

where $\epsilon_r(i)$ is the probability of i under Rater r's error process, and λ, the percentage of judgments governed by the agreement process, is assumed to be the same for all raters in the simplest model. The coefficient λ quantifies the amount of agreement between raters, and is closely related to the well-known kappa index of Cohen [1]; *see* KAPPA COEFFICIENT. In fact, basic agreement analysis can be considered as a systematization of Cohen's idea to correct for agreement by chance in the analysis of subjective categorical data.

For two raters,

$$\Pr[X_r = i, X_2 = j] = \Pr[X_1 = i]\Pr[X_2 = j] + \begin{cases} \lambda^2\tau(i)[1-\tau(i)] & \text{if } i = j, \\ -\lambda^2\tau(i)\tau(j) & \text{if } i \neq j; \end{cases}$$

model parameters can be estimated on the basis of cross-tabulation of the raters' judgments in *agreement matrices* [4]. For the special case of $K = 2$, see [3]. The model is a member of the class of *general processing tree models* [2].

The basic agreement model is a measurement error model that allows more focused analyses of experiments employing subjective categorical data from several raters, for whom ratings have measurement error distributions that can induce bias in the evaluation of scientific hypotheses of interest.

REFERENCES

1. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. and Psych. Meas.*, **20**, 37–46.
2. Hu, X. and Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, **59**, 21–47.
3. Klauer, K. C. (1996). Urteilerübereinstimmung für dichotome Kategoriensysteme. *Diagnostica*, **42**, 101–118.
4. Klauer, K. C. and Batchelder, W. H. (1996). Structural analysis of subjective categorical data. *Psychometrika*, **61**, 199–240.

See also AGREEMENT, MEASURES OF; CATEGORICAL DATA, SUBJECTIVE; and KAPPA COEFFICIENT.

KARL CHRISTOPH KLAUER

AGREEMENT, MEASURES OF

Measures of agreement are special cases of measures of association* or of correlation* that are designed to be sensitive not merely to deviations from independence, but specifically to deviations indicating agreement. These are measures most commonly used to assess reliability (*see* GROUP TESTING) or reproducibility of observations. In this context, it is usually assured that one has better than chance agreement. Consequently, the statistical problems of interest revolve not around the issue of testing the null hypothesis of independence but around estimation of the population measure of agreement.

Typically, one samples n subjects and has m observations on each, say $X_{i1}, X_{i2}, \ldots, X_{im}$ $(i = 1, 2, \ldots, n)$, where the marginal distributions of the observations are the same for all $j = 1, 2, \ldots, m$. Typically, a measure of agreement is zero when all ratings are independent and identically distributed, and 1.0 if $\Pr\{X_{ij} = X_{ij}{'}\} = 1$ for all i and $j \neq j'$.

Controversies as to the *validity* of certain measures of agreement can be generated if the assumption of equal marginal distributions is violated [1]. This assumption, however, imposes no major limitation. One need only randomly assign the m observations for each subject to positions $1, 2, \ldots, m$ prior to analysis.

Suppose that

$$X_{ij} = \mu + \xi_i + \epsilon_{ij},$$

where $\xi_i \sim N(0, \sigma_\xi^2), \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \rho = \sigma_\xi^2/(\sigma_\xi^2 + \sigma_\epsilon^2)$, and the "true" values of ξ_i and "errors" ϵ_{ij} are independent for all i and j. For this type of interval or ratio-level data, the intraclass correlation coefficient* is one such measure of agreement. This measure is most readily computed by applying a oneway analysis of variance* with each subject constituting a group. The intraclass correlation coefficient then is

$$r_I = \frac{F-1}{F+m-1},$$

where F is the F-statistic* to test for subject differences.

Since

$$F \sim \frac{(m-1)\rho + 1}{1-\rho} F_{n-1, n(m-1)},$$

nonnull tests and confidence intervals for the parameter ρ can be structured on this basis [2].

A nonparametric analogue of the same procedure is based on the use of coefficient of concordance*. In this case the n observations for each value of j are rank ordered $1, 2, \ldots, n$, with ties given the average of ranks that would have been assigned had there been no ties. One may then calculate the intraclass correlation coefficient based on the ranks r_S. This statistic, r_S, is the average Spearman rank correlation coefficient* between pairs of ratings and is related to the coefficient of concordance W by the relationship

$$r_S = \frac{mW-1}{m-1}.$$

The distribution of r_S is approximately of the same form as that of r_I [3].

These two measures, r_I and r_S, are designed to measure agreement for measurements taken on ordinal, interval, or ratio scales. For measurements taken on the nominal scale, the kappa coefficient* performs the same function [4]; its nonnull distribution is not known theoretically, and jack-knife* or bootstrap* methods are suggested for estimation and testing of this measure [5].

To illustrate these methods we use scores on a test of memory for 11 subjects ($n = 11$) each tested three times ($m = 3$). The three scores per subject are listed in random time order in Table 1. The intraclass correlation coefficient was found to be $r_I = 0.59$. The rank orders for each rating appear as superscripts in the table. The average Spearman coefficient based on these data was $r_S = 0.45$.

Finally, one may dichotomize (or classify) the scores in innumerable ways. For illustration we defined a "positive" test as a score of 50 or above, a "negative" test as a score below 50. The kappa coefficient for this dichotomization was $k = 0.05$.

This example illustrates an important point in evaluating magnitudes of measures of agreement. The measure of agreement reflects the nature of the population sampled (i.e., σ_ξ^2), the accuracy of the observation (i.e., σ_ϵ^2), and the nature of the observation itself (interval vs. ordinal vs. nominal). Consequently, poor measures of agreement are obtained if the observation is insensitive to the variations inherent in the population either because of an intrinsic scaling problem or because of inaccuracy of measurement.

Further, there are many more measures of agreement than these common ones proposed in the literature, because there are many ways of conceptualizing what constitutes agreement and how to measure disagreement. To see this, let $D(X_{ij}, X_{rs})$ be any metric reflecting agreement between two observations X_{ij} and X_{rs} (even if the X's are multivariate) with $D(X_{ij}, X_{rs}) = 1$ if and only if $X_{ij} \equiv X_{rs}$. If D_w is the mean of D's between all $n\binom{m}{2}$ pairs of observations within subjects, and D_t the mean of D's between all $\binom{nm}{2}$ pairs of observations, then a measure of agreement is

$$\frac{D_w - D_t}{1 - D_t}.$$

For example, when X_{ij} is a rank order vector, one might propose that $D(X_{ij}, X_{rs})$ be

Table 1.

	Rating		
Subject	1	2	3
1	44^{4}	48^{7}	54^{5}
2	57^{8}	45^{6}	$40^{2.5}$
3	$46^{5.5}$	32^{3}	58^{7}
4	66^{9}	50^{8}	55^{6}
5	$46^{5.5}$	43^{4}	50^{4}
6	72^{11}	83^{11}	70^{11}
7	35^{3}	28^{2}	$64^{9.5}$
8	67^{10}	66^{10}	63^{8}
9	26^{2}	44^{5}	$64^{9.5}$
10	16^{1}	21^{1}	19^{1}
11	48^{7}	69^{9}	$40^{2.5}$

a rank correlation coefficient between such vectors. Such a measure has been proposed as an extension of kappa [5] for multiple choices of categories, or as a measure of intergroup rank concordance [6].

As a result, the magnitude of a measure of agreement is determined not only by the nature of the population, the accuracy of the observation, and the nature of the observation itself, but, finally, by the metric of agreement between observations selected as the basis of the measure of agreement. *See also* SIMILARITY, DISSIMILARITY AND DISTANCE, MEASURES OF.

REFERENCES

1. Bartko, J. J. (1976). *Psychol. Bull.*, **83**, 762–765.
2. Bartko, J. J. (1966). *Psychol. Rep.*, **19**, 3–11.
3. Fleiss, J. L. and Cuzick, J. (1979). *Appl. Psychol. Meas.*, **3**, 537–542.
4. Kraemer, H. C. (1981). *Biometrika*, **68**, 641–646.
5. Kraemer, H. C. (1980). *Biometrics*, **36**, 207–216.
6. Kraemer, H. C. (1976). *J. Amer. Statist. Ass.*, **71**, 608–613.

See also ASSOCIATION, MEASURES OF; CONCORDANCE, COEFFICIENT OF; INTRACLASS CORRELATION COEFFICIENT; KAPPA COEFFICIENT; PSYCHOLOGICAL TESTING THEORY; SIMILARITY, DISSIMILARITY AND DISTANCE, MEASURES OF; SPEARMAN RANK CORRELATION COEFFICIENT; and VARIANCE COMPONENTS.

HELENA CHMURA KRAEMER

AGRICULTURAL SURVEYS

The purpose of this paper is to describe the statistical methodologies that have been developed over time that are somewhat unique to the field of agriculture. Agricultural Statisticians throughout history have been innovative in the use of available statistical theory and adapting it to their needs.

From a statistical point of view, the characteristics of agricultural surveys are similar to surveys associated with other social or physical phenomena. First, one must define the population and then prepare a frame that represents it. Then, unless one is to do a complete census*, sampling theory is used to select a sample from the frame. Estimators that reflect the sampling design are developed and include methodologies to impute for missing data (*see* IMPUTATION) and to estimate for nonresponse*. The statistics literature is rich with theory and methods to support these issues.

Agriculture has several features that pose methodological problems. One first needs to consider the definition of agriculture. Agriculture involves the use of land, the culture or raising of a living organism through different life cycles, and the concept of ownership. For example, the use of land, culture, and ownership as criteria separate fishing from aquaculture and tree farming from forestry. This definition easily points to land as the population. For example, agricultural land is simply all land used for the culture or raising of plants or animals under some form of private ownership. One of the first difficulties is that it is also necessary to describe agriculture as a business, with the need to provide measures of the economic or demographic situation of farms and the people operating the farms. This suggests that the sampling frame* consists of a list of farms or farm operators that account for the land associated with agriculture.

The need to account for both land and the farm operators has dominated the statistical thinking of agricultural statisticians and this led to the development of multiple frame sampling that will be described in greater detail.

The other unique feature of agriculture is that it provides the food and fiber that feeds and clothes a country's population. Food security becomes an issue, which means that timely and accurate forecasts of future supplies are needed for policy purposes and to ensure that markets operate efficiently. Agriculture is seasonal—especially crop production—which means that all of the supply becomes available in a short period of time and is held in storage and used until the next harvest period.

This paper will provide an overview of the history of the development of statistical procedures to forecast and measure agricultural production. The examples will mostly apply to the United States but, where appropriate, methods used in other countries will also be presented. The material will draw heavily from Reference 20, which traces the history of agricultural statistics as developed by the United States Department of Agriculture (USDA).

BEFORE PROBABILITY SAMPLING

The basic method to obtain agricultural statistics in many developing and developed countries was to use the administrative levels of government. The basic unit was an administrative unit such as a village or small agricultural region. The village or local area administrators provided subjective measures of areas planted and production and other agricultural measures. No sampling was involved as village totals were sent up the administrative line. This practice is still widespread in many developing countries, for several reasons. First, their farmers may not be educated enough to understand the concept of area or production or are unwilling to answer questions. Cost is also an issue. However, these administrative data are subject to manipulation by the various levels of government through which they pass, which means that these countries are facing the need to modernize their methodology. The administrative system used in China is typical of procedures used in many countries [21].

Basic statistics in the United States were initially provided by periodic censuses of agriculture, with surveys of voluntary reporters in intervening years to measure change from the census benchmark.

Prior to the 1880 agricultural census in the United States, only information about total crop production and livestock inventories was obtained. The 1880 census also obtained information about crop acreages. These census enumerations of acreage provided benchmarks for estimating crop acreages for years between census years. This was the beginning of the procedure that is still used to forecast and estimate crop production. Basically, it calculates crop production as the product of the two separate estimates of acreage and yield per acre. Once planted, crop averages usually do not change very much between planting and harvest. There is also less year-to-year variability between crop acres than there is between yield per acre. In general, the estimates through the nineteenth century were linked to the decennial Census of Agriculture conducted by the Bureau of the Census*. The USDA relied upon correspondents reporting their assessment of year-to-year changes in their locality to make the annual estimates. As might be suspected, small year-to-year biases in the measures of change linked to a census could grow to a widening gap over the years between the USDA estimates and the next census benchmark level. This problem led to the development of improved methodology.

The most important statistics produced were and still are the forecasts of the production of crops such as wheat, corn, soybeans, and cotton, followed by end-of-season estimates of actual production. For reasons given above, the forecasts and estimates of production were determined by separately estimating or forecasting acreage planted and average yields per acre. There was no "sampling frame" of farms; there were only lists of correspondents who would voluntarily respond to a mailed inquiry.

In the absence of probability sampling theory, much effort went into improving estimating procedures to measure crop acreages and to forecast crop yields. These procedures are discussed below in chronological order and are described more thoroughly in Reference 3.

Starting in 1912, the *Par Method* was adopted to translate farmer reported crop condition values early in the crop season into a probable yield per acre that would be realized at harvest. The par method to forecast yield (y) consisted of the following components:

$\overline{y} = CY_m/C_m,$ where

$C_m =$ the previous 10-yr average condition for the given month,

$Y_m =$ the previous 10-yr average yield per acre realized at the end of the season,

$C =$ current condition for the given month.

The forecasting model was simply a line passing through the origin and the point (C, $\overline{y}$). A separate par yield ($\overline{y}$) was established for each state, crop, and month. In actual practice, subjective modification of the means was considered necessary to remove the effects of atypical conditions. For example, a drought that may occur only once every 10 or 15 year would greatly affect the 10-yr average conditions and yield. To aid in these adjustments, 5- and 10-yr moving averages* were computed to identify unusual situations or trends, and if necessary, exclude the atypical observations.

The development of simple graphic solutions prior to the use of regression and correlation theory was a major breakthrough as a practical means to forecast crop yields, and this approach was implemented in the late 1920s. Data for a sufficient number of years had accumulated, so that final end-of-season estimates of yields could be plotted against averages of condition reports from farmers for each crop in each State.

The condition was reflected on the x-axis and end-of-season yields on the y-axis. The plotted points depicted a linear description of early season crop conditions compared to end-of-season yields. The yield forecast for the current system was obtained by entering the current condition on the chart and finding the best fit with historic yields subjectively.

Graphical regression techniques provided a consistent method to translate survey data into estimates, which in effect adjusted for persistent bias in the data caused by the purposive sampling procedures. This method quickly replaced the par method and was adopted rapidly.

The following discussion describes early attempts to estimate the acreage to be harvested.

Because the ideas of probability sampling had not yet been formed, the procedures used to estimate acres for harvest were more difficult than those to estimate average yields. The USDA used its state offices to enlarge the lists of farm operators, but there was no complete list of farms that could be used for survey purposes. Therefore, the estimating procedures relied upon establishing a base from the most recent Census of Agriculture and estimating the percent change from year to year. As the country developed with a road system and a postal service, a common data collection* approach was the Rural Carrier Survey. The postal carriers would drop a survey form in each mailbox along their route. A common procedure during that time was to include two columns in the questionnaire when asking the farmer questions about acreage planted to each crop. During the current survey, the farmer was asked to report the number of acres planted this year and the number of acres planted the previous year in each crop. This method was subject to several reporting biases, including memory bias, and it led to matching "identical" reports from year to year to remove the memory bias. The matching of identical reports did improve the estimates, but was considerably more labor intensive because the name matching had to be done by hand. The process was also complicated by problems with operations changing in size, and it was inherently biased because it did not account for new entrants to agriculture.

This methodology was subject to a potentially serious bias, caused by the selective or purposive nature of the sample. In an effort to make an allowance for this bias, a relative indicator of acreage was developed in 1922; it became known as the *ratio relative* and contained the following components:

$R_1 =$ Ratio of the acreage of a given crop to the acreage of all land in farms (or crops) for the current year as reported by the sample of farm operators;

$R_2 =$ Same as R_1 but for the previous year;

$$\hat{y} = (R_1/R_2) \times \text{(Estimated total area of the crop the previous year)}.$$

The ratio R_1/R_2 was essentially a measure of change in the crop area from the previous year. The assumption was that the previous year's area was known without error. The belief was that this ratio held the bias resulting from the purposive sampling constant from one year to the next. A reported limitation was the extreme variability in the acreage ratios between the sample units. This was countered by increasing the "number" of farms surveyed and weighting the results by size of farm.

By 1928, matched farming units that reported in both years were used to compute the ratio relative. This reduced the influence of the variability between sample units. When looking back at the variance of the estimate from a current perspective, one may examine the components (also assuming probability sampling).

$$\text{Var}(\bar{y}) = \text{Var}(R_1) + \text{Var}(R_2) - 2\ \text{cov}(R_{1,}R_2).$$

This shows why the use of matching reports improved the ratio relative estimator. However, this did not solve the problem, because by using matching reports, farms going into or out of production of a particular crop were not properly represented. Therefore, statisticians continued their efforts to develop a more objective method of gathering and summarizing survey data.

Some statisticians in the early 1920s would travel a defined route on the rural roads or via railway routes and record the number of telephone or telegraph poles opposite fields planted to each crop. The relative change in the pole count for each crop from year to year provided a measure of the average change in crop acreage. This method was generally unsatisfactory, because large portions of the United States still did not have telephone service; the pole count method was therefore not widely used.

A more refined method of estimating acreage was developed in the mid-1920s. A *crop meter* was developed and attached to an automobile speedometer to measure the linear frontage of crops along a specified route. The same routes were covered each year. This made possible a direct comparison of the number of feet in various crops along identical routes from the current year and the previous year. The properties of the estimator are described in Reference 12.

THE TWENTIETH CENTURY AFTER PROBABILITY SAMPLING

A milestone in the evolution of statistical methodology for agriculture was the development of the master sample of agriculture [13,14]. This was a cooperative project involving Iowa State University, the US Department of Agriculture, and the US Bureau of the Census. This area-sampling* frame demonstrated the advantages of probability sampling. The entire landmass of the United States was subdivided into area-sampling units using maps and aerial photographs. The sampling units had identifiable boundaries for enumeration purposes. The area-sampling frame had several features that were extremely powerful for agricultural surveys.

By design, it was complete in that every acre of land had a known probability of being selected. Using rules of association to be described presently, crops and livestock associated with the land could also be measured with known probabilities. The Master Sample of Agriculture was based on a stratified design—the strata defined to reflect the frequency of occurrence of farmsteads. Area-sampling units varied in size in different areas of the country to roughly equalize the number of farm households in each area-sampling unit.

The master sample was used for many probability surveys, but not on a recurring basis, because of added costs arising from the area samples having to be enumerated in person. The panel surveys of farm operators, while not selected using probability theory, were very much cheaper to conduct, because the collection was done by mail. It was not until 1961 that pressures to improve the precision of the official estimates resulted in the US Congress appropriating funds for a national level area frame survey on an annual recurring basis. During the early 1960s, the Master Sample of Agriculture was being replaced by a new area frame that was stratified via land use categories on the basis of the intensity of cultivation of crops. This methodology is still used. The process of developing the area frame is now much more sophisticated, relying upon satellite imagery and computer-aided stratification [18]. The method of area frame sampling described here is generally referred to as *Area Frame of Segments*.

An alternative is to select an *Area Frame of Points*. The usual practice is to divide the landmass into grids, and within each grid selecting a sample of points. Then, data collection involves identifying what is on each point or identifying the farm associated with the point; Reference 9 describes a sample design starting with 1800 × 1800 m grids. Each grid contained a total of 36 points—each point having a 3 × 3 m dimension. A random sample of 3 points was selected from each grid.

This method of sampling is easier to implement than segment sampling and is being used in Europe. However, data collected from point samples are less suitable for matching with satellite images or data. Data coming from segments are based on a drawing of the fields on aerial photographs. The use of the area frame led to the development of some new estimators, described below.

The sampling unit for the area sample frame is a segment of land—usually identified on an aerial photograph for enumeration. The segment size generally ranged from 0.5 to 2 sq mi, depending upon the availability of suitable boundaries for enumeration and the density of the farms. The basic area frame estimator was the design-based unbiased estimate of the total,

$$y_a = \sum_h \sum_i e_{hi} \bullet y_{hi}$$

where y'_{hi} was the ith segment total for an item in the hth stratum and e_{hi} was the reciprocal of the probability of selecting the ith segment in the hth stratum.

During the frame development process, the segment boundaries are determined without knowledge of farm or field boundaries. Therefore, an early (and continuing) difficulty was how to associate farms with sample segments during data collection. Three methods have evolved, which are referred to both as methods of association* and as estimators. Let y_{hil} be the value of the survey item on the lth farm having all or a portion of its land in the ith sample segment. Then different estimators arise, depending on how survey items on farms are associated with the sample segments. We present these next.

Farm (Open):

The criterion for determining whether a farm is in a sample or not is whether its headquarters are located within the boundaries of the sample segment. This estimator was most practicable when farm operations were generally homogeneous, that is, they produced a wide variety of items, some of which may not have appeared in the segment. This estimator was also useful for items such as number of hired workers and of animals born that are difficult to associate with a parcel of land. The extreme variation in size of farms and the complex rules needed to determine if the farm headquarters were in the segment resulted in large sampling and nonsampling errors,

$$y'_{hi} = \sum_l F_{hil} y_{hil},$$

where $F_{hil} = 1$, if the operator of farm l lives in the segment; 0, otherwise.

Tract (Closed):

The tract estimator is based on a rigorous accounting of all land, livestock, crops, and so on, within the segment boundaries, regardless of what part of a farm may be located within the boundaries of the segment. The method offered a significant reduction in both sample and nonsampling errors over the farm method, because reported acreages could be verified by map or photograph. The estimator is robust in that the maximum amount that can be reported for a segment is limited by its size. The estimator is especially useful for measuring acres in specific crops,

$$y'_{hi} = \sum_l T_{hil} y_{hil},$$

where

$$T_{hil} = \frac{\text{Amount of item on farm } l \text{ in segment } i}{\text{Total amount of item on farm } l}.$$

Weighted:

The difficulty with the tract estimate was that some types of information, such as economic, could only be reported on a whole-farm basis. This led to the development of the weighted procedure, in which data are obtained on a whole-farm basis, for each farm with a portion of its land inside a sample segment. The whole farm data are prorated to the segment on the basis of the proportion of each farm's land that is inside the segment. This estimator provides the advantage of a smaller sampling error than either the farm or tract procedures. On the minus side, data collection costs increased 15 to 20% because of increased interviewing times, and intractable nonsampling errors are associated with determining the weights. This estimator is also used to estimate livestock inventories, number of farm workers, and production expenditures,

$$y'_{hi} = \sum_{l} W_{hil} y_{hil},$$

where

$$W_{hil} = \frac{\text{Acres of farm } l \text{ in segment } i}{\text{Total acres in farm } l}.$$

Ratio:

The area frame sample was designed so that 50 to 80% of the segments were in the sample from year to year. This allowed the computation of the usual ratio estimators* such as the year-to-year matched segment ratios of change.

While the area frame sampling and estimating procedures were being refined, this period also saw a rapid change in the structure of agriculture. Farms became more specialized and much larger. This introduced more variability that required much larger area frame sample sizes.

The proportion of farms having livestock was decreasing rapidly during this period. The variation in numbers of livestock on such farms also had increased dramatically.

The combination of these two factors meant that either resources for an extremely large area frame sample would be needed or alternative sampling frames were needed. In the early 1960s, H.O. Hartley* at Iowa State University was approached about this problem. The result was his 1962 paper laying out the basic theory of multiple frame sampling and estimation, and summarized in Reference 6. This was followed by Reference 5, which more fully developed the concepts of multiple frame sampling and estimation methodology and which also developed multiple frame estimators with reduced variances.

As implied by its name, multiple frame sampling involves the use of two or more sampling frames. If there are two frames, there are three possible poststrata or domains—sample units belonging only to frame A, sample units belonging only to frame B, and finally the domain containing sample units belonging to both frames A and B. As pointed out by Hartley, the sampling and estimation theory to be used depended on knowing in advance of sampling whether the domain and frame sizes were known. This determined whether theories applying to poststratification or domain estimation were to be used.

In the agricultural situation, the area-sampling frame provided 100% coverage of the farm population. There was also a partial list of farms, which could be stratified by size or item characteristic before sampling. Domain membership and sizes are unknown prior to sampling, thus sample allocation is by frame and domain estimation theories apply. The theory requires that after sampling, it is necessary to separate the sampled units into their proper domain. This meant area sample units had to be divided into two domains—farms not on the list, and farms on the list.

By definition, all farms represented by the list were also in the area frame. The Hartley estimator for this situation was

$$\hat{Y}_H = N_a(\overline{y}_a + P\overline{y}'_{ab}) + N_b Q \overline{y}''_{ab},$$

where $\overline{y}_a$ represents area sample units not in the list, $\overline{y}'_{ab}$ represents area sample units overlapping the list frame, $\overline{y}''_{ab}$ represents the list frame, and $P + Q = 1$.

The weights P and Q were to be determined to minimize var $(\hat{Y}_H)$. This sampling and estimation theory was used for surveys to measure farm labor numbers and wage rates, livestock inventories, and farm production expenditure costs. Because of the considerable variation in the sizes of farms and the sampling efficiencies that occurred from the stratification in the list frame, the majority of the weight went to the list frame portion of the estimator; that is, P was small and Q was large.

An alternative estimator is suggested in Reference 10. With it, units in the list frame that are in the area frame sample are screened out of the area frame portion of the survey. In other words, $P = 0$ and

$$\hat{Y}_H = N_a\overline{y}_a + N_b\overline{y}''_{ab}.$$

Additional analysis [5] suggested that, for a fixed cost, the screening estimator would have the lower variance whenever the cost of sampling from the list frame is less than the difference between the cost of sampling from the area frame and the cost of screening the area frame sample to identify those also in the list frame.

For those reasons, the screening estimator is used exclusively. The increased use of telephone enumeration for the list sample reflects personal to telephone enumeration cost ratios of 1 to 15 in some cases. The area frame sample is surveyed in its entirety in June each year. Farms that overlap the list frame are screened out and the area domain representing the list incompleteness is defined. During the next 12-month period, a series of multiple frame quarterly surveys are conducted to measure livestock inventories, crop acreages and production, and grain in storage. Other multiple frame surveys during the year cover farm labor and production expenditures. Each survey relies upon the multiple frame-screening estimator.

Multiple frame is still the dominant sampling methodology used in the United States. Its use has also spread to many countries [23]. The methods used in other countries differ only by their choice of area frame sampling, that is, point sampling versus square segments.

As the structure of agriculture became widely diverse in the United States, the basic use of stratification of the list frame of farms and estimators such as direct and ratio estimators were becoming increasingly inefficient.

The sampling methods now used to select farms from the list frame are described in Reference 15. This procedure was named *Multiple Probability Proportional to Size** (Multiple PPS), because in effect multiple samples are selected. The frame is optimized for a particular characteristic, and a PPS sample is selected using a measure of size representing the characteristic. This process can be repeated multiple times for each variable of interest to be included in the combined sample.

The next step is to combine the samples into one overall sample and recalculate the sample weights.

The probability of selecting the ith farm is $m = \max(\pi_{i1}, \ldots, \pi_{iM})$, where from 1 to M samples have been selected, and the individual probabilities of selecting a given farm from each sample is noted. The maximum probability of selecting each farm from across the individual samples becomes the farm's sample weight.

The next step is to calibrate the sample weights so that the final estimation becomes model unbiased.

The use of Poisson Permanent Random Number (PPRN) sampling, as described in Reference 17, is used to select the PPS samples.

In these designs, every population is assigned a permanent random number between 0 and 1. Unit l is selected if its random number is less than its maximum probability of being selected.

PPRN sampling furthermore allows one to think of a sample drawn with inclusion probabilities as the union of M PPRN samples, each drawn using the same permanent random number and individual probabilities. This is convenient when one is interested in estimates of different combinations of target variables in different surveys.

For example, the USDA makes estimates for potatoes in June and December, row crops (e.g., soybeans and corn) in March, June, and December, and small grains (e.g., wheat and barley) in March, July, September, and December. It wants to contact the same farms throughout the year, but has little interest in sampling a farm for the September survey if it has not historically had small grains. Thus, PPRN samples of farms using the same permanent random number can be drawn for potatoes, row crops, and small grains, each with its own selection probabilities. The union of all three is the overall sample in June. Similarly, the union of the row-crops and small-grains samples is the overall sample in March. The use of Poisson sampling is discussed in greater detail in Reference 2.

The domain determination has been the most difficult operational aspect to tackle in developing, implementing, and using multiple frame methodology [19]. As the structure of farms becomes more complicated with complex corporate and partnership arrangements, the survey procedures require a substantial effort to minimize nonsampling errors associated with domain determination.

Since the first crop report was issued in 1863, the early season forecasts of crop production continued to be some of the most critical and market sensitive information prepared by the USDA. The development of probability sampling theory and the area-sampling frame provided a foundation upon which to replace judgement-based estimates of locality conditions to forecast yields per acre. In 1954, research was initiated to develop forecasting techniques based on objective counts and measurements that would be independent of judgment-based estimates. The use of nonrepresentative samples of farmers continued to be used to report conditions in their locality and individual farms during this period, however.

Research on the use of corn and cotton objective methods began in 1954 followed by work on wheat and soybeans in 1955 and sorghum in 1958. Early results showed that a crop-cutting survey at harvest time based on a probability sample of fields would provide estimates of yield per acre with good precision. Countries such as India and China have also used this methodology, estimating final yields and production with pioneering work described in Reference 16. The methods used in China are described in Reference 21. There were two difficulties when attempting to forecast yields. One difficulty is to forecast yield before the crop is mature, and it is even more difficult to do so before the plants have set fruit.

A two-step sampling procedure is used. First, a sample of fields is selected from those identified during the annual area frame survey as having the crop of interest. Self-weighting samples are selected. Observations within fields are made in two randomly located plots with each selected field. Selected plots for most crops include two adjacent rows of predetermined length. The probable yield per acre is a function of the number of plants, the number of fruits per plant, and the size or weight of the fruit. Early in the crop season, the number of plants is used to forecast the number of fruits, with historical averages used for fruit weights. After fruit is present, several measurements are obtained to project final fruit weight. For example, the length and diameter of corn ears are obtained from ears within the sample plots. When the crop is mature, the sample plots are harvested, and the fruit counted and weighed for the final yield estimate. The early season counts and measurements from within the sample plots are combined with the data from the harvested fruit, and become part of a database that is used to develop forecasting models in subsequent years. After the farmer harvests the sample field, another set of sample plots is located and grain left on the ground is gleaned and sent to a laboratory where it is weighed and used to measure harvest loss. During the forecast season, historical averages are used to estimate harvest losses.

Simple linear and multiple regression models are used to describe past relationships between the prediction variables and the final observations at maturity. Typically, early season counts and end-of-season harvest weights and counts from within each unit are used. They are first screened statistically for outlier* and leverage* points [4]. Once these atypical data are identified and removed, the remaining data are used to create current forecast equations.

The basic forecast models for all crops are essentially the same, in that they consist of three components: the number of fruits, average fruit weight, and harvest loss.

The net yield per acre as forecast for each sample plot is computed as follows:

$$y_i = (F_i C_i W_i) - L_i,$$

where

F_i = Number of fruit harvested or forecast to be harvested in the i^{th} sample plot,
C_i = Conversion factor using the row space measurement to inflate the plot counts to a per acre basis,
W_i = Average weight of fruit harvested or forecast to be harvested,
L_i = Harvest loss as measured from postharvest gleanings (the historical average is used during the forecast, season),
Yield forecasts = $\sum_i (y_i/n)$ for the n sample fields.

Separate models are used to forecast the number of fruits (F_i) to be harvested and the final weights (W_i). The variables used in each model vary over the season, depending upon the growth stage at the time of each survey. At the end of the crop season, F_i and W_i are actual counts and weights of fruit for harvest.

The major contributor to forecast error is the difficulty of forecasting fruit weight early in the season. Many factors such as planting date, soil moisture, and temperatures at pollination time crucially affect a plant's potential to produce fruit. While the fruit can be counted early in the season, the plant does not always display characteristics that provide an indication of final fruit weight. While each plant's potential to produce fruit is affected by previous circumstances, that information is locked inside the plant—often until fruit maturity.

Over the years, the USDA has conducted extensive research to improve the basic yield forecast models. Examples of this work appear in Reference 1. Models using weather data were continuously being developed and compared against the traditional objective yield models, but always fell short. The plant measurements reflected the impacts of weather, and the use of weather data does not add to the precision. Another effort involved an attempt to model the plant growth and to use these models, known as *plant process models*, for yield forecasting. They did not prove to be feasible in a sample survey environment.

USE OF SATELLITE IMAGERY AND DATA IN AGRICULTURAL STATISTICS

The first satellite designed to monitor land use was the land observatory (Landsat) satellite launched in 1972. Several generations of Landsat satellites have since been launched and placed in orbit, for example in 1999. The satellites are designed to travel in almost perfectly circular, near-polar orbit passes over the sunlit side of the planet several times daily. The orbit shifts westward so that every part of the surface of the earth is imaged every 16 days. The satellite contains a sensor referred to as the Thematic Mapper (TM). This camera-like device divides the images into picture elements (pixels) and measures the brightness of each pixel in seven portions of the electronic spectrum. The TM scanner pixel size is 30 m sq, for which there are measures of light reflectance for seven bands of the electromagnetic spectrum. The French government launched the SPOT satellite in 1986, which contains a sensor that provides a 20-m resolution. The sensor is pointable, which allows the satellite to observe the same area on the ground several days in a row. Data from both LANDSAT and SPOT satellites are available in either photographic or digital form. The net result is large amounts of data about the land and the vegetation it carries.

The set of measurements for each pixel, its signature, can be used to separate crop areas by type or by different types of land use. It is only to the degree that the spectral signatures for different crops and land uses can be separated that satellite data become useful.

It soon became evident that the accuracy of the crop and land use classifications as derived from the satellite data would be greatly improved by using ground truth data. The methodology developed to use data from the area frame surveys as ground truth to improve the accuracy of the classification of pixels by land use or crop cover is described in Reference 11. This process first involves obtaining ground-to-pixel registration. Discriminant functions are developed from pixels matched to ground truth data. The discriminant functions are then used to classify all elements in a satellite scene. In other words, every pixel is assigned to a target crop or land use. Regression estimators are used to estimate the population parameters.

This complete classification of pixels by crop or land use in effect provides complete coverage of a given land area. A popular product is the set of cropland data layers prepared for entire states. Since each pixel is georeferenced, these cropland data in a geographic information system can be linked to transportation corridors, watershed boundaries, or any other georeferenced data.

The National Oceanic and Atmospheric Administration (NOAA) has launched a series of weather satellites that also carry an imaging system, the Advanced Very High Resolution Radiometer (AVHRR). This imaging has a pixel size of 1.1 km versus the 30-m pixel size of the TM. A key result, however, is that the entire globe can be imaged daily instead of once every 16 days. The AVHRR images provide both data and images on vegetation conditions [22]. The daily images or weekly composites are used by governmental agencies and marketing boards to monitor crop conditions over large areas to make decisions for agricultural marketing and for early warning of food aid requirements.

Satellite data classified by land use categories are used extensively to design and prepare area-sampling frames [7]. The satellite data are also spatially referenced using latitude/longitude coordinates. Therefore, they can be used along with mapping products showing natural boundaries such as roads and rivers.

THE FUTURE

The future holds many challenges. There is a growing need to understand and measure agriculture's affect on the environment. There is a related need for policy makers and others to know how their decisions about trade and the environment affect the production decisions made by farmers and their resulting economic situation. There is a growing concern about the demographics of farm operations as they shrink in number with the increasing size of farms. The interrelationship between these variables will need to be measured, pointing to an integration of sampling methods with the use of satellite data. Statistical theory will pave the way.

REFERENCES

1. Arkin, G. F., Vanderlip, R. L., and Ritchie, J. T. (1976). A dynamic Grain sorghum growth model. *Trans. Am. Soc. Agricult. Engrs.*, **19**(4), 622–630.
2. Bailey, J. and Kott, P. (1997). An application of multiple frame sampling for multipurpose surveys, *Proceedings of American Statistical Association*.
3. Becker, J. A. and Harlan, C. L. (1939). Developments in the crop and livestock reporting service since 1920. *J. Farm Econ.*, **21**, 799–827.
4. Beckman, R. J. and Cook, R. D. (1983). Outliers. *Technometrics*, **25**, 119–149.
5. Cochran, R. S. (1965). *Theory and Application of Multiple Frame Sampling.* Ph.D. Dissertation, Iowa State University, Ames, Iowa.
6. Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. John Wiley, New York.
7. Cotter, J. and Tomczac, G., (1994). An image analysis system to develop area sampling frames for agricultural surveys. *Photogrammetr. Eng. Remote Sens.*, **60**(3), 299–306.
8. Food and Agricultural Organization of the United Nations. (1998). *Multiple Frame Agricultural Surveys.* FAO Statistical Development Series, Vol. 2.
9. Fuller, W. A. and Burnmeister, L. F. (1972). "Estimators for Samples Selected from Two Overlapping Frames". *Proceedings of the Social Statistics Section*. American Statistical Association, pp. 245–249.
10. Gallego, F. J. (1995). Sampling Frames of Square Segments. Report EUR 16317, Office for Official Publications of the European Communities, Luxembourg, ISBN 92-827-5106-6.
11. Hartley, H. O. (1962). "Multiple Frame Surveys". *Proceedings of the Social Statistics Section*. American Statistical Association, pp. 203–206.

12. Hanuschak, G. R., Sigman, R., Craig, M., Ozga, M., Luebbe, R., Cook, P., Kleweno, D., and Miller, C. (1979). *Obtaining Timely Crop Area Estimates Using Ground-Gathered and Landsat Data.* Technical Bulletin No. 1609, USDA, Washington, D.C.
13. Hendricks, W. A. (1942). *Theoretical Aspects of the Use of the Crop Meter.* Agricultural Marketing Service, USDA.
14. King, A. J. and Jessen, R. J. (1945). The master sample of agriculture. *J. Am. Stat. Assoc.*, **4D**, 38–56.
15. King, A. J. and Simpson, G. D. (1940). New developments in agricultural sampling. *J. Farm Econ.*, **22**, 341–349.
16. Kott, P. S. and Bailey, J .T. (2000). The theory and practice of maximal Brewer selection with Poisson PRN sampling. *International Conference on Establishment Surveys II*, Buffalo, New York.
17. Mahalanobis, P. C. (1946). Sample surveys of crop yields in India. *Sankhya*, 269–280.
18. Ohlsson, E. (1995). "Coordination of Samples Using Permanent Random Numbers". *Business Survey Methods*. Wiley, New York, pp. 153–169.
19. Tortora, R. and Hanusachak, G., (1988). Agricultural surveys and technology. *Proceedings of 1998 American Statistical Association Meeting*.
20. Vogel, F. A. (1975). "Surveys with Overlapping Frames—Problems in Application". *Proceedings of the Social Statistics Section*. American Statistical Association, pp. 694–699.
21. Vogel, F. A. (1995). The evolution and development of agricultural statistics at the United States Department of Agriculture. *J. Off. Stat.*, **11**(2), 161–180.
22. Vogel, F. A. (1999). *Review of Chinese Crop Production Forecasting and Estimation Methodology.* U.S. Department of Agriculture Miscellaneous Publication No. 1556.
23. Wade, G., Mueller, R., Cook, P., and Doraiswamy, T. (1994). AVHRR map products for crop condition assessment: a geographic system approach. *Photogrammetr. Eng. Remote Sens.*, **60**(9), 1145–1150.

See also AGRICULTURE, STATISTICS IN and CROP AREA ESTIMATION, LANDSAT DATA ANALYSIS IN.

FREDERIC A. VOGEL (Retired)

AGRICULTURE, STATISTICS IN

The area covered by this topic is so vast that whole volumes have been written on it, e.g., ref. 8—and that is only an introduction. The use of statistical techniques in agricultural research goes back many years, and indeed agriculture was one of the areas in which modern analytical techniques were first devised. The interchange of ideas between statistical and agricultural science has been of mutual benefit to both subjects. This continues to the present day, and all that is done here is to point out particular topics of joint importance to the two sciences.

HISTORY

The earliest paper describing what may be thought of as a statistically designed agricultural experiment appears to be that of Cretté de Palluel [4]. This concerned an experiment on the fattening of sheep in which 16 animals, four each of four different breeds, were fed on four diets, one of each breed per diet. The animals were killed at four monthly intervals so that the experiment could be regarded, in modern terms, either as a $\frac{1}{4}$ replicate* of a 4^3 factorial* or a 4×4 Latin square*. This experiment, which antedates the founding of modern agricultural research stations by more than half a century, shows in a simple form the principles of good experimental design* and analysis.

Agricultural journals have been in existence in their present form since the early years of this century, and many now have statisticians on their editorial boards. Thus the *Journal of Agricultural Science* has been published in Cambridge since 1905 and deals with many branches of agriculture and animal husbandry. The *Journal of Agricultural Research* was founded in Washington in 1913 and changed to *Agronomy Journal* in 1949, reflecting its prime concern with crops. *Tropical Agriculture* has been published since 1924 in Trinidad, and the *Indian Journal of Agricultural Science* since 1931 in New Delhi. These two deal primarily with tropical agriculture, as does *Experimental Agriculture*, which started in 1930 in Oxford as the *Empire Journal of Experimental Agriculture* and dropped its imperial connections in 1965. All these journals have a long and honorable history of statistical writing, from early papers on the methodology of the analysis of field data in the *Journal of Agricultural Science* in the 1920s to several papers on the techniques of intercropping trials in *Experimental Agriculture* in the late 1970s.

Courses on statistical methods applied to agriculture have been taught for many years, one of the first being those by G. W. Snedecor* at Iowa State College as early as 1915. However, the first statistician appointed to work at an agricultural research station was R. A. Fisher*, who went to Rothamsted Experimental Station in 1919. Within a few years Fisher had developed the technique of analysis of variance* for use in analyzing the results of agricultural experiments; he was also quick to emphasize the importance of replication and randomization in field trials and introduced the randomized block design*. A good summary of Fisher's early work is given by Yates [12].

PRESENT POSITION

From the 1930s onward, statistical methods for agricultural use have been greatly extended, both by the introduction of new techniques and by their use in agricultural research throughout the world. Thus, at Rothamsted, Fisher's colleague and successor F. Yates introduced more complex experimental designs. Among others, Yates recognized the importance of extensive experimentation: Crowther and Yates [5] gave a comprehensive summary of fertilizer trials in northern Europe from 1900 to that time. Yates also used statistical methods in surveys of agricultural practice, from 1944 onward [14]. Again, these statistical techniques were initially employed in agronomy and crop husbandry, but similar principles were soon applied to experiments with animals, despite their often greater expense and difficulty. A comprehensive statement of the part statistics, and statisticians, can play in planning field experiments was given by Finney [7], and the position since then has changed only in detail, not in broad outline.

METHODS OF EXPERIMENTAL DESIGN AND ANALYSIS

The main techniques used in practice for design and analysis of agricultural experiments continue to be based largely on Fisherian principles. Thus, since all agricultural work is subject to biological variability, treatments in comparative experiments are replicated in space, and sometimes in time also. Further, the application of any treatment to a particular set of plants or animals, or piece of ground, is usually randomized, possibly with some restrictions, although systematic designs* are sometimes used for particular purposes. These same principles are used, to a lesser degree, in the design of surveys, the random element occurring in the selection of units to be sampled.

The most commonly used experimental design for field trials is the randomized block* design, in which the area of land available for experimentation is divided into *blocks*, within which it is hoped that soil conditions are reasonably uniform; the blocks are subdivided into *plots* to which treatments* are applied. (The names "block" and "plot," now widely used in experimental design, reflect the agricultural context in which they were first applied.) There are three main lines of development of practical designs, in the directions of factorial experimentation*, incomplete block designs*, and row and column designs*. Full details are given in the relevant articles elsewhere, but there are whole books devoted to the topic of experimental design, e.g., Cochran and Cox [2] for ways of allocating treatments to plots and Cox [3] for other aspects of the planning of practical experiments.

The standard technique for analyzing the results of agricultural experiments is the analysis of variance*. Although this has its critics and is certainly not universally applicable, it remains the usual method for assessing whether the variation among a group of treatments is greater than would occur if all the observed effects were due to chance. However, this technique occupies only the middle range of the examination of experimental results: it is first necessary to summarize observed data to see whether they have any meaning at all, and it is frequently desirable to synthesize the observed results into more formal models, which may advance agricultural theory as well as practice.

Since it is common to take many records on an agricultural crop or animal, the first task is to sort out those on which to conduct a formal statistical analysis. For example, if a crop is harvested over a long period of time (e.g., tomatoes or coffee), does one wish to analyze total yield, or early yield, or indeed the proportion of the total yield in a specified time? Again, there may be derived variables of interest: in experiments with animals, it could be the digestibility of the feed or the butterfat content of the milk. In pest and disease control trials it is often far more important to determine the damage on a crop than to assess the total yield, damaged and undamaged together. All these preliminaries are a vital part of the statistical assessment of a trial*; also, noting apparently anomalous values may help to pinpoint errors in recording, or alternatively, lead to the discovery of quite unsuspected effects.

Formal statistical analysis is not always necessary when the main purpose of a trial is just to obtain preliminary information for use in a further trial, for example at an early stage in a plant breeding project. However, it is common to conduct analyses of variance on trial results, if only to provide an assessment of residual variation* after allowing for treatment effects*. Some trials have treatments that are quantitative in nature, and the technique of regression* as well as analysis of variance will be useful at this formal stage. With two variables, judicious use of analysis of covariance* permits the effect of one variable on another to be assessed and allowed for. When, as is common, many variables have been recorded, multivariate methods of analysis (*see* MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)) may be used as an alternative to the separate analysis of each record.

Although analysis of variance and its derivatives are undoubtedly the methods most commonly used for data analysis, they are not the only ones; many other techniques may be used to supplement or replace them. Thus an important area for studying experimental techniques is the investigation of best plot sizes. An early study here was that by Fairfield Smith [6] of the relation between plot size and variability. Subsequent work has shown that it is also often necessary to take account of possible variation due to individual plants as well as the environment, while there are many nonstatistical factors that have to be considered in practice. Studies of animal breeding trials and components of variance (*see* VARIANCE COMPONENTS) have proceeded together ever since the work of Henderson [9] dealing with a nonorthogonal set of data on dairy cows. Many agricultural experiments are now conducted to provide data for testing a mathematical model, and there are biologically important models that do not fall conveniently into the linear form suitable for analysis-of-variance techniques. One example among many is the set of models describing the relations between crop yield and plant density, work on which is conveniently summarized by Willey and Health [11]. There is now much interest in plant disease epidemiology, and although the earlier theoretical work, both biological and mathematical, was not relevant to practical agriculture, some of the more recent studies are, e.g., ref. 1. Finally, the design and analysis of series of trials often present problems different in kind from those for a single trial: for trials of crop varieties, references range in time from 1938 [13] to 1980 [10].

APPLICATION AREAS

There is now scarcely an agricultural experimental station anywhere in the world that does not use statistical techniques of the types outlined here; indeed, many have their own statisticians. This is true not only of the United States and the United Kingdom, where these methods started, but of other countries in the English-speaking world. The language barrier has proved no impediment, and striking advances have been made in many European countries, including the Netherlands, East Germany, and Poland. Further, the methods, although originating largely in the more developed countries with a temperate climate, have been used in tropical developing countries, such as India, Israel, and others in Asia, together with those in Africa and Latin America.

Experiments on many crops now use statistical methods; these include a wide range of temperate cereals, fruit, vegetables, and forage crops, and an even wider range of tropical cereals and plantation crops. Experiments in the area of animal husbandry and disease control also use statistical techniques (although the methods used on large and expensive long-lived animals cannot be identical with those on short-term annual crops). Surveys using statistical methods have been conducted on an equally wide range of temperate and tropical practices in agriculture and animal husbandry. Indeed, the use of statistical methods now permeates the whole of research and development in agriculture and related disciplines throughout the world.

REFERENCES

1. Butt, D. J. and Royle, D. J. (1974). In *Epidemics of Plant Diseases*, J. Kranz, ed., pp. 78–114.
2. Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd ed. Wiley, New York.
3. Cox, D. R. (1958). *Planning of Experiments.* Wiley, New York.
4. Cretté de Palluel (1788; English version by A. Young, 1790). *Ann. Agric.*, **14**, 133–139.
5. Crowther, E. M. and Yates, F. (1941). *Emp. J. Exper. Agric.*, **9**, 77–97.
6. Fairfield Smith, H. (1938). *J. Agric. Sci. Camb.*, **28**, 1–23.
7. Finney, D. J. (1956). *J. R. Statist. Soc. A*, **119**, 1–27.
8. Finney, D. J. (1972). *An Introduction to Statistical Science in Agriculture*, 4th ed. Blackwell, Oxford.
9. Henderson, C. R. (1953). Biometrics, **9**, 226–252.
10. Patterson, H. D. and Silvey, V. (1980). *J. R. Statist. Soc. A*, **143**, 219–240.
11. Willey, R. W. and Heath, S. B. (1969). *Adv. Agron.*, **21**, 281–321.
12. Yates, F. (1964). *Biometrics*, **20**, 307–321.
13. Yates, F. and Cochran, W. G. (1938). *J. Agric. Sci. Camb.*, **28**, 556–580.
14. Yates, F., Boyd, D. A., and Mathison, I. (1944). *Emp. J. Exp. Agric.*, **12**, 164–176.

See also ANALYSIS OF COVARIANCE; ANALYSIS OF VARIANCE; and FISHER, RONALD AYLMER.

G. H. FREEMAN

AIC. See MODEL SELECTION: AKAIKE'S INFORMATION CRITERION

AIDS STOCHASTIC MODELS

AIDS is an infectious disease caused by a retrovirus called human immunodeficiency virus (HIV) [17]. The first AIDS case was diagnosed in Los Angeles, CA, USA in 1980 [14]. In a very short period, the AIDS epidemic has grown into dangerous proportions. For example, the World Health Organization (WHO) and the United Nation AIDS Program (UNAIDS) estimated that 5 million had acquired HIV in 2003, and about 40 million people are currently living with AIDS. To control AIDS, in the past 10 years, significant advances had been made in treating AIDS patients by antiviral drugs through cocktail treatment protocol [10]. However, it is still far from cure and the disease is still spreading, especially in Africa and Asia. For preventing the spread of HIV, for controlling AIDS, and for understanding the HIV epidemic, mathematical models that take into account the dynamic of the HIV epidemic and the HIV biology are definitely needed. From this perspective, many mathematical models have been developed [1,8,11,22,24–26,29,42,55]. Most of these models are deterministic models in which the state variables (i.e., the numbers of susceptible people, HIV-infected people, and AIDS cases) are assumed as deterministic functions, ignoring completely the random nature of these variables. Because the HIV epidemic is basically a stochastic process, many stochastic models have been developed [5–7,18,33–37,39–44,46–49,53–54,57–64, 66–68,73–76]. This is necessary because as shown by Isham [23], Mode et al. [40,41], Tan [48,54], Tan and Tang [62], and Tan et al. [63], in some cases, the difference between the mean numbers of the stochastic models and the results of deterministic models could be very substantial; it follows that in these cases, the deterministic models would provide very poor approximation to the corresponding mean numbers of the stochastic models, leading to misleading and sometimes confusing results.

STOCHASTIC TRANSMISSION MODELS OF HIV EPIDEMIC IN POPULATIONS

Stochastic transmission models for the spread of HIV in populations were first developed by Mode et al. [40,41] and by Tan and Hsu [59,60] in homosexual populations; see also References 5–7, 18, 47–49, 54, 57, 66–68, 73, and 74. These models have been extended to IV drug populations [19,53,62] and to heterosexual populations [39,46,75–76]. Many of these models have been summarized in books by Tan [55] and by Mode and Sleeman [42]. Some applications of these models have been illustrated in Reference 55.

To illustrate the basic procedures for deriving stochastic transmission models for the HIV epidemic, consider a large population at risk for AIDS. This population may be a population of homosexual men, or a population of IV drug users, or a population of single males, single females, and married couples, or a mixture of these populations. In the presence of HIV epidemic, then there are three types of people in the population: S people (susceptible people), I people (infective people), and A people (AIDS patients). S people are healthy people but can contract HIV to become I people through sexual contact and/or IV drug contact with I people or A people or through contact with HIV-contaminated blood. I people are people who have contracted HIV and can pass the HIV to S people through sexual contact or IV drug contact with S people. According to the 1993 AIDS case definition [15] by the Center of Disease Control (CDC) at Atlanta, GA, USA, an I person will be classified as a clinical AIDS patient (A person) when this person develops at least one of the AIDS symptoms specified in Reference 15 and/or when his/her CD4$^{(+)}$ T-cell counts fall below 200/mm^3. Then, in this population, one is dealing with a high-dimensional stochastic process involving the numbers of S people, I people and AIDS cases.

To develop realistic stochastic models for this process, it is necessary to incorporate many important risk variables and social and behavior factors into the model, and to account for the dynamics of the epidemic process. Important risk variables that have significant impacts on the HIV epidemic are age, race, sex, and sexual (or IV drug use) activity levels defined by the average number of sexual (or IV drug use) partners per unit time; the important social and behavior factors are the IV drug use, the mixing patterns between partners, and the condom use that may reduce the probability of transmission of the HIV viruses. To account for these important risk variables and for IV drug use, the population is further stratified into subpopulations.

Given that the population has been stratified into sub populations by many risk factors, the stochastic modeling procedures of the HIV epidemic essentially boils down to two steps. The first step involves modeling the transmission of HIV from HIV carriers to susceptible people. This step would transform S people to I people ($S \longrightarrow I$). This step is the dynamic part of the HIV epidemic process and is influenced significantly by age, race, social behavior, sexual level, and many other factors. This step is referred to as the transmission step. The next step is the modeling of HIV progression until the development of clinical AIDS and death in people who have already contracted HIV. This step is basically the step transforming I people into A people ($I \longrightarrow A$) and is influenced significantly by the genetic makeup of the individual and by the person's infection duration that is defined by the time period elapsed since he/she first contracts the HIV. This step is referred to as the HIV progression step. By using these two steps, the numbers of S people, I people, and AIDS cases are generated stochastically at any time given the numbers at the previous time. These models have been referred by Tan and his associates [48–49,54,57,61,63,66–68,73–76] as chain multinomial models since the principle of random sampling dictates that aside from the distributions of recruitment and immigration, the conditional probability distributions of the numbers of S people, I people, and AIDS cases at any time, given the numbers at the previous time, are related to multinomial distributions.

THE TRANSMISSION STEP: THE PROBABILITY OF HIV TRANSMISSION

The major task in this step is to construct the probabilities of HIV transmission from infective people or AIDS cases to S people by taking into account the dynamic aspects of the HIV epidemic. These probabilities are functions of the mixing pattern that describes how people from different risk groups mix together and the probabilities of transmission of HIV from HIV carriers to susceptible people given contacts between these people. Let $S_i(t)$ and $I_i(u,t)(u = 0, \ldots, t)$ denote the numbers of S people and I people with infection duration u at time t in the ith risk group respectively. Let the time unit be a month and denoted by $p_i(t; S)$, the conditional probability that an S person in the ith risk group contracts HIV during the tth month given $\{S(t), I(u,t), u = 0, 1, \ldots, t\}$. Let $\rho_{ij}(t)$ denote the probability that a person in the ith risk group selects a person in the jth risk group as a partner at the tth month and $\alpha_{ij}(u,t)$ the probability of HIV transmission from an infective person with infection duration u in the jth risk group to the susceptible person in the ith risk group given contacts between them during the tth month. Assume that because of the awareness of AIDS, there are no contacts between S people and AIDS cases and that there are n risk groups or subpopulations. Then $p_i(t; S)$ is given by:

$$p_i(t; S) = 1 - \{1 - \psi_i(t)\}^{X_i(t)} \tag{1}$$

where $X_i(t)$ is the number of partners of the S person in the ith risk group during the tth month and $\psi_i(t) = \sum_{j=1}^{n} \rho_{ij}(t)\{I_j(t)/T_j(t)\}\overline{\alpha}_{ij}(t)$ with $T_j(t) = S_j(t) + \sum_{u=0}^{t} I_j(u,t)$ and $\overline{\alpha}_{ij}(t) = \frac{1}{I_j(t)} \sum_{u=0}^{t} I_j(u,t)\alpha_{ij}(u,t)$ with $I_j(t) = \sum_{u=0}^{t} I_j$ (u,t).

If the $\alpha_{ij}(u,t)$ are small, then $\{1 - \psi_i(t)\}^{X_i(t)} \cong \{1 - X_i(t)\psi_i(t)\}$ so that

$$\begin{aligned} p_i(t; S) &\cong X_i(t)\psi_i(t) \\ &= X_i(t) \sum_{j=1}^{n} \rho_{i,j}(t) \frac{I_j(t)}{T_j(t)} \overline{\alpha}_{ij}(t). \end{aligned} \tag{2}$$

Notice that in Equations 1 and 2, the $p_i(t; S)$ are functions of $\{S_i(t), I_i(u,t), u = 0, 1, \ldots, t\}$ and hence in general are random variables. However, some computer simulation studies by Tan and Byer [57] have indicated that if the $S_i(t)$ are very large, one may practically assume $p_i(t; S)$ as deterministic functions of time t and the HIV dynamic.

THE PROGRESSION STEP: THE PROGRESSION OF HIV IN HIV-INFECTED INDIVIDUALS

The progression of HIV inside the human body involves interactions between CD4$^{(+)}T$ cells, CD8$^{(+)}T$ cells, free HIV, HIV antibodies, and other elements in the immune system, which will eventually lead to the development of AIDS symptoms as time increases. It is influenced significantly by the dynamics of the interactions between different types of cells and HIV in the immune system, treatment by antiviral drugs, and other risk factors that affect the speed of HIV progression. Thus, it is expected that the progression of I to A depends not only on the calendar time t but also on the infection duration u of the I people as well as the genetic makeup of the I people. This implies that the transition rate of $I \longrightarrow A$ at time t for I people with infection duration u is in general a function of both u and t; this rate will be denoted by $\gamma(u,t)$. Let T_{inc} denote the time from HIV infection to the onset of AIDS. Given $\gamma(u,t)$, the probability distribution of T_{inc} can readily be derived [55, chapter. 4]. In the AIDS literature, T_{inc} has been referred to as the HIV incubation period and the probability distribution of T_{inc} the HIV incubation distribution. These distributions have been derived by Bachetti [4], Longini et al. [33], and by Tan and his associates [50–53,57,61,64–65] under various conditions. These probability distributions together with many other distributions, which have been used in the literature have been tabulated and summarized in Reference 55, chapter 4.

STOCHASTIC MODELING OF HIV TRANSMISSION BY STATISTICAL APPROACH

To develop stochastic models of HIV transmission, it is necessary to take into account the dynamic of the HIV epidemic to construct the probabilities $p_S(i,t) = \beta_i(t)\Delta t$ of HIV transmission. To avoid the dynamic aspect, statisticians assume $p_i(t; S)$ as deterministic functions of i and t and proceed to estimate these probabilities. This is a nonparametric procedure, which has ignored all information about the dynamics of the HIV epidemic; on the other hand, it has minimized the misclassification or misspecification of the dynamic of the HIV epidemic. In the literature, these probabilities are referred to as the infection incidence.

To illustrate the statistical approach, let T_I denote the time to infection of S people and $f_i(t)$ the probability density of T_I in the ith risk group. Then, $f_i(t) = \beta_i(t) exp\{-\int_0^t \beta_i(x)dx\}$. The $f_i(t)$ have been referred by statisticians as the HIV infection distributions.

To model the HIV progression, let T_A denote the time to AIDS of S people and $h_i(t)$ the probability density of T_A in the ith risk group. Then $T_A = T_I + T_{inc}$, where T_{inc} is the HIV incubation period. If T_I and T_{inc} are independently distributed of each other, then

$$h_i(t) = \int_0^t f_i(x) g_i(x, t) dx, \tag{3}$$

where $g_i(s, t)$ is the density of the HIV incubation distribution given HIV infection at time s in the ith risk group. Notice that since the transition rates of the infective stages are usually independent of the risk group, $g_i(s, t) = g(s, t)$ are independent of i. In what follows, it is thus assumed that $g_i(s, t) = g(s, t)$ unless otherwise stated.

Let ω_i denote the proportion of the ith risk group in the population. For an individual taken randomly from the population, the density of T_A is given by:

$$\begin{aligned} h(t) &= \sum_{i=1}^{n} w_i h_i(t) = \int_0^t \left\{ \sum_{i=1}^{n} w_i f_i(x) g_i(x, t) \right\} dx \\ &= \int_0^t f(x) g(x, t) dx, \end{aligned} \tag{4}$$

where $\sum_{i=1}^{n} \omega_i f_i(t) = f(t)$ is the density of the HIV infection distribution for people taken randomly from the population.

Equation 4 is the basic equation for the backcalculation method. By using this equation and by interchanging the order of summation and integration, it can easily be shown that the probability that a S person at time 0 taken randomly from the population will become an AIDS case for the first time during $(t_{j-1}, t_j]$ is

$$\begin{aligned} P(t_{j-1}, t_j) &= \int_{t_{j-1}}^{t_j} \int_0^t f(u) g(u, t)\, du dt \\ &= \left\{ \int_0^{t_j} - \int_0^{t_{j-1}} \right\} \int_0^t f(u) g(u, t)\, du dt \\ &= \int_0^{t_j} f(u) \left\{ G(u, t_j) - G(u, t_{j-1}) \right\}, du, \end{aligned} \tag{5}$$

where $G(u, t) = \int_u^t g(u, x)\, dx$ is the cumulative distribution function (cdf) of the HIV incubation period, given HIV infection at time u.

Equation 5 is the basic formula by means of which statisticians tried to estimate the HIV infection or the HIV incubation based on AIDS incidence data [8,55]. This has been illustrated in detail in Reference 8 and in Reference 55. There are two major difficulties in this approach, however. One difficulty is that the problem is not identifiable in the sense that one cannot estimate simultaneously the HIV infection distribution and the HIV incubation distribution. Thus, one has to assume the HIV incubation distribution as known if one wants to estimate the HIV infection distribution; similarly, one has to assume the HIV infection distribution as known if one wants to estimate the HIV incubation distribution. Another difficulty is that one has to assume that there are no immigration, no competing death, and no other disturbing factors for Equation 5 to hold; see Reference 55, chapter 5. These difficulties can readily be resolved by introducing state space models; see References 55, 56, 73, and 74.

STOCHASTIC TRANSMISSION MODELS OF HIV EPIDEMIC IN HOMOSEXUAL POPULATIONS

In the United States, Canada, Australia, New Zealand, and Western European countries, AIDS cases have been found predominantly amongst the homosexual, bisexual, and intravenous drug-user community with only a small percentage of cases being due to heterosexual contact. Thus, most of the stochastic models for HIV spread were first developed in homosexual populations [5–7, 40–41,47,49,57,59–60,62,66–68,73–74].

To illustrate how to develop stochastic models for the HIV epidemic, consider a large homosexual population at risk for AIDS that has been stratified into n risk groups of different sexual activity levels. Denote by $A_i(t)$ the number of new AIDS cases developed during the tth month in the ith risk group and let the time unit be a month. Then, one is entertaining a high-dimensional discrete-time stochastic process $\underset{\sim}{U}(t) = \{S_i(t), I_i(u,t), u = 0, \ldots, t, A_i(t), i = 1, \ldots, n\}$. To derive basic results for this process, a convenient approach is by way of stochastic equations. This is the approach proposed by Tan and his associates in modeling the HIV epidemic [49,54–57,63–64,66–68,73–76].

THE STOCHASTIC DIFFERENCE EQUATIONS AND THE CHAIN MULTINOMIAL MODEL

To develop the stochastic model for the above process, let $\{\Lambda_i(S,t), \mu_i(S,t)\}$ and $\{\Lambda_i(u,t), \mu_i(u,t)\}$ denote the recruitment and immigration rate, and the death and migration rate of S people and $I(u)$ people at the tth month in the ith risk group respectively. Then, given the probability $p_i(t; S)$ that the S person in the ith risk group would contract HIV during the tth month, one may readily obtain $\underset{\sim}{U}(t+1)$ from $\underset{\sim}{U}(t)$ by using multinomial distributions, for $t = 0, 1, \ldots,$. This procedure provides stochastic difference equations for the numbers of S people, $I(u)$ people, and the number of new A people at time t. These models have been referred to by Tan [48,49,54,55] and by Tan and his coworkers [57,61,63,64, 66–68,73,74] as chain multinomial models.

To illustrate, let $R_i(S,t), F_i(S,t)$, and $D_i(S,t)$ denote respectively the number of recruitment and immigrants of S people, the number of $S \to I(0)$ and the total number of death of S people during the tth month in the ith risk group. Similarly, for $u = 0, 1, \ldots, t$, let $R_i(u,t)$, $F_i(u,t)$ and $D_i(u,t)$ denote respectively the number of recruitment and immigrants of $I(u)$ people, the number of $I(u) \to A$ and the total number of death of $I(u)$ people during the tth month in the ith risk group. Then, the conditional probability distribution of $\{F_i(S,t), D_i(S,t)\}$ given $S_i(t)$ is multinomial with parameters $\{S_i(t), p_i(t; S), \mu_i(S,t)\}$ for all $i = 1, \ldots, n$; similarly, the conditional probability distribution of $\{F_i(u,t), D_i(u,t)\}$ given $I_i(u,t)$ is multinomial with parameters $\{I_i(u,t), \gamma_i(u,t), \mu_i(u,t)\}$, independently of $\{F_i(S,t), D_i(S,t)\}$ for all $\{i = 1, \ldots, n, u = 0, 1, \ldots, t\}$. Assume that $E\{R_i(S,t)|S(t)\} = S_i(t)\Lambda_i(S,t)$ and $E\{R_i(u,t)|I_i(u,t)\} = I_i(u,t)\Lambda_i(u,t)$. Then, one has the following stochastic difference equations for $\{S_i(t), I_i(u,t), i = 1, \ldots, n\}$:

$$\begin{aligned} S_i(t+1) &= S_i(t) + R_i(S,t) \\ &\quad - F_i(S,t) - D_i(S,t) \\ &= S_i(t)\{1 + \Lambda_i(S;t) - p_i(t;S) \\ &\quad - \mu_i(S,t)\} + \epsilon_i(S,t+1), \quad (6) \end{aligned}$$

$$\begin{aligned} I_i(0,t+1) &= F_i(S,t) = S_i(t)p_i(t;S) \\ &\quad + \epsilon_i(0,t+1), \quad (7) \end{aligned}$$

$$\begin{aligned} I_i(u+1,t+1) &= I_i(u,t) + R_i(u,t) \\ &\quad - F_i(u,t) - D_i(u,t) \\ &= I_i(u,t)\{1 + \Lambda_i(u,t) \\ &\quad - \gamma_i(u,t) - \mu_i(S,t)\} \\ &\quad + \epsilon_i(u+1,t+1), \quad (8) \end{aligned}$$

$$\begin{aligned} A_i(t+1) &= \sum_{u=0}^{t} F_i(u,t) \\ &= \sum_{u=0}^{t} I_i(u,t)\gamma_i(u,t) + \epsilon_i(A,t). \quad (9) \end{aligned}$$

In Equations 6 to 9, the random noises $\{\epsilon_i(S,t), \epsilon_i(u,t), u = 0, 1, \ldots, t, \epsilon_i(A,t)\}$ are derived by subtracting the conditional mean numbers from the corresponding random variables. It can easily be shown that these random noises have expected values 0 and are uncorrelated with the state variables.

Using the above stochastic equations, one can readily study the stochastic behaviors of the HIV epidemic in homosexual populations and assess effects of various risk factors on the HIV epidemic and on some intervention procedures. Such attempts have been made by Tan and Hsu [59,60], Tan [48,54], and Tan, Tang and Lee [63] by using some simplified models. For example, Tan and Hsu [59–60] have shown that the effects of intervention by decreasing sexual contact rates depend heavily on the initial number of infected people; when the initial number is small, say 10, then the effect is quite significant. On the other hand, when the initial number is large, say 10000, then the effect of decreasing sexual contact rates is very small. The Monte Carlo studies by Tan and Hsu [59,60] have also revealed some effects of "regression on the mean" in the sense that the variances are linear functions of the expected numbers. Thus, although the variances are much larger than their respective mean numbers, effects of risk factors on the variance curves are quite similar to those of these risk factors on the mean numbers.

By using the above equations, one may also assess the usefulness of deterministic models in which $S_i(t), I_i(r,t), r = 1, \cdots, k, A_i(t)$ are assumed as deterministic functions of i and t, ignoring completely randomness of the HIV epidemic process. The system of equations defining the deterministic models is derived by ignoring the random noises from Equations 6 to 9. Thus, one may assume that the deterministic models are special cases of the stochastic models. However, the above equations for $S_i(t)$ and $I_i(0,t)$ are not the same as the equations for the mean numbers of $S_i(t)$ and $I_i(0,t)$ respectively. This follows from the observation that since the $p_S(i,t)$ are functions of $S_i(t)$ and $I_i(u,t), u = 1, \ldots, k, E[S_i(t)p_S(i,t)] \neq E[S_i(t)] \times E[p_S(i,t)]$. As shown in Reference 55, the equations for the mean numbers of $S_i(t)$ and $I_i(0,t)$ differ from the corresponding ones of the deterministic model in that the equations for the means of $S_i(t)$ and $I_i(0,t)$ contain additional terms involving covariances $Cov\{S_i(t), p_S(i,t)\}$ between $S_i(t)$ and $p_S(i,t)$. Thus, unless these covariances are negligible, results of the deterministic models of the HIV epidemic would in general be very different from the corresponding mean numbers of the stochastic models of the HIV epidemic. As shown by Isham [23], Mode et al. [40,41], Tan [48,54], Tan and Tang [62], and Tan, Tang and Lee [63], the difference between results of deterministic models and the mean numbers of the stochastic models could be very substantial. The general picture appears to be that the stochastic variation would in general speed up the HIV epidemic. Further, the numbers of I people and A people computed by the deterministic model would underestimate the true numbers in the short run but overestimate the true numbers in the long run. These results imply that, in some cases, results of the deterministic model may lead to misleading and confusing results.

THE PROBABILITY DISTRIBUTION OF THE STATE VARIABLES

Let $\boldsymbol{X} = \{\boldsymbol{X}(0), \ldots, \boldsymbol{X}(t_M)\}$ with $\boldsymbol{X}(t) = \{S_i(t), I_i(u,t), u = 0, 1, \ldots, t, i = 1, \ldots, n\}$. To estimate the unknown parameters and to assess stochastic behaviors of the HIV epidemic, it is of considerable interest to derive the probability density $P(\boldsymbol{X}|\Theta)$ of $\boldsymbol{X}$. By using multinomial distributions for $\{R_i(S,t), F_i(S,T)\}$ and for $\{R_i(u,t), F_i(u,T)\}$ as above, this probability density can readily be derived. Indeed, denoting by $g_i, s\{j; t|\boldsymbol{X}(t)\}$ and $g_{i,u}\{j; t|\boldsymbol{X}(t)\}$ the conditional densities of $R_i(S,t)$ given $\boldsymbol{X}(t)$ and of $R_i(u,t)$ given $BX(t)$ respectively, one has

$$\begin{aligned} P(\boldsymbol{X}|\Theta) &= P\{\boldsymbol{X}(0)|\Theta\} \prod_{t=1}^{t_M} P\{\boldsymbol{X}(t)|\boldsymbol{X}(t-1)\} \\ &= P\{\boldsymbol{X}(0)|\Theta\} \prod_{t=1}^{t_M} \prod_{i=1}^{n} P\{S_i(t)|\boldsymbol{X}(t-1)\} \\ &\quad \times \left\{ \prod_{u=0}^{t} P[I_i(u,t)|\boldsymbol{X}(t-1)] \right\}. \qquad (10) \end{aligned}$$

In Equation 10, the $P\{S_i(t+1)|\boldsymbol{X}(t)\}$ and the $P\{I_i(u+1,t+1)|\boldsymbol{X}(t)\}$ are given respectively by

$$\begin{aligned} &P\{S_i(t+1)|\boldsymbol{X}(t)\} \\ &\quad = \binom{S_i(t)}{I_i(0,t+1)} [p_i(S,t)]^{I_i(0,t+1)} h_i(t|S), \qquad (11) \\ &P\{I_i(u+1,t+1)|\boldsymbol{X}(t)\} \\ &\quad = \sum_{r=0}^{I_i(u,t)} \binom{I_i(u,t)}{r} [\gamma_i(u,t)]^r h_{i,r}(u,t|I) \\ &\quad \text{for } u = 0, 1, \ldots, t, \qquad (12) \end{aligned}$$

where the $h_i(t|S)$ and the $h_{i,r}(u,t|I)$ are given respectively by:

$$h_i(t|S) = \sum_{j=0}^{S_i(t+1)-S_i(t)+I_i(0,t+1)} g_{i,S}\{j,t|\boldsymbol{X}(t)\} \times \binom{S_i(t)-I_i(0,t+1)}{a_{i,S}(j,t)} [d_i(S,t)]^{a_{i,S}(j,t)} \times\{1-p_i(S,t)-d_i(S,t)\}^{b_{i,S}(j,t)},$$

with $a_{i,S}(j,t) = S_i(t) - S_i(t+1) - I_i(0,t+1) + j$ and $b_{i,S}(j,t) = S_i(t+1) - j$, and

$$h_{i,r}(u,t|I) = \sum_{j=0}^{I_i(u+1,t+1)-I_i(u,t)+r} g_{i,u}[j,t|\boldsymbol{X}(t)] \times \binom{I_i(u,t)-r}{a_{i,u}(r,j,t)} [d_i(u,t)]^{a_{i,u}(r,j,t)} \times\{1-\gamma_i(u,t)-d_i(u,t)\}^{b_{i,u}(r,j,t)},$$

with $a_{i,u}(r,j,t) = I_i(u,t) - I_i(u+1,t+1) - r + j$ and $b_{i,u}(r,j,t) = I_i(u+1,t+1) + r - 2j$.

In the above equations, notice that aside from the immigration and recruitment, the distribution of $\boldsymbol{X}$ is basically a product of multinomial distributions. Hence, the above model has been referred to as a chain multinomial model; see References 48, 49, 54, 57, 61, 63, 64, 66–68, 73, and 74. Tan and Ye [74] have used the above distribution to estimate both the unknown parameters and the state variables via state space models.

THE STAGED MODEL OF HIV EPIDEMIC IN HOMOSEXUAL POPULATIONS

In the above model, the number of state variables $I_i(u,t)$ and hence the dimension of the state space increases as time increases; that is, if the infection duration is taken into account, then the size of the dimension of the state space increases as time increases. To simplify matters, an alternative approach is to partition the infective stage into a finite number of substages with stochastic transition between the substages and assume that within the substage, the effects of duration is the same. This is the approach proposed by Longini and his associates [33–37]. In the literature, such staging is usually achieved by using the number of CD4$^{(+)}$ T-cell counts per mm^3 blood. The staging system used by Satten and Longini [43,44] is I_1, CD4 counts $\geqslant 900/mm^3$; I_2, $900/mm^3 >$ CD4 counts $\geqslant 700/mm^3$; I_3, $700/mm^3 >$ CD4 counts $\geqslant 500/mm^3$; I_4, $500/mm^3 >$ CD4 counts $\geqslant 350/mm^3$; I_5, $350/mm^3 >$ CD4 counts $\geqslant 200/mm^3$; I_6, $200/mm^3 >$ CD4 counts. (Because of the 1993 AIDS definition by CDC [15], the I_6 stage is merged with the AIDS stage (A stage).)

The staging of the infective people results in a staged model for the HIV epidemic. Comparing these staged models with the previous model, the following differences are observed: *(i)* Because of the infection duration, the nonstaged model is in general not Markov. On the other hand, if one assumes that the transition rates of the substages are independent of the time at which the substage were generated, the staged models are Markov. *(ii)* For the nonstaged model, the number of different type of infectives always increase nonstochastically as time increases. These are referred to as expanding models by Liu and Chen [32]. On the other hand, the number of substages of the infective stage in the staged model is a fixed number independent of time. *(iii)* For the nonstaged model, the infective people increase its infection duration nonstochastically, always increasing by one-time unit with each increase of one-time unit. However, they transit directly to AIDS stochastically. On the other hand, for the staged model, the transition from one substage to another substage is stochastic and can either be forward or backward, or transit directly to AIDS. Because of the random transition between the infective substages, one would expect that the staging has introduced more randomness into the model than the nonstaged model.

Assuming Markov, then one may use some standard results in Markov chain theory to study the HIV epidemic in staged models. This has been done by Longini and his associates [33–37]. Alternatively, by using exactly the same procedures as in the previous model, one may derive the stochastic equation for the state variables as well as the probability distributions of the state variables. By using these equations and the probability distributions, one can then study the stochastic behaviors and to assess effects of risk variables and the impact of some intervention procedures. This has been done by Tan and his associates [48–49,54–55,57,63–64,66–68]. By using the San Francisco homosexual population as an example, they have shown that the staged model gave similar results as the previous model and hence the same conclusions. These results indicates that the errors of approximation and the additional variations due to stochastic transition between the substages imposed by the staging system are in general quite small for the HIV epidemic in homosexual populations. On the other hand, because of the existence of a long asymptomatic infective period with low infectivity, it is expected that the staged model would provide a closer approximation to the real world situations then the nonstaged model.

Another problem in the staged model is the impacts of measurement error as the CD4 T-cell counts are subject to considerable measurement error. To take this into account, Satten and Longini [44] have proposed a hidden Markov model by assuming the measurement errors as Gaussian variables. The calculations done by them did not reveal a significant impact of these errors on the HIV epidemic, however, indicating that the effects of measurement errors on the HIV epidemic of the staged model is not very significant.

STOCHASTIC MODELS OF HIV TRANSMISSION IN COMPLEX SITUATIONS

In Africa, Asia, and many south American countries, although homosexual contact and IV drug use may also be important avenues, most of the HIV epidemic are developed through heterosexual contacts and prostitutes [38]. The dynamics of HIV epidemic in these countries are therefore very different from those in the United States, Canada, and the western countries, where the major avenues of HIV transmission are homosexual contact and sharing needles and IV drug use. It has been documented that even in homosexual populations, race, age, and risk behaviors as well as many other risk variables would significantly affect the HIV epidemic [66–67,75–76]. To account for effects of many risk factors such as sex, race, and age, the above simple stochastic model has been extended into models under complex situations [19,39,46,53,66–67,75–76].

To develop stochastic models of HIV transmission in complex situations, the basic procedures are again the same two steps as described above: *(i)* Stratifying the population into subpopulations by sex, race, and risk factors, derive the probabilities of HIV transmission from infective people to susceptible people in each subpopulation. These probabilities usually involve interactions between people from different risk groups and the structure of the epidemic. *(ii)* Develop steps for HIV progression within each subpopulation. It appears that because the dynamics of HIV epidemic are different under different situations, the first step to derive the probabilities of HIV transmission varies from population to population depending on different situations. Given that the probabilities of HIV transmission from infective people to susceptible people have been derived for each subpopulation, the second step is similar to the progression step of the procedures described above. That is, the only major difference between different models lies in the derivation of the probabilities $p_i(t;S)$ for the ith subpopulation. Assuming that there are no sexual contacts with AIDS cases, the general form of $p_i(t;S)$ is

$$p_i(t;S) = X_i(t) \sum_j \rho_{ij}(t) \frac{I_j(t)}{T_j(t)} \overline{\alpha}_{ij}(t) \qquad (13)$$

Notice that Equation 13 is exactly of the same form as Equation 2; yet, because of the different dynamics in different models, the $\rho_{ij}(t)$'s are very different between different models. In populations of IV drug users, HIV spread mainly through sharing IV needles in small parallel groups. In these populations, therefore, $\rho_{ij}(t)$ is derived by first forming small groups and then spread HIV by sharing needles between members within the group. This is the basic formulation by means of which Capasso et al. [9], Gani and Yakowitz [19], Kaplan [28], and Kaplan et al. [30] derived the probabilities of HIV infection of S people by infective people. Along this line, Tan [55, Chap. 4] has formulated a general procedure to derive this probability.

In populations stratified by race, sex, age, sexual activity levels, and risk behaviors and involving married couples, to derive $\rho_{ij}(t)$ one needs to take into account some realistic preference patterns. These realities include *(i)* People tend to mix more often with people of the same race, same age group, and same sexual activity level; *(ii)* people with high sexual activity levels and/or old age tend to select sexual partners indiscriminately; *(iii)* if the age difference between the two partners is less than five years, then age is not an important factor in selecting sexual partners, and *(vi)* race and sexual activity level may interact with each other to affect the selection of sexual partners; *(v)* a happy marriage would reduce external marital relationship. Taking many of these factors into account and assuming that members of small populations select members from larger populations, Tan and Xiang [66,67] and Tan and Zhu [75,76] have proposed a selective mixing pattern through the construction of acceptance probabilities. Intuitively, this mixing can be expressed as a product of two probability measures: The first is the probability of selecting members from subpopulations with larger effective population size via acceptance probabilities; the second is the conditional probability of selecting members of infective people with different infection duration from the selected population.

By using the selective mixing pattern, Tan and Xiang [66,67] have developed stochastic models for the HIV epidemic in homosexual populations taking into account race, age, and sexual activity levels. Their results indicate that race and age affect the HIV epidemic mainly through the numbers of different sexual partners per partner per month and their interactions with the mixing pattern. Increasing the transition rates of infective people by race and/or by age seems to have some impact on the HIV progression but the effects are much smaller than those from the average numbers of different sexual partners per partner per month and the mixing patterns. Thus, the observed result that there is a much larger proportion of AIDS cases from black people than from white people in the US population ([16]) is a consequence of the following observations: *(i)* Black people in general have larger number of sexual partners per unit time than white people. *(ii)* There is a large proportion of restricted mixing pattern and mixing patterns other than proportional mixing while under these mixing patterns, black people appear to contract HIV much faster than white people.

By using selective mixing pattern, Tan and Zhu [75,76] have developed stochastic models involving single males, single females, married couples, and prostitutes. Their results indicate that the prostitute factor may be the main reason for the rapid growth of the HIV epidemic in some Asian countries such as Thailand and India, which have a large prostitute population. Their Monte Carlo studies also suggest that rapid growth of the HIV epidemic in some Asian countries may be arrested or controlled by promoting extensive use of condoms by prostitutes combined with a campaign of AIDS awareness in the younger and sexually active populations.

STATE SPACE MODELS OF THE HIV EPIDEMIC

State space models of stochastic systems are stochastic models consisting of two submodels: The stochastic system model, which is the stochastic model of the system, and the observation model, which is a statistical model based on available observed data from the system. That is, the state space model adds one more dimension to the stochastic model and to the statistical model by combining both of these models into one model. This is a convenient and efficient approach to combine information from both stochastic models and statistical models. It takes into account the basic mechanisms of the system and the random variation of the system through its stochastic system model and incorporate all these into the observed data from the system; and it validates and upgrades the stochastic model through its observation model and the observed data of the system and the estimates of the state variables. It is advantageous over both the stochastic model and the statistical model when used alone since it combines information and advantages from both of these models. Specifically, one notes that *(i)* Because of additional information, many of the identifiability problems in statistical analysis are nonexistent in state space models; see References 73 and 74 for some examples. *(ii)* It provides an optimal procedure to update the model by new data that may become available in the future. This is the smoothing step of the state space models; see References 2, 12, and 20. *(iii)* It provides an optimal procedure via Gibbs sampling to estimate simultaneously the unknown parameters and the state variables of interest; see References 55, chapter 6; 56, chapter 9; 73; and 74. *(iv)* It provides a general procedure to link molecular events to critical events in population and at the cellular level; see Reference 58.

The state space model (Kalman-filter model) was originally proposed by Kalman and his associates in the early 60s for engineering control and communication [27]. Since then it has been successfully used as a powerful tool in aerospace research, satellite research, and military missile research. It has also been used by economists in econometrics research [21] and by mathematician and statisticians in time series research [3] for solving many difficult problems that appear to be extremely difficult from other approaches. In 1995, Wu and Tan [78,79] had attempted to apply the state space model and method to AIDS research. Since then many papers have been published to develop state space models for the HIV epidemic and the HIV pathogenesis; see References 13, 68–73, 77, and 78. Alternatively, by combining the Markov staged model with Gaussian measurement errors for the CD4 T-cell counts, Satten and Longini [44] have proposed a Hidden Markov model for the HIV epidemic; however, it is shown by Tan [56] that this is a special case of the state space models.

Although Tan and Ye [74] have applied the state space model for the HIV epidemic in the Swiss population of IV drug users, to date, the state space models for HIV epidemic are primarily developed in homosexual populations. To illustrate how to develop state space models for the HIV epidemic, we will thus use the San Francisco homosexual population as an example, although the general results apply to other populations as well.

A STATE SPACE MODEL FOR THE HIV EPIDEMIC IN THE SAN FRANCISCO HOMOSEXUAL POPULATION

Consider the San Francisco homosexual population, in which HIV spread primarily by sexual contact [16]. For this population, Tan and Ye [73] have developed a state space model for the HIV epidemic. For this state space model, the stochastic system model was represented by stochastic difference equations. The observation model of this state space model is based on the monthly AIDS incidence data (i.e., data of new AIDS cases developed during a month period). This data is available from the gofer network of CDC. This is a statistics model used by statistician through the backcalculation method. Combining these two models into a state space model, Tan and Ye [73] have developed a general Bayesian procedure to estimate the HIV infection, the HIV incubation, as well as the numbers of susceptible people, infective people, and AIDS cases. Notice that this is not possible by using the stochastic model alone or by using the statistic model alone because of the identifiability problem.

THE STOCHASTIC SYSTEM MODEL

To develop a stochastic model for the San Francisco population, Tan and Ye [74] have made two simplifying assumptions: *(i)* By visualizing the infection incidence and hence the infection distribution as a mixture of several sexual levels, one sexual activity level may be assumed. *(ii)* Because the population size of the city of San Francisco changes very little, for the S people and I people it is assumed that the number of immigration and recruitment is about the same as the number of death and migration. Tan and Ye [74] have shown that these assumptions have little impacts on the probability distributions of the HIV infection and the HIV incubation. On the basis of these assumptions, then the state variables are $\underset{\sim}{U}(t) = \{S(t), I(u,t), u = 0, 1, \ldots, t, A(t)\}$. Assuming that the probability $p_S(t)$ of HIV infection of S people and the probability $\gamma(s,t) = \gamma(t-s)$ of $I(u) \rightarrow AIDS$ as deterministic functions of time, then the parameters are $\underset{\sim}{\theta}(t) = \{p_S(t), \gamma(u,t) = \gamma(t-u), u = 0, 1, \ldots, t\}'$. The densities of the HIV infection and the HIV incubation are given by $f_I(t) = p_S(t)\Pi_{i=0}^{t-1}[1 - p_S(i)] = G_S(t-1)p_S(t), t = 1, \ldots, \infty$ and $g(t) = \gamma(t)\Pi_{j=0}^{t-1}[1 - \gamma(j)] = G_I(t-1)\gamma(t), t = 1, \ldots, \infty$ respectively.

Let $F_S(t)$ be the number of S people who contract HIV during the *t-th* month and $F_I(u,t)$ the number of $I(u,t) \to A$ during the tth month. Then,

$$S(t+1) = S(t) - F_S(t), \tag{14}$$

$$I(0,t+1) = F_S(t), \tag{15}$$

$$I(u+1,t+1) = I(u,t) - F_I(u,t), \tag{16}$$

$$u = 0,\ldots,t,$$

where $F_S(t)|S(t) \sim B\{S(t),p_S(t)\}$ and $F_I(u,t)| I(ut) \sim B\{I(u,t),\gamma(u)\}, u = 0,1,\ldots,t$.

Put $\Theta = \{\underset{\sim}{\theta}(t), t = 1,\ldots,t_M\}, \underset{\sim}{X}(t) = \{S(t), I(u,t), u = 0,1,\ldots t\}$ and $\boldsymbol{X} = \{\underset{\sim}{X}(1),\ldots, \underset{\sim}{X}(t_M)\}$, where t_M is the last time point. Then, the probability distribution of $\boldsymbol{X}$ given Θ and given $\underset{\sim}{X}(0)$ is

$$P\{\boldsymbol{X}|\underset{\sim}{X}(0)\} = \prod_{j=0}^{t_M-1} P\{\underset{\sim}{X}(j+1)|\underset{\sim}{X}(j),\Theta\}$$

where

$$\begin{aligned} &Pr\{\underset{\sim}{X}(j+1)|\underset{\sim}{X}(j),\Theta\} \\ &\quad = \binom{S(t)}{I(0,t+1)} [p_S(t)]^{I(0,t+1)} \\ &\qquad \times [1-p_S(t)]^{S(t)-I(0,t+1)} \\ &\qquad \times \prod_{u=0}^{t} \binom{I(u,t)}{I(u,t)-I(u+1,t+1)} \\ &\qquad \times [\gamma(u)]^{I(u,t)-I(u+1,t+1)} \\ &\qquad \times [1-\gamma(u)]^{I(u+1,t+1)}. \end{aligned} \tag{17}$$

Notice that the above density is a product of binomial densities and hence has been referred to as the chain binomial distribution.

For the HIV epidemic in the San Francisco homosexual population, Tan and Ye [73] have assumed January 1, 1970, as $t = 0$ since the first AIDS case in San Francisco appeared in 1981 and since the average incubation period for HIV is about 10 years. It is also assumed that, in 1970, there are no infective people but to start the HIV epidemic, some HIV were introduced into the population in 1970. Thus, one may take $I(0,0) = 36$ because this is the number of AIDS in San Francisco in 1981. Tan and Ye [73] have assumed the size of the San Francisco homosexual population in 1970 as 50000 because with a 1% increase in population size per year by the US census survey [77], the estimate of the size of the San Francisco homosexual population is $58{,}048 = 50{,}000 \times (1.01)^{15}$ in 1985, which is very close to the estimate 58,500 of the size of the San Francisco homosexual population in 1985 by Lemp et al. [31].

THE OBSERVATION MODEL.

Let $y(j+1)$ be the observed AIDS incidence during the jth month and $A(t+1)$ the number of new AIDS cases developed during the tth month. Then the stochastic equation for the observation model is

$$\begin{aligned} y(j+1) &= A(j+1) + \xi(j+1) \\ &= \sum_{u=0}^{j} F_I(u,j) + \xi(j+1) \\ &= \sum_{u=0}^{j} [I(u,t) - I(u+1,t+1)] \\ &\quad + \xi(t+1) \\ &= \sum_{u=0}^{j} I(u,j)\gamma(u) + \epsilon_A(j+1) + \xi(j+1) \\ &= \sum_{u=0}^{j} I(u,j)\gamma(u) + e(j+1), \end{aligned} \tag{18}$$

where $\xi(t+1)$ is the random measurement error associated with observing $y(j+1)$ and $\epsilon_A(j+1) = [F_s(t) - S(t)p_s(t)] + \sum_{u=1}^{j}[F_I(u,t) - I(u,t)\gamma(u)]$.

Put $\mathbf{Y} = \{y(j), j = 1,\ldots,t_M\}$. Assuming that the $\xi(j)$ are independently distributed as normal with means 0 and variance σ_j^2, then the likelihood function $P\{\mathbf{Y}|\mathbf{X},\Theta\} = L(\Theta|\mathbf{Y},\mathbf{X})$ given the state variables is

$$\begin{aligned} &P\{\mathbf{Y}|\mathbf{X},\Theta\} \\ &\quad \propto \prod_{j=1}^{t_M} \left(\sigma_j^{-1} \exp\left\{-\frac{1}{2\sigma_j^2}[y(j) - A(j)]^2\right\}\right). \end{aligned} \tag{19}$$

Notice that under the assumption that $\{p_s(t), \gamma(t)\}$ are deterministic functions of t,

$$\begin{aligned}
E[S(t+1)] &= E[S(t)][1-p_s(t)] \\
&= E[S(t-1)]\prod_{i=t-1}^{t}\{1-p_s(i)\} \\
&= E[S(0)]\prod_{i=0}^{t}\{1-p_s(i)\} \\
&= E[S(0)]G_s(t), \qquad (20) \\
E[F_s(t)] &= E[S(t)]p_s(t) \\
&= E[S(0)]G_s(t-1)p_s(t) \\
&= E[S(0)]f_I(t), \qquad (21) \\
E[I(u+1,t+1)] &= E[I(u,t)][1-\gamma(u)] \\
&= E[I(0,t-u)]\prod_{j=0}^{u}[1-\gamma(j)] \\
&= E[I(0,t-u)]G_I(u) \\
&= E[S(0)]f_I(t-u)G_I(u). \qquad (22)
\end{aligned}$$

Hence,

$$\begin{aligned}
&E[I(u,t)-I(u+1,t+1)] \\
&\quad = E[S(0)]f_I(t-u)\{G_I(u-1)-G_I(u)\} \\
&\quad = E[S(0)]f_I(t-u)G_I(u-1)\gamma(u) \\
&\quad = E[S(0)]f_I(t-u)g(u). \qquad (23)
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sum_{u=0}^{t}E[I(u,t)-I(u+1,t+1)] \\
&\qquad = E[S(0)]\sum_{u=0}^{t}f_I(t-u)g(u) \\
&\qquad = E[S(0)]\sum_{u=0}^{t}f_I(u)g(t-u), \qquad (24)
\end{aligned}$$

so that

$$y(j+1) = E[S(0)]\sum_{u=0}^{t}f_I(u)g(t-u)+e(j+1). \qquad (25)$$

Notice that Equation 25 is the convolution formula used in the backcalculation method [4,55]. This implies that the backcalculation method is the observation model in the state space model. The backcalculation method is not identifiable because using Equation 25 alone and ignoring information from the stochastic system model, the information is not sufficient for estimating all the parameters.

A GENERAL BAYESIAN PROCEDURE FOR ESTIMATING THE UNKNOWN PARAMETERS AND THE STATE VARIABLES

By using the state space model, Tan and Ye [73,74] have developed a generalized Bayesian approach to estimate the unknown parameters and the state variables. This approach will combine information from three sources: *(i)* Previous information and experiences about the parameters in terms of the prior distribution of the parameters, *(ii)* biological information via the stochastic system equations of the stochastic system, and *(iii)* information from observed data via the statistical model from the system.

To illustrate the basic principle of this method, let $P(\Theta)$ be the prior distribution of Θ. Then, the joint distribution of $\{\Theta, \boldsymbol{X}, \boldsymbol{Y}\}$ is given by $P(\Theta, \boldsymbol{X}, \boldsymbol{Y}) = P(\Theta)P(\boldsymbol{X}|\Theta)P(\boldsymbol{Y}|\boldsymbol{X}, \Theta)$. From this, the conditional distribution $P(\boldsymbol{X}|\Theta, \boldsymbol{Y})$ of $\boldsymbol{X}$ given $(\Theta, \boldsymbol{Y})$ and the conditional posterior distribution $P(\Theta|\boldsymbol{X}, \boldsymbol{Y})$ of Θ given $(\boldsymbol{X}, \boldsymbol{Y})$ are given respectively by

$$\begin{aligned}
&\text{(A):} \quad P(\boldsymbol{X}|\Theta, \boldsymbol{Y}) \propto P(\boldsymbol{X}|\Theta)P(\boldsymbol{Y}|\boldsymbol{X}, \Theta) \\
&\text{(B):} \quad P(\Theta|\boldsymbol{X}, \boldsymbol{Y}) \propto P(\Theta)P(\boldsymbol{X}|\Theta)P(\boldsymbol{Y}|\boldsymbol{X}, \Theta)
\end{aligned}$$

Given these probability densities, one may use the multilevel Gibbs sampling method to derive estimates of Θ and $\boldsymbol{X}$ given $\boldsymbol{Y}$ [45]. This is a Monte Carlo sequential procedure alternating between two steps until convergence: *(i)* Given $\{\Theta, \boldsymbol{Y}\}$, one generates $\boldsymbol{X}$ by using $P(\boldsymbol{X}|\Theta, \boldsymbol{Y})$ from (A). These are the Kalman-filter estimates. *(ii)* Using the Kalman-filter estimates of $\underset{\sim}{X}$ from (A) and given $\boldsymbol{Y}$, one generates values of Θ by using $P(\Theta|\boldsymbol{X}, \boldsymbol{Y})$ from (B). Iterating between these two steps until convergence, one then generates random samples from the conditional probability distribution $P(\boldsymbol{X}|\boldsymbol{Y})$ independently of Θ, and from the posterior distribution $P(\Theta|\boldsymbol{Y})$ independently of $\boldsymbol{X}$, respectively. This provides the Bayesian estimates of Θ given data and the Bayesian estimates of $\boldsymbol{X}$ given data, respectively. The proof of the convergence can be developed by using basic theory of stationary distributions in irreducible and aperiodic Markov chains; see Reference 56, chapter 3.

Using the above approach, one can readily estimate simultaneously the numbers of S people, I people, and AIDS cases as well as the parameters $\{p_S(t), \gamma(t)\}$. With the estimation of $\{p_S(t), \gamma(t)\}$, one may then estimate the HIV infection distribution $f_I(t)$ and the HIV incubation distribution $g(t)$. For the San Francisco homosexual population, the estimates of $f_I(t)$ and $g(t)$ are plotted in Figs. 1 and 2. Given below are some basic findings by Tan and Ye [73]:

(a) From Fig 1; the estimated density of the HIV infection clearly showed a mixture of distributions with two obvious peaks. The first peak (the higher peak) occurs around January 1980 and the second peak around March 1992.

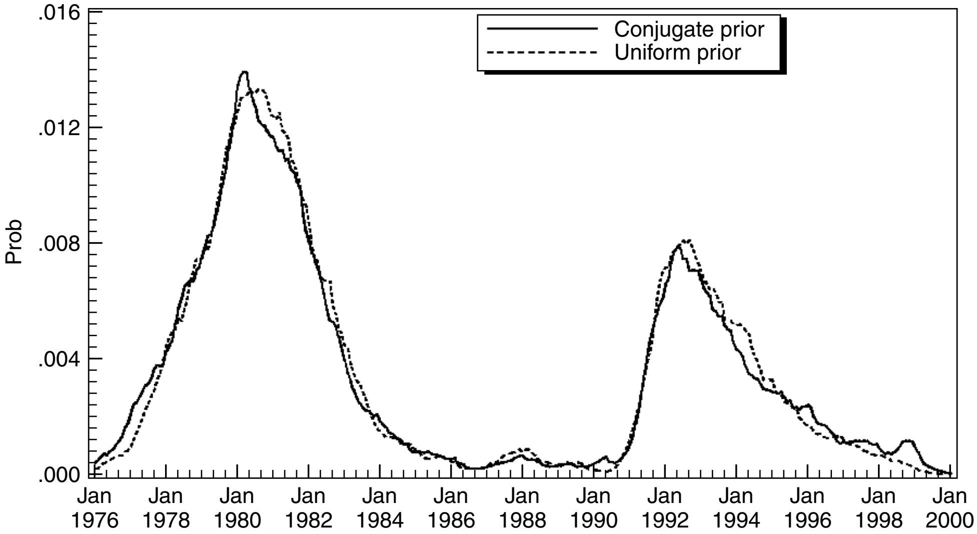


Figure 1. Plots of the estimated HIV infection distribution

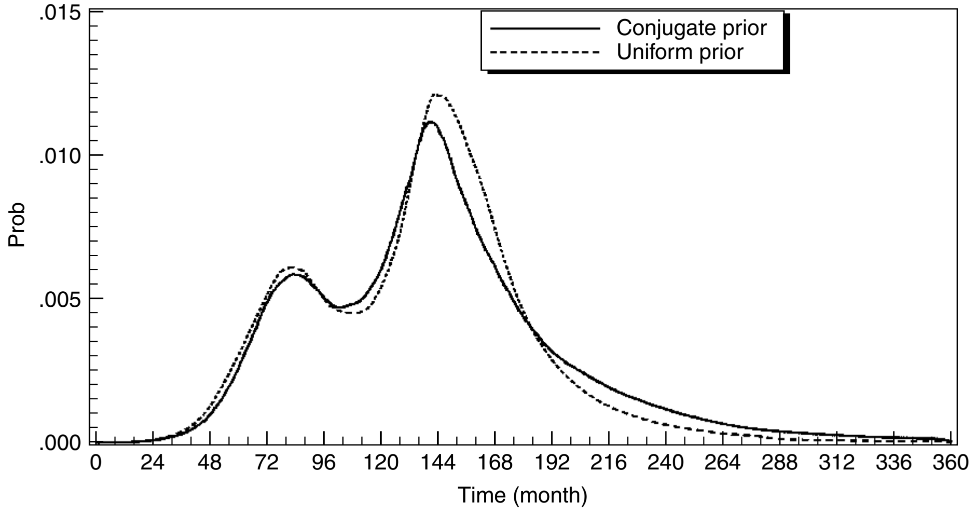


Figure 2. Plots of the estimated HIV incubation distribution

The mixture nature of this density implies that there are more than one sexual activity levels with a high proportion of restricted mixing (like-with-like) mixing. The second peak also implies a second wave of infection although the infection intensity is much smaller than the first wave.

(b) From Fig 2, the estimated density of the HIV incubation distribution also appeared to be a mixture of distributions with two obvious peaks. The first peak is around 75 months after infection and is much lower than the second peak which occurs around 140 months after infection. This result suggests a multistage nature of the HIV incubation.

(c) The Kalman-filter estimates of the AIDS incidence by the Gibbs sampler are almost identical to the corresponding observed AIDS incidence respectively. This result indicates that the Kalman-filter estimates can trace the observed numbers very closely if observed numbers are available.

(d) Figuring a 1% increase in the population size of San Francisco yearly, Tan and Ye [73] have also estimated the number of S people and the I people in the San Francisco population. The estimates showed that the total number of S people before January 1978 were always above 50,000 and were between 31,000 and 32,000 during January 1983 and January 1993. The total number of people who do not have AIDS were estimated around 50,000 before January 1993. It appeared that the total number of infected people reached a peak around the middle of 1985 and then decreased gradually to the lowest level around 1992. The estimates also showed that the number of infected people had two peaks, with the higher peak around the middle of 1985, the second peak around the year 2000 and with the lowest level around 1992.

Extending the above state space model to include immigration and death, Tan and Ye [74] have also analyzed the data of the Swiss population of homosexual men and IV drug users by applying the above generalized Bayesian method. The estimated density of the HIV infection in the Swiss homosexual population is quite similar to that of the San Francisco population except that in the Swiss population, the first peak appears about 6 months earlier and the second peak about 2 years earlier. Similarly, the estimated density of the HIV incubation distribution in the Swiss population is a mixture of distributions with two obvious peaks. The higher peak occurs around 320 months after infection and the lower peak occurs around 232 months after infection. In the Swiss population, the estimates of the immigration and recruitment rates are about 10 times greater than those of the estimates of the death and retirement rates of the I people, suggesting that the size of the Swiss homosexual and bisexual population is increasing with time. Another interesting point is that, in the Swiss population, the estimates of the death and retirement rates of infective people were much greater (at least 100 times greater) than those of S people, suggesting that HIV infection may have increased the death and retirement rates of HIV infected people.

REFERENCES

1. Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, U.K.
2. Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J.
3. Aoki, M. (1990). *State Space Modeling of Time Series*, 2nd ed. Spring-Verlag, New York.
4. Bacchetti, P. R. (1990). Estimating the incubation period of AIDS comparing population infection and diagnosis pattern. *J. Amer. Statist. Assoc.*, **85**, 1002–1008.
5. Billard, L. and Zhao, Z. (1991). Three-stage stochastic epidemic model: an application to AIDS. *Math. Biosci.*, **107**, 431–450.
6. Billard, L. and Zhao, Z. (1994). Multi-stage non-homogeneous Markov models for the acquired immune deficiency syndrome epidemic. *J. R. Stat. Soc. B*, **56**, 673–686.
7. Blanchard, P., Bolz, G. F., and Krüger, T. (1990). "Modelling AIDS-Epidemics or Any

Veneral Disease on Random Graphs". In *Stochastic Processes in Epidemic Theory*, Lecture Notes in Biomathematics, No. 86, eds. J. P. Gabriel, C. Lefévre, and P. Picard, eds. Springer-Verlag, New York, pp. 104–117.

8. Brookmeyer, R. and Gail, M. H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford, U.K.
9. Capassor, V., Di Somma, M., Villa, M., Nicolosi, A., and Sicurello, F. (1997). "Multistage Models of HIV Transmission Among Injecting Drug Users via Shared Injection Equipment". In *Advances in Mathematical Population Dynamics- Molecules, Cells and Man, Part II. Population Dynamics in Diseases in Man*, O. Arino, D. Axelrod, and M. Kimmel, eds. World Scientific, Singapore, Chapter 8, pp. 511–528.
10. Carpenter, C. C. J., Fischl, M. A., Hammer, M. D., Hirsch, M. S., Jacobsen, D. M., Katzenstein, D. A., Montaner, J. S. G., Richman, D. D., Saag, M. S., Schooley, R. T., Thompson, M. A., Vella, S., Yeni, P. G., and Volberding, P. A. (1996). Antiretroviral therapy for the HIV infection in 1996. *J. Amer. Med. Ass.*, **276**, 146–154.
11. Castillo-Chavez, C., ed. (1989). *Mathematical and Statistical Approaches to AIDS Epidemiology*, Lecture Notes in Biomathematics 83. Springer-Verlag, New York.
12. Catlin, D. E. (1989). *Estimation, Control and Discrete Kalman Filter*. Springer-Verlag, New York.
13. Cazelles, B. and Chau, N. P. (1997). Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Math. Biosci.*, **140**, 131–154.
14. CDC. (1981). Pneumocystis pneumonia - Los Angeles. *MMWR*, **30**, 250–252.
15. CDC, (1992). Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR*, **41**(RR-17), 1–19.
16. CDC. (1997). *Surveillance Report of HIV/AIDS*. Atlanta, Ga., June 1997.
17. Coffin, J., Haase, J., Levy, J. A., Montagnier, L., Oroszlan, S., Teich, N., Temin, H., Toyoshima, K., Varmus, H., Vogt, P., and Weiss, R. (1986). Human immunodeficiency viruses. *Science*, **232**, 697.
18. Gani, J. (1990). "Approaches to the Modelling of AIDS". In *Stochastic Processes in Epidemic Theory*, Lecture Notes in Biomathematics No. 86, J. P. Gabriel, C. Leférre, and P. Picards, eds. Springer-Verlag, Berlin, pp. 145–154.
19. Gani, J. and Yakowitz, S. (1993). Modeling the spread of HIV among intravenous drug users. *IMA J. Math. Appl. Med. Biol.*, **10**, 51–65.
20. Gelb, A. (1974). *Applied Optimal Estimation*. MIT Press, Cambridge, Mass.
21. Harvey, A. C. (1994). *Forcasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, U.K.
22. Hethcote, H. W. and Van Ark, J. M. (1992). *Modeling HIV Transmission and AIDS in the United States*, Lecture Notes in Biomathematics 95. Springer-Verlag, New York.
23. Isham, V. (1991). Assessing the variability of stochastic epidemics. *Math. Biosci.*, **107**, 209–224.
24. Isham, V. and Medley, G., eds. (1996). *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Cambridge University Press, Cambridge, U.K.
25. Jager, J. C. and Ruittenberg, E. J., eds. (1988). *Statistical Analysis and Mathematical Modelling of AIDS*. Oxford University Press, Oxford, U.K.
26. Jewell, N. P., Dietz, K., and Farewell, V. T. (1992). *AIDS Epidemiology: Methodological Issues*. Birkhäuser, Basel.
27. Kalman, R. E. (1960). A new approach to linear filter and prediction problems. *J. Basic Eng.*, **82**, 35–45.
28. Kaplan, E. H. (1989). Needles that kill: modeling human immuno-deficiency virus transmission via shared drug injection equipment in shooting galleries. *Rev. Infect. Dis.*, **11**, 289–298.
29. Kaplan, E. H. and Brandeau, M. L., eds. (1994). *Modeling the AIDS Epidemic*. Raven Press, New York.
30. Kaplan, E. H., Cramton, P. C., and Paltiel, A. D. (1989). "Nonrandom Mixing Models of HIV Transmission". In *Mathematical and Statistical Approaches to AIDS Epidemiology*, Lecture Notes in Biomathematics 83, C. Castillo-Chavez, ed. Springer-Verlag, Berlin, pp. 218–241.
31. Lemp, G. F., Payne, S. F., Rutherford, G. W., Hessol, N. A., Winkelstein, W. Jr., Wiley, J. A., Moss, A. R., Chaisson, R. E., Chen, R. T., Feigal, D. W. Jr., Thomas, P. A., and Werdegar, D. (1990). Projections of AIDS morbidity and mortality in San Francisco. *J. Am. Med. Assoc.*, **263**, 1497–1501.
32. Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo method for dynamic systems. *J. Am. Stat. Assoc.*, **93**, 1032–1044.

33. Longini Ira, M. Jr., Byers, R. H., Hessol, N. A., and Tan, W. Y. (1992). Estimation of the stage-specific numbers of HIV infections via a Markov model and backcalculation. *Stat. Med.*, **11**, 831–843.

34. Longini Ira, M. Jr., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F. and Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Stat. Med.*, **8**, 831–843.

35. Longini Ira, M. Jr., Clark, W. S., Gardner, L. I., and Brundage, J. F. (1991). The dynamics of CD4+ T-lymphocyte decline in HIV-infected individuals: a Markov modeling approach. *J. AIDS*, **4**, 1141–1147.

36. Longini Ira, M. Jr., Clark, W. S., and Karon, J. (1993). Effects of routine use of therapy in slowing the clinical course of human immunodeficiency virus (HIV) infection in a population based cohort. *Am. J. Epidemiol.*, **137**, 1229–1240.

37. Longini Ira, M. Jr., Clark, W. S., Satten, G. A., Byers, R. B., and Karon, J. M., (1996). "Staged Markov Models Based on CD4$^{(+)}$ T-Lymphocytes for the Natural History of HIV Infection. In *Models for Infectious Human Diseases: Their Structure and Relation to Data*, V. Isham and G. Medley, eds. Cambridge University Press, Cambridge, U.K., pp. 439–459.

38. Mann, J. M. and Tarantola, D. J. M. (1998). HIV 1998: The global picture. *Sci. Am.*, **279**, 82–83.

39. Mode, C. J. (1991). A stochastic model for the development of an AIDS epidemic in a heterosexual population. *Math. Biosci.*, **107**, 491–520.

40. Mode, C. J., Gollwitzer, H. E., and Herrmann, N. (1988). A methodological study of a stochastic model of an AIDS epidemic. *Math. Biosci.*, **92**, 201–229.

41. Mode, C. J., Gollwitzer, H. E., Salsburc, M. A., and Sleeman, C. K. (1989). A methodological study of a nonlinear stochastic model of an AIDS epidemic with recruitment. *IMA J. Math. Appl. Med. Biol.*, **6**, 179–203.

42. Mode, C. J. and Sleeman, C. K. (2000). *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases and Computers*. World Scientific, River Edge, N.J.

43. Satten, G. A. and Longini Ira, M. Jr. (1994). Estimation of incidence of HIV infection using cross-sectional marker survey. *Biometrics*, 50, 675–68.

44. Satten, G. A. and Longini Ira, M. Jr. (1996). Markov chain with measurement error: estimaxting the 'True' course of marker of the progression of human immunodeficiency virus disease. *Appl. Stat.*, **45**, 275–309.

45. Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.

46. Tan, W. Y. (1990). A Stochastic model for AIDS in complex situations. *Math. Comput. Modell.*, **14**, 644–648.

47. Tan, W. Y. (1991). Stochastic models for the spread of AIDS and some simulation results. *Math. Comput. Modell.*, **15**, 19–39.

48. Tan, W. Y. (1991). "On the Chain Multinomial Model of HIV Epidemic in Homosexual Populations and Effects of Randomness of Risk Factors". In *Mathematical Population Dynamics 3*, O. Arino, D. E. Axelrod, and M. Kimmel, eds. Wuerz Publishing, Winnepeg, Manitoba, Canada, pp. 331–353.

49. Tan, W. Y. (1993). The chain multinomial models of the HIV epidemiology in homosexual populations. *Math. Comput. Modell.*, **18**, 29–72.

50. Tan, W. Y. (1994). On the HIV Incubation Distribution under non-Markovian Models. *Stat. Probab. Lett.*, **21**, 49–57.

51. Tan, W. Y. (1994). On first passage probability in Markov models and the HIV incubation distribution under drug treatment. *Math. Comput. Modell.*, **19**, 53–66.

52. Tan, W. Y. (1995). On the HIV incubation distribution under AZT treatment. *Biometric. J.*, **37**, 318–338.

53. Tan, W. Y. (1995). A stochastic model for drug resistance and the HIV incubation distribution. *Stat. Probab. Lett.*, **25**, 289–299.

54. Tan, W. Y. (1995). "On the Chain Multinomial Model of HIV Epidemic in Homosexual Populations and Effects of Randomness of Risk Factors". In *Mathematical Population Dynamics 3*, O. Arino, D. E. Axelrod, and M. Kimmel, eds. Wuerz Publishing, Winnipeg, Manitoba, Canada, pp. 331–356.

55. Tan, W. Y. (2000). *Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis*. World Scientific, River Edge, N.J.

56. Tan, W. Y. (2002). *Stochastic Models With Applications to genetics, Cancers, AIDS and Other Biomedical Systems*. World Scientific, River Edge, N.J.

57. Tan, W. Y. and Byers, R. H. (1993). A stochastic model of the HIV epidemic and the HIV infection distribution in a homosexual population. *Math. Biosci.*, **113**, 115–143.

58. Tan, W. Y., Chen, C. W., and Wang, W. (2000). *A generalized state space model of carcinogenesis*. Paper Presented at the *2000 International*

Biometric Conference at UC. Berkeley, Calif., July 2–7, 2000.

59. Tan, W. Y. and Hsu, H. (1989). Some stochastic models of AIDS spread. *Stat. Med.*, **8**, 121–136.
60. Tan, W. Y. and Hsu, H. (1991). ”A Stochastic Model for the AIDS Epidemic in a Homosexual Population”. In *“Mathematical Population Dynamics*, eds. O. Arion, D. E. Axelrod, and M. Kimmel, eds. Marcel Dekker, New York, Chapter 24, pp. 347–368.
61. Tan, W. Y, Lee, S. R., and Tang, S. C. (1995). Characterization of HIV infection and seroconversion by a stochastic model of HIV epidemic. *Math. Biosci.*, **126**, 81–123.
62. Tan, W. Y. and Tang, S. C. (1993). Stochastic models of HIV epidemics in homosexual populations involving both sexual contact and IV drug use. *Math. Comput. Modell.*, **17**, 31–57.
63. Tan, W. Y., Tang, S. C., and Lee, S. R. (1995). Effects of Randomness of risk factors on the HIV epidemic in homo-sexual populations. *SIAM J. Appl. Math.*, **55**, 1697–1723.
64. Tan, W. Y., Tang, S. C., and Lee, S. R. (1996). Characterization of the HIV incubation distributions and some comparative studies. *Stat. Med.*, **15**, 197–220.
65. Tan, W. Y., Tang, S. C., and Lee, S. R. (1998). Estimation of HIV seroconversion and effects of age in San Francisco homosexual populations. *J. Appl. Stat.*, **25**, 85–102.
66. Tan, W. Y. and Xiang, Z. H. (1996). A stochastic model for the HIV epidemic and effects of age and race. *Math. Comput. Modell.*, **24**, 67–105.
67. Tan, W. Y. and Xiang, Z. H.(1997). ”A Stochastic Model for the HIV Epidemic in Homosexual Populations and Effects of Age and Race on the HIV Infection”. In *Advances in Mathematical Population Dynamics- Molecules, Cells and Man”. Part II. Population Dynamics in Diseases in Man*, O. Arino, D. Axelrod, and M. Kimmel, eds. World Scientific, River Edge, N.J, Chapter 8, pp. 425–452.
68. Tan, W. Y. and Xiang, Z. H. (1998). State space models of the HIV epidemic in homosexual populations and some applications. *Math. Biosci.*, **152**, 29–61.
69. Tan, W. Y. and Xiang, Z. H. (1998). ”Estimating and Predicting the numbers of T Cells and Free HIV by Non-Linear Kalman Filter”. In *Artificial Immune Systems and Their Applications*, D. DasGupta, ed. Springer-Verlag, Berlin, pp. 115–138.
70. Tan, W. Y. and Xiang, Z. H.(1998). ”State Space Models for the HIV Pathogenesis”. In *“Mathematical Models in Medicine and Health Sciences*, M. A. Horn, G. Simonett, and G. Webb, eds. Vanderbilt University Press, Nashville, Tenn., pp. 351–368.
71. Tan, W. Y. and Xiang, Z. H. (1999). Modeling the HIV epidemic with variable infection in homosexual populations by state space models. *J. Stat. Inference Plann.*, **78**, 71–87.
72. Tan, W. Y. and Xiang, Z. H. (1999). A state space model of HIV pathogenesis under treatment by anti-viral drugs in HIV-infected individuals. *Math. Biosci.*, **156**, 69–94.
73. Tan, W. Y. and Ye, Z. Z. (2000). Estimation of HIV infection and HIV incubation via state space models. *Math. Biosci.*, **167**, 31–50.
74. Tan, W. Y. and Ye, Z. Z. (2000). Simultaneous estimation of HIV infection, HIV incubation, immigration rates and death rates as well as the numbers of susceptible people, infected people and AIDS cases. *Commun. Stat. (Theory Methods)*, **29**, 1059–1088.
75. Tan, W. Y. and Zhu, S. F. (1996). A stochastic model for the HIV infection by heterosexual transmission involving married couples and prostitutes. *Math. Comput. Modell.*, **24**, 47–107.
76. Tan, W. Y. and Zhu, S. F. (1996). A stochastic model for the HIV epidemic and effects of age and race. *Math. Comput. Modell.*, **24**, 67–105.
77. U.S. Bureau of the Census. (1987). *Statistical Abstract of the United States: 1980*, 108th ed. Washington, D.C.
78. Wu, H. and Tan, W. Y. (1995). ”Modeling the HIV Epidemic: A State Space Approach”. *ASA 1995 Proceedings of the epidemiology Section*. ASA, Alexdria, Va., pp. 66–71.
79. Wu, H. and Tan, W. Y. (2000). Modeling the HIV epidemic: a state space approach. *Math. Comput. Modell.*, **32**, 197–215.

See also Clinical Trials; Epidemics; and Epidemiological Statistics—I.

Wai-Yuan Tan

AITCHISON DISTRIBUTIONS

These form a class of multivariate distributions with density functions:

$$
\begin{aligned}
f_{\mathrm{X}}(\mathbf{X}|\alpha,\beta) \propto {} & \left[\prod_{i=1}^{p} x_i^{\alpha_i-1}\right] \\
& \times \exp\left[-\frac{1}{2}\sum^{p}\sum_{i<j} \beta_{ij}(\log x_i - \log x_j)^2\right],
\end{aligned}
$$

$$0 < x_i, \quad \sum_{i=1}^{p} x_i = 1,$$

$$\alpha_i \geqslant 0, \quad \boldsymbol{\beta} \text{ nonnegative definite.}$$

Although there are p variables $x_1, \ldots, x_p$, the distribution is confined to the $(p-1)$-dimensional simplex: $0 \leqslant x_i, \sum_{i=1}^{p} x_i = 1$.

If $\boldsymbol{\beta} = \mathbf{0}$ we have a *Dirichlet distribution**; if $\alpha_i = 0$ for all i we have a *multivariate logistic-normal distribution*.

Methods of estimating the parameters (α and β) are described by Aitchison [1, pp. 310–313].

REFERENCE

1. Aitchison, J. (1986). *Statistical Analysis of Compositional Data*. Chapman and Hall, London and New York.

See also Compositional Data; Dirichlet Distribution; Frequency Surfaces, Systems of; and Logistic-Normal Distribution.

AITKEN, ALEXANDER CRAIG

Alexander Craig Aitken was born April 1, 1885 in Dunedin, New Zealand, and died in Edinburgh, Scotland, on November 3, 1967, where he had spent his working life. Dunedin is a rather Scottish community on the southern tip of New Zealand. His father was a farmer. Aitken made important contributions to statistics, numerical analysis, and algebra. His extraordinary memory, musical gift, attractive personality, work, and teaching talents are well described in the obituary articles [1,2].

After attending Otago Boys High School, he studied at Otago University classical languages for two years. In April, 1915 he enlisted in the New Zealand Infantry and served in Gallipoli, Egypt, and France in World War I—experiences movingly described in a manuscript (written while he was recovering from wounds in 1917), but not turned into a book [3] until 1962—his last publication. Aitken had total recall of his past, and this section of it always gave him great pain. His platoon was all but wiped out in the battle of the Somme along with all records. He was able to remember and write down all the information in the records of all these men. Upon recovery, he returned to Otago University. He could not study mathematics there, though it was realized that he had a gift for it. He then taught languages in his old school for three years. Upon graduating in 1920, he married a fellow student, and later they had a son and a daughter.

Fortunately, in 1923 he was given a scholarship to study mathematics under E. T. Whittaker in Edinburgh. The first edition of Whittaker and Robinson's *The Calculus of Observations* [4] (W&R) appeared in 1924. Whittaker had earlier considered *graduation** or *smoothing of data* as a statistical problem; his motivation was mainly actuarial. So arose—almost—what we now call splines. The function chosen to be minimized was the sum of the squares of the differences on the observed u_n and "true" values u_n' plus a multiple of the sum of squares of the *third* differences of the "true" values. How to execute this was Aitken's Ph.D. problem. Whittaker was so pleased with Aitken's results that he was awarded a D.Sc. instead and a staff appointment. His method is given in W&R. In a preceding section in W&R, they speak of the *method of interlaced parabolas*, in which they fit a cubic polynomial to each successive four graduated values u_n'—this allows interpolation. Had the second term in the minimand been the integral of the square of the third derivative of the interpolating function, they would have invented the modern method of getting a spline.

The Mathematics Department in Edinburgh had over many years very broad interests that spanned all of applied mathematics, and Whittaker was a towering figure in many fields. In particular, it was then the only place in Britain that taught determinants and matrices, and these immediately appealed to Aitken. In 1932 he published with H. W. Turnbull, *An Introduction of the Theory of Canonical Matrices* [5]. Its last chapter gives, among other applications, some to statistics. So by then he had shown his real interests—algebra, numerical analysis, and statistics. Aitken succeeded Whittaker in the Chair in Mathematics in 1946, holding it until he retired in 1965.

Aitken was a renowned teacher. In 1939 he published the first two volumes in the Oliver & Boyd series (of which he was a joint editor with D. E. Rutherford) *University Mathematical Texts*. They were *Statistical Mathematics* [6] and *Determinants and Matrices* [7]. These two books were my first acquaintance with his work. When I was an undergraduate there were very few books on these topics with any style to them.

Around 1949, as a graduate student, I became aware of his research, largely written in the preceding ten years, in statistics, especially of his matrix treatment of least squares (see e.g. ref. [8])—idempotents like $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, the use of the trace operator (e.g. $E\mathbf{x}^t\mathbf{A}\mathbf{x} = \mathrm{Tr}\, E\mathbf{A}\mathbf{x}\mathbf{x}^t$), etc. This has now become standard. Indeed, with the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ where the random error vector has mean $\mathbf{0}$ and nonsingular covariance matrix $\mathbf{V}$, the estimating equations $\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}$ are known as the *Aitken equations**. But he never mentioned vector spaces, although he would have been aware of the background geometry. He wrote many papers about least squares and especially about the fitting of polynomials and on other statistical topics. We have only singled out here three areas to mention.

In the late thirties he gave a student from New Zealand, H. Silverstone, a Ph.D. topic—the optimal estimation of statistical parameters—which he had apparently already worked out for a scalar parameter. (See Aitken & Silverstone (1941) [9].) In a 1947 paper [10] Aitken completed the work for many parameters. But he has never received credit for this work, which seems unfair to me. *For the time at which it was written*, the formulation is correct, as are the answers. But instead of a direct proof of a minimum, they simply give the "Euler equation" that the calculus of variations (C of V) throws up, though with the full knowledge that that method has difficulties. This was very natural, as the study of minimal surfaces was then of great interest, and much use was made of the C of V in mathematical physics.

To take the simpler problem with his notation, let $\Phi(\mathbf{x}, \theta)$ be the density of the vector $\mathbf{x}$, which ranges over a region which is the same for all θ, and be uniformly differentiable with respect to θ. Suppose there exists $t(\mathbf{x})$, a minimum-variance estimator of θ. Then $\int t\partial\Phi/\partial\theta\, d\mathbf{x} = 1$, and $\int(t-\theta)^2\Phi d\mathbf{x}$ must be a minimum. Writing $I(t) = \int(t-\theta)^2\Phi d\mathbf{x} - 2\lambda\int t\partial\Phi/\partial\theta d\mathbf{x}$, consider any other estimator $t + h$. Then

$$I(t+h) = I(t) + \int h\left(2(t-\theta)\Phi - 2\lambda\frac{\partial\Phi}{\partial\theta}\right) d\mathbf{x} + \int h^2\Phi\, d\mathbf{x} \geqslant I(t)$$

if and only if $t - \theta = \lambda\partial\Phi/\partial\theta$, where λ may be a function of θ. This is the "Euler" equation they give (and, I would guess, the proof they had), and from which they correctly draw all the now well-known conclusions. It took however many other statisticians many years to clarify these questions and to find what other assumptions are necessary. M. S. Bartlett has published some correspondence [11] with Aitken.

Several other features of Aitken's life are his love of music, his awesome memory, and his arithmetic ability. He played the violin well and for a time was leader of the Edinburgh University Musical Society Orchestra, which was sometimes conducted by his close friend Sir Donald Tovey. His violin was particularly important to him when in the Army, and it now is displayed in his old school. He said that 75% of the time his thoughts were musical. On occasion he would demonstrate his arithmetic feats. He wrote a book [12] against the decimalization of the English coinage—the use of twelve, which has so many factors, appealed to him. He was able to dictate rapidly the first 707 digits of π.

Among his honors, he was a Fellow of the Royal Societies of Edinburgh and London and of the Royal Society of Literature.

REFERENCES

1. Copson, E. T., Kendall, D. G., Miller, J. C. P., and Ledermann, W. (1968). Obituary articles. *Proc. Edinburgh Math. Soc.*, **16**, 151–176.

2. Whittaker, J. M. and Bartlett, M. H. (1968). Alexander Craig Aitken. *Biogr. Mem. Fell. R. Soc. Lond.*, **14**, 1–14.
3. Aitken, A. C. (1962). *Gallipoli to the Somme*. Oxford University Press, London. 177 pp.
4. Whittaker, E. T. and Robinson, G. (1924). *The Calculus of Observations*. Blackie & Son, London and Glasgow (6th impression, 1937).
5. Turnbull, H. W. and Aitken, A. C. (1932). *An Introduction to the Theory of Canonical Matrices*. Blackie & Son, London and Glasgow. 192 pp.
6. Aitken, A. C. (1939). *Statistical Mathematics*. Oliver & Boyd, Edinburgh. 153 pp.
7. Aitken, A. C. (1939). *Determinants and Matrices*. Oliver & Boyd, Edinburgh. 144 pp.
8. Aitken, A. C. (1934–35). On least squares and linear combinations of observations. *Proc. Roy. Soc. Edinburgh A*, **55**, 42–48.
9. Aitken, A. C. and Silverstone, H. (1941). On the estimation of statistical parameters. *Proc. Roy. Soc. Edinburgh A*, **61**, 186–194.
10. Aitken, A. C. (1947). On the estimation of many statistical parameters. *Proc. Roy. Soc. Edinburgh A*, **62**, 369–370.
11. Bartlett, M. S. (1994). Some pre-war statistical correspondence. In *Probability, Statistics and Optimization*, F. P. Kelly, ed. Wiley, Chichester and New York, pp. 297–413.
12. Aitken, A. C. (1962). *The Case against Decimalization*. Oliver & Boyd, Edinburgh, 22 pp.

G. S. Watson

AITKEN EQUATIONS

One of the most important results in the theory of linear models is the Gauss—Markov theorem*. It establishes that for the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$E(\mathbf{e}) = \boldsymbol{\phi} \quad \text{and} \quad E(\mathbf{e}\mathbf{e}^T) = \sigma^2\mathbf{I} \qquad (1)$$

the minimum variance linear unbiased estimator for an estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is given by $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a solution of the normal equations*

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}. \qquad (2)$$

The case of equal variances and zero correlations for the elements of **e**, as given in (1), was generalized first by Aitken* [1] to the case where the observations and hence the elements of **e** have different variances and/or are correlated, i.e., (1) is replaced by

$$E(\mathbf{e}) = \boldsymbol{\phi}, E(\mathbf{e}\mathbf{e}^T) = \Sigma = \sigma^2\mathbf{V}. \qquad (3)$$

In (3), Σ represents a known (positive definite) variance—covariance matrix, but since it needs to be known only up to a constant, Σ is often rewritten as $\sigma^2\mathbf{V}$, where $\mathbf{V}$ represents a known matrix (note that for the case of equal variances $\sigma^2, \mathbf{V}$ is a correlation matrix). As a consequence of the variance—covariance structure (3) the equations (2) are replaced by the Aitken equations

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \qquad (4)$$

which are obtained by minimizing the expression (see [1])

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

It is for this reason that the method is also referred to as weighted least squares* or generalized least squares, as compared to ordinary least squares (OLS), which leads to (2).

With the conditions (3) the minimum-variance linear unbiased estimator for an estimable function $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is given by $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a solution to the Aitken equations (4). A proof can be found in ref. [2].

An interesting and important question is: For $\mathbf{V} \neq \mathbf{I}$. when are the OLS and the weighted least squares estimators for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ the same? The answer is (see e.g. [2]), when $\mathbf{V}\mathbf{X} = \mathbf{X}\mathbf{Q}$ for some matrix $\mathbf{Q}$. This condition holds, for example, for mixed linear models with balanced data.

REFERENCES

1. Aitken, A. C. (1934/35). On least squares and linear combination of observations. *Proc. Roy. Soc. Edinburgh*, **55**, 42–48.
2. Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments*, Vol. 1. Wiley, New York.

See also General Linear Model and Weighted Least Squares.

Klaus Hinkelmann

AKAIKE'S INFORMATION CRITERION. See MODEL SELECTION: AKAIKE'S INFORMATION CRITERION

ALEATORY VARIABLE

An obsolete term for *random variable*.

ALGEBRA OF EVENTS

Let Ω be a space whose points correspond to the possible outcomes of a random experiment. Certain subsets of Ω are called *events*, and *probability* is assigned to these subsets. A collection $\mathscr{F}$ of subsets of Ω is called an *algebra* (the term *field* is also used) if the following conditions are satisfied:

(a) The space Ω belongs to $\mathscr{F}(\Omega \in \mathscr{F})$.

(b) The collection $\mathscr{F}$ is closed under complementation and finite union. Formally:

$\mathbf{b}_1$: If $A \in \mathscr{F}$, then the complement $\overline{A}$ (also denoted as A^c) belongs to $\mathscr{F}$.

$\mathbf{b}_2$: If $A_1, \ldots, A_n \in \mathscr{F}$, then union $\cup_{i=1}^n A_i$ (also denoted as $A_1 \cup \ldots \cup A_n) \in \mathscr{F}$.

[Since $\overline{(\cup_{i=1}^n \overline{A}_i)} = \cap_{i=1}^n A_i$, b_1 and b_2 imply that an algebra is also closed under finite intersection.]

If in place of b_2 we require $\mathscr{F}$ to be closed under *countable* union, namely, if $A_1, A_2, \ldots \in \mathscr{F}$, then $\cup_{i=1}^\infty A_i \in \mathscr{F}$, the collection $\mathscr{F}$ is called a *σ-algebra* (or *σ-field*). The notion of the σ-algebra of events is a basic concept for theoretical probability theory.

See also AXIOMS OF PROBABILITY.

ALGORITHM

An algorithm is a rule for performing a calculation—usually, although not necessarily, numerical. For example, one might have algorithms for classificatory purposes, as well as for evaluation of roots of determinantal equations. Algorithms do not provide any background for the calculations to which they refer, either in terms of motivation or justification.

Algorithms for specific purposes are described in separate entries, in particular in the article ALGORITHMS, STATISTICAL.

ALGORITHMIC INDEPENDENCE. See ALGORITHMIC INFORMATION THEORY

ALGORITHMIC INFORMATION THEORY

The Shannon entropy* concept of classical information theory* [9] is an ensemble notion; it is a measure of the degree of ignorance concerning which possibility holds in an ensemble with a given a priori probability distribution*

$$H(p_1, \ldots, p_n) \equiv -\sum_{k=1}^{n} p_k \log_2 p_k.$$

In algorithmic information theory the primary concept is that of the *information content* of an individual object, which is a measure of how difficult it is to specify or describe how to construct or calculate that object. This notion is also known as *information-theoretic complexity*. For introductory expositions, see refs. 1, 4, and 6. For the necessary background on computability theory and mathematical logic, see refs. 3, 7, and 8. For a more technical survey of algorithmic information theory and a more complete bibliography, see ref. 2. See also ref. 5.

The original formulation of the concept of algorithmic information is independently due to R. J. Solomonoff [22], A. N. Kolmogorov* [19], and G. J. Chaitin [10]. The information content $I(x)$ of a binary string x is defined to be the size in bits (binary digits) of the smallest program for a canonical universal computer U to calculate x. (That the computer U is universal means that for any other computer M there is a prefix μ such that the program μp makes U do exactly the same computation that the program p makes M do.) The *joint information* $I(x,y)$ of two strings is defined to be the size of the smallest program that makes U calculate both of them. And the *conditional* or *relative information* $I(x|y)$ of x

given y is defined to be the size of the smallest program for U to calculate x from y. The choice of the standard computer U introduces at most an $O\,(1)$ uncertainty in the numerical value of these concepts. [$O(f)$ is read "order of f" and denotes a function whose absolute value is bounded by a constant times f.]

With the original formulation of these definitions, for most x one has

$$I(x) = |x| + O(1) \tag{1}$$

(here $|x|$ denotes the length or size of the string x, in bits), but unfortunately

$$I(x,y) \leqslant I(x) + I(y) + O(1) \tag{2}$$

holds only if one replaces the $O(1)$ error estimate by $O(\log I(x)I(y))$.

Chaitin [12] and L. A. Levin [20] independently discovered how to reformulate these definitions so that the subadditivity property (2) holds. The change is to require that the set of meaningful computer programs be an instantaneous code, i.e., that no program be a prefix of another. With this modification, (2) now holds, but instead of (1) most x satisfy

$$\begin{aligned} I(x) &= |x| + I(|x|) + O(1) \\ &= |x| + O(\log |x|). \end{aligned}$$

Moreover, in this theory the decomposition of the joint information of two objects into the sum of the information content of the first object added to the relative information of the second one given the first has a different form than in classical information theory. In fact, instead of

$$I(x,y) = I(x) + I(y|x) + O(1), \tag{3}$$

one has

$$I(x,y) = I(x) + I(y|x, I(x)) + O(1). \tag{4}$$

That (3) is false follows from the fact that $I(x, I(x)) = I(x) + O(1)$ and $I(I(x)|x)$ is unbounded. This was noted by Chaitin [12] and studied more precisely by Solovay [12, p. 339] and Gač [17].

Two other concepts of algorithmic information theory are *mutual* or *common information* and *algorithmic independence*. Their importance has been emphasized by Fine [5, p. 141]. The mutual information content of two strings is defined as follows:

$$I(x:y) \equiv I(x) + I(y) - I(x,y).$$

In other words, the mutual information* of two strings is the extent to which it is more economical to calculate them together than to calculate them separately. And x and y are said to be algorithmically independent if their mutual information $I(x:y)$ is essentially zero, i.e., if $I(x,y)$ is approximately equal to $I(x) + I(y)$. Mutual information is symmetrical, i.e., $I(x:y) = I(y:x) + O(1)$. More important, from the decomposition (4) one obtains the following two alternative expressions for mutual information:

$$\begin{aligned} I(x:y) &= I(x) - I(x|y, I(y)) + O(1) \\ &= I(y) - I(y|x, I(x)) + O(1). \end{aligned}$$

Thus this notion of mutual information, although it applies to individual objects rather than to ensembles, shares many of the formal properties of the classical version of this concept.

Up until now there have been two principal applications of algorithmic information theory: (a) to provide a new conceptual foundation for probability theory and statistics by making it possible to rigorously define the notion of a *random sequence**, and (b) to provide an information-theoretic approach to metamathematics and the limitative theorems of mathematical logic. A possible application to theoretical mathematical biology is also mentioned below.

A random or patternless binary sequence x_n of length n may be defined to be one of maximal or near-maximal complexity, i.e., one whose complexity $I(x_n)$ is not much less than n. Similarly, an infinite binary sequence x may be defined to be random if its initial segments x_n are all random finite binary sequences. More precisely, x is random if and only if

$$\exists c \forall n [I(x_n) > n - c]. \tag{5}$$

In other words, the infinite sequence x is random if and only if there exists a c such that for all positive integers n, the algorithmic information content of the string consisting of the first n bits of the sequence x, is bounded from below by $n - c$. Similarly, a *random real number* may be defined to be one having the property that the base 2 expansion of its fractional part is a random infinite binary sequence.

These definitions are intended to capture the intuitive notion of a lawless, chaotic, unstructured sequence. Sequences certified as random in this sense would be ideal for use in Monte Carlo* calculations [14], and they would also be ideal as one-time pads for Vernam ciphers or as encryption keys [16]. Unfortunately, as we shall see below, it is a variant of Göel's famous incompleteness theorem that such certification is impossible. It is a corollary that no pseudorandom number* generator can satisfy these definitions. Indeed, consider a real number x, such as $\sqrt{2}$, π, or e, which has the property that it is possible to compute the successive binary digits of its base 2 expansion. Such x satisfy

$$I(x_n) = I(n) + O(1) = O(\log n)$$

and are therefore maximally nonrandom. Nevertheless, most real numbers are random. In fact, if each bit of an infinite binary sequence is produced by an independent toss of an unbiased coin, then the probability that it will satisfy (5) is 1. We consider next a particularly interesting random real number, Ω, discovered by Chaitin [12, p. 336].

A. M. Turing's theorem that the halting problem is unsolvable is a fundamental result of the theory of algorithms [4]. Turing's theorem states that there is no mechanical procedure for deciding whether or not an arbitrary program p eventually comes to a halt when run on the universal computer U. Let Ω be the probability that the standard computer U eventually halts if each bit of its program p is produced by an independent toss of an unbiased coin. The unsolvability of the halting problem is intimately connected to the fact that the halting probability Ω is a random real number, i.e., its base 2 expansion is a random infinite binary sequence in the very strong sense (5) defined above. From (5) it follows that Ω is normal (a notion due to E. Borel [18]), that Ω is a Kollectiv* with respect to all computable place selection rules (a concept due to R. von Mises and A. Church [15]), and it also follows that Ω satisfies all computable statistical tests of randomness* (this notion being due to P. Martin-Löf [21]). An essay by C. H. Bennett on other remarkable properties of Ω, including its immunity to computable gambling schemes, is contained in ref. 6.

K. Gödel established his famous incompleteness theorem by modifying the paradox of the liar; instead of "This statement is false" he considers "This statement is unprovable." The latter statement is true if and only if it is unprovable; it follows that not all true statements are theorems and thus that any formalization of mathematical logic is incomplete [3,7,8]. More relevant to algorithmic information theory is the paradox of "the smallest positive integer that cannot be specified in less than a billion words." The contradiction is that the phrase in quotes only has 14 words, even though at least 1 billion should be necessary. This is a version of the Berry paradox, first published by Russell [7, p. 153]. To obtain a theorem rather than a contradiction, one considers instead "the binary string s which has the shortest proof that its complexity $I(s)$ is greater than 1 billion." The point is that this string s cannot exist. This leads one to the metatheorem that although most bit strings are random and have information content approximately equal to their lengths, it is impossible to prove that a specific string has information content greater than n unless one is using at least n bits of axioms. See ref. 4 for a more complete exposition of this information-theoretic version of Gödel's incompleteness theorem, which was first presented in ref. 11. It can also be shown that n bits of assumptions or postulates are needed to be able to determine the first n bits of the base 2 expansion of the real number Ω.

Finally, it should be pointed out that these concepts are potentially relevant to biology. The algorithmic approach is closer to the intuitive notion of the information content of a biological organism than is the classical ensemble viewpoint, for the role of a computer program and of deoxyribonucleic acid (DNA) are roughly analogous. Reference 13 discusses possible applications of the concept of mutual algorithmic information to theoretical biology; it is suggested that a living organism might be defined as a highly correlated region, one whose parts have high mutual information.

GENERAL REFERENCES

1. Chaitin, G. J. (1975). *Sci. Amer.*, **232** (5), 47–52. (An introduction to algorithmic information theory emphasizing the meaning of the basic concepts.)
2. Chaitin, G. J. (1977). *IBM J. Res. Dev.*, **21**, 350–359, 496. (A survey of algorithmic information theory.)
3. Davis, M., ed. (1965). *The Undecidable—Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions*. Raven Press, New York.
4. Davis, M. (1978). In *Mathematics Today: Twelve Informal Essays*, L. A. Steen, ed. Springer-Verlag, New York, pp. 241–267. (An introduction to algorithmic information theory largely devoted to a detailed presentation of the relevant background in computability theory and mathematical logic.)
5. Fine, T. L. (1973). *Theories of Probability: An Examination of Foundations*. Academic Press, New York. (A survey of the remarkably diverse proposals that have been made for formulating probability mathematically. Caution: The material on algorithmic information theory contains some inaccuracies, and it is also somewhat dated as a result of recent rapid progress in this field.)
6. Gardner, M. (1979). *Sci. Amer.*, **241** (5), 20–34. (An introduction to algorithmic information theory emphasizing the fundamental role played by Ω.)
7. Heijenoort, J. van, ed. (1977). *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge, Mass. (This book and ref. 3 comprise a stimulating collection of all the classic papers on computability theory and mathematical logic.)
8. Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, New York. (The longest and most lucid introduction to computability theory and mathematical logic.)
9. Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Ill. (The first and still one of the very best books on classical information theory.)

ADDITIONAL REFERENCES

10. Chaitin, G. J. (1966). *J. ACM*, **13**, 547–569; **16**, 145–159 (1969).
11. Chaitin, G. J. (1974). *IEEE Trans. Inf. Theory*, **IT-20**, 10–15.
12. Chaitin, G. J. (1975). *J. ACM*, **22**, 329–340.
13. Chaitin, G. J. (1979). In *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus, eds. MIT Press, Cambridge, Mass., pp. 477–498.
14. Chaitin, G. J. and Schwartz, J. T. (1978). *Commun. Pure Appl. Math.*, **31**, 521–527.
15. Church, A. (1940). *Bull. AMS*, **46**, 130–135.
16. Feistel, H. (1973). *Sci. Amer.*, **228** (5), 15–23.
17. Gač, P. (1974). *Sov. Math. Dokl.*, **15**, 1477–1480.
18. Kac, M. (1959). *Statistical Independence in Probability, Analysis and Number Theory*. Mathematical Association of America, Washington, D.C.
19. Kolmogorov, A. N. (1965). *Problems of Inf. Transmission*, **1**, 1–7.
20. Levin, L. A. (1974). *Problems of Inf. Transmission*, **10**, 206–210.
21. Martin-Löf, P. (1966). *Inf. Control*, **9**, 602–619.
22. Solomonoff, R. J. (1964). *Inf. Control*, **7**, 1–22, 224–254.

See also ENTROPY; INFORMATION THEORY AND CODING THEORY; MARTINGALES; MONTE CARLO METHODS; PSEUDO-RANDOM NUMBER GENERATORS; STATISTICAL INDEPENDENCE; and RANDOMNESS, TESTS OF.

G. J. CHAITIN

ALGORITHMS, STATISTICAL

Traditionally, in mathematics, the term "algorithm"* means "some special process for solving a certain type of problem" [3].[1] With the advent of automatic computing, the term was adopted to refer to the description of a process in a form suitable for implementation on a computer. Intuitively, an algorithm is useful in mathematics or in computing if the "type of problem" is well defined and if the "special process" can be used effectively for these problems. A reasonable definition of the term for our purposes is:

> An algorithm is a process for the solution of a type of problem, such that the process can be

> implemented computationally without significant difficulty and that the class of problems treated is computationally specific and well understood.

Statistical algorithms are those algorithms having useful application to problems encountered in statistics. They are not, it should be emphasized, restricted to algorithms written by or specifically for statisticians. Such a restriction would exclude a wide range of useful work and, unfortunately, would still include a number of inferior approaches to some problems.

In the general process of using computers to assist in statistical analysis of data, three aspects are frequently important: the recognition of the need for an algorithm (more generally, the role of algorithms in the overall approach); the attempt to find or implement a suitable algorithm; and judgments about the quality of an algorithm. Let us consider each of these questions in turn.

ALGORITHMS AND STATISTICAL COMPUTING

The importance of good algorithms derives from their role as building blocks supporting reliable, flexible computing. Statisticians (and equally, physicists, chemists, engineers, and other users of computers) have tended to plunge in with ad hoc attacks on specific computing problems, with relatively little use of existing algorithms or research in computing. Many arguments, some of them quite sound, support this approach. The end user is interested in the "answer" (in our case, the statistical analysis), not in the process that produces it. Particularly at early stages of statistical computing, the statistician was often not familiar with computing, either in the sense of a user or in the more important sense of understanding some of the basic principles of computation. The problem to be solved often appeared straightforward, with a solution that was qualitatively obvious to the statistician. In this case, finding or developing an algorithm seems a waste of valuable time. Furthermore, it may not be at all obvious how a statistical problem can be formulated in appropriate terms for algorithms which frequently were devised for other problem areas.

Paradoxically, some of the statistical systems and packages developed to assist statisticians aggravate the tendency to take ad hoc rather than algorithmic approaches. The many conveniences of using high-level systems make it tempting to rig an intuitively plausible solution within the system rather than reach outside to find a high-quality algorithm for the problem. In many systems, the process of integrating such an algorithm into the system may require a high degree of programming skill and knowledge of the system's inner workings.

Although arguments for casual solutions to statistical computing problems have some force, there are stronger arguments that statisticians should try to integrate high-quality algorithms into their computing. Two arguments are particularly important. First, the use of good computational algorithms generally improves the use of our own time, in spite of the widely held intuition to the contrary. Second, the quality and the defensibility of the statistical analysis of data is eventually inseparable from the quality of the underlying computations.

Support for the first argument is that well-chosen algorithms will not only increase the chance that a particular statistical computation succeeds relatively quickly, but will usually greatly simplify the (inevitable) process of adapting the computation to new data or to a change in the analysis. As for the second argument, this is asserting both that the statistician should understand what an analysis has produced, in clear and precise terms, and also that the operational steps should be communicable and independently reproducible by others. Well-defined and correct computations are needed if statistical analysis is to satisfy fundamental criteria of scientific validity. This requires, in turn, that computations in statistical systems and specially programmed data analysis be based on algorithms that are accepted as correct implementations of valid computational methods.

As computing evolves, statisticians should be able to combine the convenience of statistical systems with the use of high-quality algorithms. Statistical systems increasingly incorporate good algorithms for the common operations. Advances in techniques of language design and implementation can simplify the process of integrating new algorithms into such systems, gradually merging the process of user programming and system extension [2].

SPECIFYING ALGORITHMS

(This section is directed largely to persons writing algorithms.) Our definition of an algorithm requires that it be suitable for implementation on a computer but deliberately does not restrict the form in which the algorithm is specified. The specification may be translatable mechanically into computer steps, i.e., may be given in a programming language. Alternatively, the specification may be instructions that someone familiar with a programming language can understand and implement. An important goal of computer science is to merge the two forms by improving programming languages and the art of algorithm design to the point that algorithms can be presented in a computer-readable form which is at the same time comprehensible to reasonably well informed human beings. Steps toward this goal, such as techniques of structured programming, are of value in that they increase the chance that one can understand what an algorithm does and hence the degree of confidence in its correctness.

The most convenient specification of an algorithm would intuitively seem to be in a programming language that is locally available, so that a running program could in principle be generated directly. This convenience has to be tempered by the need to understand the algorithm and occasionally by the unsuitability of the common programming languages (e.g., FORTRAN) to handle certain problems (e.g., random number generation*). Clear verbal descriptions are still important as supplements to program code. In some cases, semiverbal presentations can be used either as supplements or as replacement for actual code. Two styles of semiverbal description are used: natural language statements organized into numbered steps, usually with iteration among the steps; and "pidgin" programming languages, with most of the description identical to some language, but with natural language inserted where the actual code would be harder to understand and with details omitted.

Given that an algorithm is to be presented in a programming language, which one will be most helpful? The overwhelming majority of published and otherwise generally circulated algorithms are written in FORTRAN, at least for scientific computing. Presenting an algorithm in this language is then likely to make it implementable widely (at least on the larger computers) and allow it to be used with many existing programs. Other, older languages are less frequently used. ALGOL60 was designed specifically to bridge the previously mentioned gap between readability and implementability; however, at the time of its design, it could take only a partial step in this direction. Although many early published algorithms were written in ALGOL60 (for a time the only accepted language for algorithm sections), FORTRAN has largely taken over, in spite of its deficiencies in generality and readability. Many of the ALGOL60 algorithms were subsequently translated into FORTRAN (some examples will be mentioned in the section "Finding Algorithms"). Other languages, such as PL-1 and COBOL, have at most marginal relevance to algorithms for statistics.

Three other languages do, however, need to be considered: APL, BASIC, and PASCAL. These are all important for interactive computing, particularly on the smaller machines. APL is widely used in statistical analysis; its advantages are its interactive nature and a general, convenient approach to arrays*. Some existing algorithms written in APL have been ad hoc and poorly designed. Nevertheless, there is a large community of users. Also, improvements in APL have made the description of some calculations more attractive (in particular, the inclusion in APL of some key operators to support the kind of calculations done in regression* and multivariate analysis*).

BASIC is also interactive, but in appearance is a (simplified) language of the FORTRAN family. It shares the advantages of availability on small computers, relative ease of initial learning, and a sizable user community. As with APL, the language has suffered at times from algorithms written without enough understanding of the problem. Both languages have been somewhat neglected by the computer-science community involved in developing high-quality algorithms. The neglect is a combination of professional isolation and some intrinsic flaws in the languages themselves. For example, both languages are rather clumsy for expressing the iterative calculations that most algorithms involve. BASIC, in addition, may make the process of separate definition of algorithms difficult.

PASCAL is again oriented to the use of small, interactive computers and can be learned fairly easily. It derives, however, from the ALGOL family of languages. PASCAL is a simple, structured language, well adapted to writing many types of algorithms in a clear and readable form. One of its attractions, in fact, is to the portion of the computer-science community interested in writing programs whose correctness can be formally verified. For these reasons, PASCAL is perhaps the most attractive new language for the specification of algorithms. At the time of writing, however, its applications to statistical computing are minimal. Applications of PASCAL are mostly to non-numerical problems. Its importance as a vehicle for statistical algorithms is largely in the future.

FINDING ALGORITHMS

There are several sources for statistical algorithms, with no simple process for searching them all. Algorithms from the various sources will tend to differ in reliability and in the convenience of implementation. Roughly in descending order of overall reliability, the major sources are:

Published Algorithm Sections

Several computing journals have published algorithms in one of a set of accepted programming languages (typically FORTRAN and ALGOL60). These algorithms have been independently referred and (in principle) tested. They should conform to specified requirements for quality (see the next section) established by journal policy. The journal *Applied Statistics** publishes such an algorithm section specifically for statistical computing. Some statistical algorithms have also appeared in *Communications in Statistics** (*B*). Major general algorithm sections appear in *Transactions on Mathematical Software* and *The Computer Journal*. The publication *Collected Algorithms of the Association of Computing Machinery* reprints the former set of general algorithms and contains an important cumulative index, covering most published algorithm sections as well as many algorithms published separately in scientific journals.

General Algorithm Libraries

These are collections of algorithms, usually distributed in machine-readable form, for a wide range of problems. Although the algorithms are often the work of many people, the libraries usually exert some central editorial control over the code. As a result, from the user's viewpoint, greater uniformity and simplicity can be achieved. However, the distributors may not be as disinterested judges of the library contents as are editors of algorithm sections. Confidence in the quality of the library rests to a large extent on evaluation of the organization distributing it. The International Mathematical and Statistical Library (IMSL), specifically oriented to statistical algorithms, is distributed by an independent organization in suitable FORTRAN source for many computer systems. The National Algorithm Group (NAG) is a publicly sponsored British organization designed to coordinate the distribution and development of algorithms. In this work it has had the cooperation of a number of professional groups, including the Royal Statistical Society*. A number of scientific laboratories also maintain and distribute general algorithm libraries, e.g., the PORT library (Bell Laboratories), Harwell Laboratory (U.K. Atomic Energy Research Establishment), and the National Physical Laboratory.

Specialized Algorithm Packages

These are less general collections of algorithms than the previous. They provide a range of solutions to a set of related problems, frequently in greater detail than that provided by general libraries. In addition, they attack some problem areas that tend to be ignored by published algorithms, such as graphics*. Questions of reliability will be similar to the general algorithm libraries. A series of specialized packages has been developed with the cooperation of Argonne National Laboratories, covering topics such as eigenvalue problems, linear equations, and function approximation. Graphics packages include the GR-Z package (Bell Laboratories) for data analysis and the DISSPLA package (a general-purpose system distributed commercially).

Scientific Journals

In addition to published algorithm sections, many published papers contain algorithm descriptions, either in one of the semiverbal forms or in an actual programming language. A qualified referee should have examined the paper, but unless given an explicit statement, it is probably unwise to assume that the algorithm has been independently implemented and tested. Nevertheless, there are a number of problems for which the only satisfactory published algorithms are of this form (e.g., some random number generation* techniques).

Unpublished Papers; Program Sharing

These categories are perhaps last resorts—least in average quality but certainly not least in quantity. It may be that more algorithms exist in these forms than in all other categories combined. They are usually not referred, except unintentionally by users, and one should expect to spend time testing them before putting them into regular use. Simply finding out about the algorithms requires considerable effort. Of most help are library search techniques and centralized clearinghouses for technical reports (such as the National Technical Information Service in the United States).

With increased familiarity, the process of searching the various sources will become more straightforward. Services provided by technical libraries, such as literature searches (now often computerized and relatively inexpensive) and centralized listings of papers, books, and memoranda, are extremely valuable. Modern library personnel are often very knowledgeable and helpful in searching through the jungle of technical literature. Of course, once one or more algorithms have been found, there remains the question of whether they are adequate and, if not, what steps can be taken to improve or replace them.

THE QUALITY OF ALGORITHMS

The problem of evaluating algorithms has no simple solution. For most statistical applications, a sensible judgment about algorithms requires some understanding of the computational methods being used. The discussion in Chambers [1] and in the further references cited there provides background to some of the computational methods important for statistics. Although it is tempting to hope that some mechanical evaluation of algorithms could resolve their quality thoroughly, this is rarely the case. Most problems are too complex for an evaluation that treats the algorithm as a black box*; i.e., as a phenomenon to be judged only by its empirical performance, without regard for the techniques used. A tendency to use only this approach to evaluate statistical software is regrettable, particularly since it reinforces the overall ad hoc approach which has been detrimental to statistical computing in the past.

In the process of evaluating algorithms, both empirically and in terms of the method used, one may apply some general guidelines. Four helpful classes of questions are the following.

Is The Algorithm Useful? Does it solve the problem at hand? Is it general enough for all the cases likely to be encountered? Will it adapt to similar problems to be encountered later, or will a new algorithm have to be found essentially from scratch?

Is The Algorithm Correct? Will it run successfully on all the cases? If not, will it detect and clearly indicate any failures? For numerical calculations, what guarantees of accuracy are available? (If not theoretical estimates beforehand, are there at least reliable measures of accuracy after the fact?) We emphasize again that such judgments require understanding of what numerical methods can do to solve the problem.

How Hard Is It To Implement And Use? Does the form of the algorithm require considerable local effort (e.g., because the algorithm is written in English or in a programming language not locally available)? Does the algorithm as implemented make inconvenient assumptions (such as limits on the size of problem that can be handled)? Is it written portably, or are there features that will need to be changed locally? Most important, is the algorithm comprehensible, so that there is some hope of fixing problems or making modifications after one is committed to its use?

Is The Algorithm Efficient? Will its requirements for storage space, running time, or other computer resources be modest enough to make it practical for the problems at hand? Are there convenient, general estimates of these requirements? Issues of efficiency are often overemphasized, in the sense that the human costs involved in the previous questions are far more important in most applications. Nevertheless, we can still encounter problems that exceed the capacity of current computing, and it is good to be careful of such situations. As with accuracy, it is important to understand what computing science can do for the problem. Both theoretical estimates (of the order of difficulty, frequently) and empirical estimates are helpful.

NOTE

1. For the general reader, the complete Oxford English Dictionary gives algorithm (preferably algorism) as meaning the arabic numerals, with the chastening added meaning of a cipher or nonentity.

REFERENCES

1. Chambers, J. M. (1977). *Computational Methods for Data Analysis*. Wiley, New York.
2. Chamber, J. M. (1980). *Amer. Statist.*, 34, 238–243.
3. James, G. and James, R. C. (1959). *Mathematics Dictionary*. D. Van Nostrand, Princeton, N.J.

BIBLIOGRAPHY

The following is a selection of references, by subject, to some of the more useful algorithms for statistical applications. In most cases, the algorithms are presented in the form of subprograms or procedures in some programming language. A few are (reasonably precise) verbal descriptions to be followed by the reader in implementing the procedure. There is, of course, no assertion that these are "best" algorithms. Most of them do present a good combination of reliability, generality, and simplicity.

In addition to these references, several organizations provide algorithm libraries. Two sources that should be mentioned specifically for statistical and scientific computing are:

International Mathematical and Statistical Libraries, Inc. (IMSL), 7500 Bellaire Boulevard, Houston, Tex. 77036, USA.

The Numerical Algorithms Group (NAG), 7 Banbury Road, Oxford OX2 6NN, England.

Fourier Transforms

Singleton, R. C. (1968). *Commun. ACM*, **11**, 773–779.

Singleton, R. C. (1969). *IEEE Trans. Audio Electroacoust.*, **17**, 93–103.

Graphics

Akima, H. (1978). *ACM Trans. Math. Software*, **4**, 148–159.

Becker, R. A. and Chambers, J. M. (1977). *The GR-Z System of Graphical Subroutines for Data Analysis*. Write to Computer Information Library, Bell Laboratories, Murray Hill, N. J. 07974.

Crane, C. M. (1972). *Computer J.*, **15**, 382–384.

Doane, D. P. (1976). *Amer. Statist.*, **30**, 181–183.

Lewart, C. R. (1973). *Commun. ACM*, **16**, 639–640.

Newman, W. M. and Sproull, R. F. (1979). *Principles of Interactive Computer Graphics*, 2nd ed. McGraw-Hill, New York.

Scott, W. (1979). *Biometrika*, **66**, 605–610.

Nonlinear Models

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, N. J.

Gill, P. E. and Murray, W. (1970). *A Numerically Stable Form of the Simplex Algorithm*. *Rep. Math. No. 87*, National Physical Laboratory, Teddington, England.

Lill, A. S. (1970). *Computer J.*, **13**, 111–113; also, *ibid.*, **14**, 106, 214 (1971).

O'Neill, R. (1971). *Appl. Statist.*, **20**, 338–345.

Shanno, D. F. and Phua, K. H. (1976). *ACM Trans. Math. Software*, **2**, 87–94.

Numerical Approximation

Cody, W. J., Fraser, W., and Hart, J. F. (1968). *Numer. Math.*, **12**, 242–251.

Hart, J. F., Cheney, E. W., Lawson, C. L., Maehly, H. J., Mesztenyi, C. K., Rice, J. R., Thacher, H. C., and Witzgall, C. (1968). *Computer Approximations*. Wiley, New York.

Numerical Integration

Blue, J. L. (1975). *Automatic Numerical Quadrature: DQUAD. Comp. Sci. Tech. Rep. No*. 25, Bell Laboratories, Murray Hill, N. J.

Gentleman, W. M. (1972). *Commun. ACM*, **15**, 353–355.

Numerical Linear Algebra

Businger, P. A. and Golub, G. H. (1969). *Commun. ACM*, **12**, 564–565.

Wilkinson, J. H. and Reinsch, C. eds. (1971). *Handbook for Automatic Computation* Vol. 2: *Linear Algebra*. Springer-Verlag, Berlin. (Contains a wide selection of algorithms, many subsequently used in the EISPAK package developed at Argonne Laboratories.)

Programming

Ryder, B. G. (1974). *Software—Pract. Exper.*, **4**, 359–377.

Sande, G. (1975). *Proc. 8th Comp. Sci. Statist. Interface Symp.* Health Sciences Computing Facility, UCLA, Los Angeles, Calif., pp. 325–326.

Random Numbers

Ahrens, J. H. and Dieter, U. (1972). *Commun. ACM*, **15**, 873–882.

Ahrens, J. H. and Dieter, U. (1974). *Computing*, **12**, 223–246.

Ahrens, J. H. and Dieter, U. (1974). *Acceptance-Rejection Techniques for Sampling from the Gamma and Beta Distributions*. *Tech. Rep. No. AD*-782478, Stanford University, Stanford, Calif. (Available from National Technical Information Service.)

Chambers, J. M., Mallows, C. L., and Stuck, B. W. (1976). *J. Amer. Statist. Ass.*, **71**, 340–344.

Kinderman, A. J. and Monahan, J. F. (1977). *ACM Trans. Math. Software*, **3**, 257–560.

Regression

Barrodale, I. and Roberts, F. D. K. (1974). *Commun. ACM*, **17**, 319–320.

Chambers, J. M. (1971). *J. Amer. Statist. Ass.*, **66**, 744–748.

Daniel, J. W., Gragg, W. B., Kaufman, L., and Stewart, G. W. (1976). *Math. Comp.*, **30**, 772–795.

Gentleman, W. M. (1974). *Appl. Statist.*, **23**, 448–454.

Wampler, R. H. (1979). *ACM Trans. Math. Software*, **5**, 457–465.

Sorting

Brent, R. P. (1973). *Commun. ACM*, **16**, 105–109.

Chambers, J. M. (1971). *Commun. ACM*, **14**, 357–358.

Loesser, R. (1976). *ACM Trans. Math. Software*, **2**, 290–299.

Singleton, R. C. (1969). *Commun. ACM*, **12**, 185–186.

Utilities

Fox, P., Hall, A. D., and Schryer, N. L. (1978). The PORT Mathematical Subroutine Library. *ACM. Trans. Math. Software*, **4**, 104–126. (Also Bell Laboratories, Murray Hill, N. J., *Computing Sci. Tech. Rep. No*. 47.)

Kernighan, B. W. and Plauger, P. J. (1976). *Software Tools*. Addison-Wesley, Reading, Mass.

See also COMPUTERS AND STATISTICS; STATISTICAL PACKAGES; and STATISTICAL SOFTWARE.

JOHN M. CHAMBERS

ALIAS

When two (or more) parameters affect the distribution of a test statistic* in similar ways, each is said to be an alias of the other(s). The term is especially associated with fractional factorial designs*, in the analysis of which certain sums of squares have distributions that can reflect the existence of any one, or some, of a number of different effects.

See also CONFOUNDING and FRACTIONAL FACTORIAL DESIGNS.

ALIAS GROUP. See Fractional Factorial Designs

ALIASING. See Confounding

ALIAS MATRIX. See Fractional Factorial Designs

ALLAN VARIANCE. See Successive Differences

ALLGEMEINES STATISTISCHES ARCHIV

The *Allgemeines Statistisches Archiv* is the Journal of the Deutsche Statistische Gesellschaft* (German Statistical Society) and began publication in 1890.

The journal provides an international forum for researchers and users from all branches of statistics. The first part (Articles) contains contributions to statistical theory, methods, and applications. There is a focus on statistical problems arising in the analysis of economic and social phenomena. All papers in this part are refereed. In order to be acceptable, a paper must either present a novel methodological approach or a result, obtained by substantial use of statistical methods, which has a significant scientific or societal impact. For further information, readers are referred to the parent society website, www.dstatg.de.

See also Statistische Gesellschaft, Deutsche.

ALLOKURTIC CURVE

An allokurtic curve is one with "unequal" curvature, or a skewed*, as distinguished from an isokurtic* curve (which has equal curvature and is symmetrical). This term is seldom used in modern statistical literature.

See also Kurtosis and Skewness: Concepts and Measures.

ALLOMETRY

It is rare, in nature, to observe variation in size without a corresponding variation in shape. This is true during the growth of an organism when radical shape changes are commonplace; when comparing different species from the same family; and even when comparing mature individuals from the same species. The quantitative study of this relationship between size and shape is known loosely as *allometry* and the main tool is the log-log plot. If X and Y are two dimensions that change with size, then the way each changes relative to the other is best studied by plotting $\log X$ vs. $\log Y$. In the past this was usually done on special log-log graph paper, but calculators have rendered such devices obsolete. Natural algorithms will be used in this article (and are recommended).

Some examples are shown in Figs. 1 to 3. In Fig. 1 the points represent different individuals, each measured at one point during growth. In Fig. 2 the points refer to mature individuals. In Fig. 3 the points refer to different species, and X and Y now refer to mean values (or some other typical values) for the species. The value of the log-log plot is that it provides a simple summary of departures from *isometric** size variation, i.e., variation in which geometric similarity is maintained. If X and Y are both linear dimensions, then isometric variation corresponds to a constant ratio Y/X, which in turn corresponds to a line

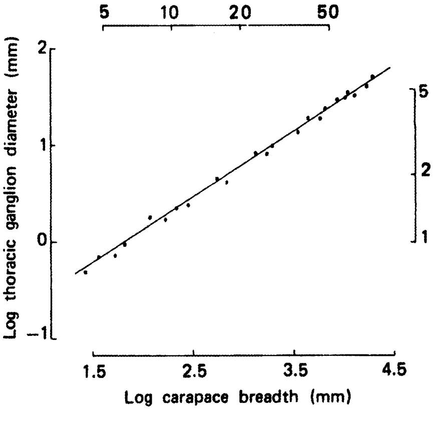


Figure 1. Growth of crabs (redrawn from Huxley [10]). The slope of the line is approximately 0.6.

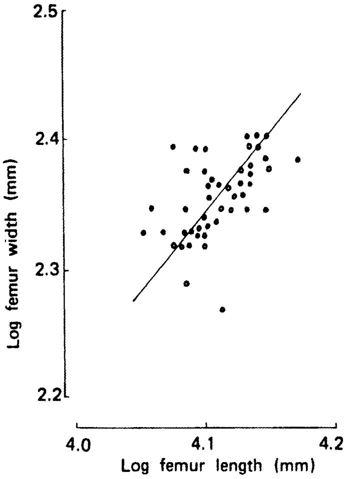


Figure 2. Variation between mature pine martens (redrawn from Jolicoeur [11]). The slope of the line is approximately 1.3.

of slope 1 on the log-log plot. If Y is linear but X is a volume (or weight), then isometric variation corresponds to a constant value for $Y/X^{1/3}$, i.e., a line of slope 0.33 in the log-log plot. In a similar fashion the slope is 0.50 when Y is linear and X an area and 0.67 when Y is an area and X a volume. Slopes of X on Y are the reciprocals of the slopes of Y on X.

The log-log plot was first used systematically by Huxley [9,10] and Teissier [23]. They found that for growth studies, using a wide variety of dimensions and organisms, the plot could often be adequately summarized by a straight line, i.e., a power law of the type $Y = bX^{\alpha}$ in original units. The coefficient α is the slope in the log-log plot and $\log b$ is the intercept: $\log Y = \alpha \log X + \log b$. For linear dimensions $\alpha = 1$ corresponds to isometry* and $\alpha \neq 1$ is referred to as allometric growth: positive if $\alpha > 1$ and negative if $\alpha < 1$. Departures from a simple straight-line relationship are not uncommon, the best known being where there is a sudden change in relative growth rate during development.

Some importance was initially attached to providing a theoretical basis for the power law, but today the relationship is recognized as being purely empirical. It is the "roughly linear relationship" so widely used in biology, but it happens to be in log units because this is the natural scale on which to study departures from isometric size variation. Linear plots using original units have also been used, but these have the disadvantage that isometry now corresponds to a line passing through the origin rather than a line of slope 1 (in log units), and there are usually no measurements in the region of the origin.

The log-log plot has also been widely used when comparing species in the general study of the effect of scale on form and function. Once again a linear relationship often proves adequate, but it is not only departures from isometry that are important but also deviations of individual species from the allometric line.

Gould [6] should be consulted for further details about the history of allometry and for a review of applications.

STATISTICAL METHODS FOR ESTIMATING A STRAIGHT LINE

The central problem is to estimate the straight-line relationship displayed in the log-log plot. To avoid too many "logs," the values of $\log X$ and $\log Y$ will be referred

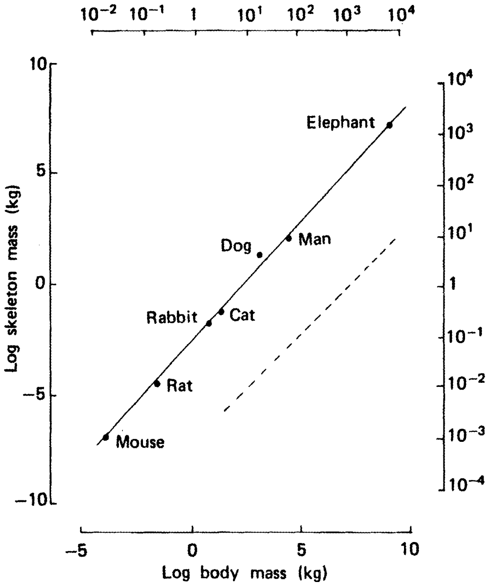


Figure 3. "Mouse-to-elephant" line (redrawn from Schmidt-Nielsen, p. 5, in Pedley [21]). The dashed line has slope 1.

to as x and y. The equation of a straight line passing through (x_0, y_0) with slope $\alpha(y$ on $x)$ is $y - y_0 = \alpha(x - x_0)$. This corresponds to the power law $Y = bx^{\alpha}$, where $\log b = y_0 - \alpha x_0$. Figure 4 shows four different approaches to fitting a line. In the first two the lines are chosen to minimize the sum of squares of deviations* $\sum d^2$ and correspond to regression* of y on x, and x on y. In the third the line minimizing $\sum d^2$ is called the major axis and in the fourth the value of $\sum d_1 d_2$ is minimized to produce the reduced major axis [14] or D-line [25]. All four lines pass through the centroid of points $(\overline{x}, \overline{y})$ and differ only in their values for α the slope of y on x. All the estimates may be expressed in terms of r, S_x, and S_y, the sample correlation* and standard deviations* of x and y. They are:

1. $\hat{\alpha} = rS_y/S_x$ (regression of y on x)
2. $\hat{\alpha} = r^{-1}S_y/S_x$ (regression of x on y)
3. $\hat{\alpha} =$ slope of the major axis (given below)
4. $\hat{\alpha} = S_y/S_x$ with sign the same as that of r.

The two regression estimates differ from one another and there is no natural way of resolving the question of which to choose. For this reason regression theory is not as helpful in allometry as in other branches of applied statistics. The slope of the major axis is found by first obtaining the positive root of the equation in t,

$$t/(1 - t^2) = r\lambda/(1 - \lambda^2), \qquad (1)$$

where $\lambda = S_y/S_x$ or S_x/S_y, whichever is less than 1. If $\lambda = S_y/S_x < 1$, then t is the slope in units of y on x. If $\lambda = S_x/S_y < 1$, then t is the slope in units of x on y and t^{-1} is the slope in units of y on x. When r is negative, the same numerical value is used for t but the sign is now negative.

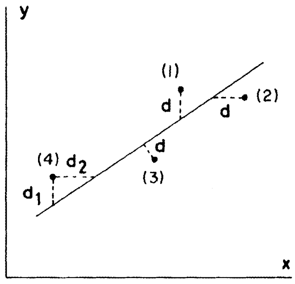


Figure 4. Four methods of fitting a straight line.

Of the last two estimates the most popular has been the reduced major axis. The major axis has been criticized because it depends on the ratio $\lambda = S_y/S_x$ in a nonlinear way so that a change of units for x and y does not affect t in the same way as it affects λ. This has no validity when x and y refer to log measurements because a change of units for the measurements X and Y leaves r, S_x, and S_y unchanged. The major axis has also been criticized as being more difficult to compute than the others, but in fact (1) is easy to solve either as a quadrate or graphically. For values of λ in the range 0.2 to 0.8 a good approximation is $t \simeq \lambda + 0.3 \log r$. As λ approaches 1, the right-hand side of (1) tends to ∞, so t tends to 1. There is no general agreement as to which of these two estimates is to be preferred. Fortunately, when r is high, they give very similar answers.

Statistical sampling properties of the estimates are based on the bivariate normal distribution*. Approximate large sample standard errors* are given by Kermack and Haldane [14]:

$$SE(\hat{\alpha}) \simeq \sqrt{\left(\frac{1 - r^2}{r^2 n}\right)} t$$

for the major axis

$$SE(\hat{\alpha}) \simeq \sqrt{\left(\frac{1 - r^2}{n}\right)} \frac{S_y}{S_x}$$

for the reduced major axis.

Since $t > S_y/S_x$, the standard error for the major axis is always greater than that for the reduced major axis (considerably greater if r is low). This is an argument in favor of the reduced major axis, but such an estimate would make little sense if r were too low. When $S_x = S_y$ and $r = 0$, it amounts to drawing a line of slope 1 through a circular cloud of points. At least the large sampling error of the major axis serves as a warning that the line is difficult to determine when r is low. As a numerical example, consider the case $r = 0.8$, $n = 20$, $\lambda = S_y/S_x =$

0.6. The slope of the major axis is found from $t/(1-t^2) = 0.75$, a quadratic equation yielding $t = 0.5352$ ($t = \tan\{\frac{1}{2}\tan^{-1}(2 \times 0.75)\}$ on a calculator). The slope of the reduced major axis is $\lambda = 0.6$ and the approximation for t is $\lambda + 0.3 \log 0.8 = 0.5331$. The two estimates for α are 0.54 ± 0.09 (major axis) and 0.60 ± 0.08 (reduced major axis). When $r = 0.4$ they become 0.33 ± 0.17 and 0.60 ± 0.12, respectively.

An alternative to basing estimates directly on the bivariate normal distribution* is to assume that x and y would lie exactly on a straight line if it were not for "errors." The errors are due to biological variation rather than to measurement errors* in this context. The model is called the linear functional relationship model. It is possible to estimate α provided that the ratio of error variances is assumed known. In fact, if they are assumed equal, the estimate turns out to be the major axis. However, no sampling theory based on the model is available and the approach has not proved very fruitful. Further details are given in Sprent [22] and Kendall and Stuart [13]. Bartlett's estimate of slope also belongs to this class [2]. It is an example of the use of an instrumental variate* (see Kendall and Stuart [13]). the estimate has enjoyed some popularity in allometry, but unfortunately the assumptions underlying its use are not usually fulfilled unless the scatter* is low. A detailed numerical example of the use of Bartlett's estimate has been given by Simpson et al. [21].

STATISTICAL TESTS

The most commonly required test is for departure from isometry. For both the major axis and reduced major axis, $\alpha = 1$ implies that $\sigma_y = \sigma_x$ (in terms of population parameters), and this is most conveniently tested for by computing $z = y - x$ and $w = y + x$ for each individual. Since $\mathrm{cov}(z, w) = \sigma_y^2 - \sigma_x^2$, the test for $\sigma_y = \sigma_x$ is equivalent to testing for zero correlation* between z and w. If the isometric value of α is 0.33 (for example) rather than 1, then $z = y - 0.33x$ and $w = y + 0.33x$ are used to test whether $\sigma_y = 0.33\sigma_x$ in the same way.

Confidence intervals* for α may be obtained using the approximate large sample standard errors for $\log\hat{\alpha}$. These are easily derived from those given earlier and are

$$SE(\log\hat{\alpha}) \simeq \sqrt{\left(\frac{1-r^2}{r^2 n}\right)}$$

for the major axis

$$SE(\log\hat{\alpha}) \simeq \sqrt{\left(\frac{1-r^2}{n}\right)}$$

for the reduced major axis.

There is some advantage to using the SE of $\log\hat{\alpha}$ rather than $\hat{\alpha}$ since the formulas do not involve the population value α. A better approximation for the reduced major axis has been given by Clarke [4]. See also Jolicoeur [12].

An important question in allometry is whether or not one group of individuals or species is an allometric extension of another. Figure 5 illustrates a case where this is so and also shows why the use of different lines can give different answers to this question. Group A is a linear extension of group B only along the major axis, not along the regression line of y on x. There seems to be no generally recommended statistical test for this situation. Analysis of covariance* would be used with regression lines,* and this suggests a rough test for use with the major axis. First obtain a pooled estimate* of the common slope of the axis in the two groups as $\hat{\alpha} = \frac{1}{2}(\hat{\alpha}_1 + \hat{\alpha}_2)$ and then test for differences between groups in a direction perpendicular to this slope, ignoring any sampling error in $\hat{\alpha}$. This is equivalent to calculating $z = y - \hat{\alpha}x$ for each individual (or species) in each group and then testing whether the mean of z differs between groups.

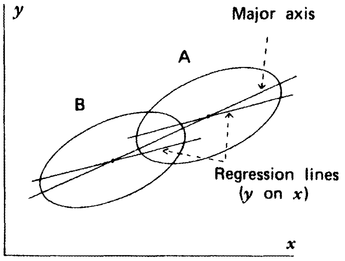


Figure 5. Allometric extension.

Statistical tests are often omitted in allometry and there are a number of reasons for this. One is that in many growth studies the random error* is small compared to the systematic change and the conclusions are apparent. Another is that where points represent species, which cannot be thought of as randomly sampled from a population, the relevance of statistical sampling theory is slight. Even in growth studies the individuals are rarely a proper random sample from a defined population, so that significance tests* play a less important role than they do in randomized experiments*. Finally, the inadequacy of statistical methods based on the bivariate normal distribution or the linear functional relationship model is an important factor.

MULTIVARIATE ALLOMETRY

When more than two measurements of an organism are studied, an overall view of the joint variation is desirable. For p measurements $X_1, \ldots, X_p$ the log-log plot generalizes to a plot of $x_1 = \log X_1, \ldots, x_p = \log X_p$ in p dimensions, a useful concept even though it is not possible to actually plot the points. Isometry corresponds to the direction vector $\boldsymbol{\alpha}_0 = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$ with the usual provision that if X_i represents an area or volume it is replaced by $X_i^{1/2}$ or $X_i^{1/3}$, respectively. Allometry again corresponds to departures from isometry, but clearly the possibilities are considerably wider in p dimensions than in two dimensions. A further problem is that there is no longer a large body of empirical evidence to suggest that departures from isometry are usually adequately summarized by a straight line. However, such a departure is a sensible starting point, and it may be summarized by a direction vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, where $|\boldsymbol{\alpha}| = 1$ (i.e., $\sum \alpha_i^2 = 1$). The angle between $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$ is given by $\cos\theta = \sum \alpha_i/\sqrt{p}$.

The direction $\boldsymbol{\alpha}$ may be estimated using either the major or the reduced major axis. In the latter $\hat{\alpha}$ is taken to be the direction vector proportional to $(S_1, \ldots, S_p)$ where S_i is the sample standard deviation of x_i. Using the major axis, $\hat{\alpha}$ is taken equal to the direction of the first eigenvector* of the covariance matrix* of $x_1, \ldots, x_p$. This eigenvector is often referred to as the first principal component*. Corruccini and Henderson [5] give a good example of the use of the major axis. The linear functional relationship model has also been generalized to many variables [8,22], but there are few examples of its actual use.

The statistical sampling theory for many variables encounters a major difficulty that does not exist in the bivariate case. This is the need to specify the error structure about the allometric line. In the bivariate case variation perpendicular to the line is in one dimension only and may be summarized by a standard deviation. In p dimensions the space perpendicular to the allometric line has $p - 1$ dimensions, so that covariation as well as variation must be specified. The simplest thing is to assume zero covariance perpendicular to the line, which implies that all eigenvalues* of the covariance matrix apart from the first are equal. A fairly straightforward test for this hypothesis has been described by Morrison [16, p. 250]. The eigenvector corresponding to the first eigenvalue is the allometric direction, and for this to be a sensible summary of the data the first eigenvalue must be considerably greater than the $p - 1$ (equal) eigenvalues. This is the same qualitative judgment that has to be made in the bivariate case. The first eigenvalue is the variance in the allometric direction and the others are the variances in directions perpendicular to the line. Variation along the line should be much greater than that about the line. Isometry corresponds to a further specialization of this situation to the case where the first eigenvector is $(1/\sqrt{p}, \ldots, 1/\sqrt{p})$. Kshirsagar [15] gives an overall χ^2 test* for both the isometric direction and independent variation about the line and also shows how to partition the χ^2 into its two components. Morrison [16] describes a test for the isometric direction which does not rest on the assumption that all other eigenvalues are equal. Anderson [1] should be consulted for the detailed derivation of the tests described in Morrison's book.

Mosimann [17,18] has proposed a more general approach which avoids the very restrictive assumption of independent variation orthogonal to $\boldsymbol{\alpha}_0$ or $\boldsymbol{\alpha}$. He defines size as a function G of the original measurements $\mathbf{X}$, with the property that $G(\mathbf{X}) > 0$ and $G(a\mathbf{X}) = aG(\mathbf{X})$ for $a > 0$. A shape vector $\mathbf{Z}(\mathbf{X})$

is defined to be any dimensionless vector with $p-1$ components. Isometry is then defined to be statistical independence* of shape and size, and it is shown that for a given choice of size, either all shape vectors are independent of size or none are. Mosimann also shows that if shape is independent of one measure of size, then it cannot be independent of any other measure of size.

A test for isometry of shape with respect to a given size can be carried out on a log scale using multiple regression* techniques to test independence*. For example, if $(x_1, \ldots, x_p)$ are the measurements on a log scale, then a common measure of size is $\overline{x} = \sum x_i/p$ and a possible shape vector is $(x_1 - \overline{x}, \ldots, x_{p-1} - \overline{x})$. If R^2 is the multiple correlation* between $\overline{x}$ and $(x_1 - \overline{x}, \ldots, x_{p-1} - \overline{x})$ based on n individuals, then $\{R^2(n-p)\}/\{(1-R^2)(p-1)\}$ may be used to test the independence of shape and size. Provided that $(x_1, \ldots, x_p)$ have a multivariate normal distribution*, the statistic has the F-distribution* with $p-1$, $n-p$ degrees of freedom when the null hypothesis is true. Mosimann and James [19] give an example of this test, but no attempt is made to quantify departures from isometry, and this seems to be a weakness of the approach so far.

The problem of whether one group of points is an allometric extension of another in p dimensions does not seem to have been dealt with explicitly. There have been some advances in the wider problem of making comparisons between groups, ignoring variation in certain given directions (e.g., the isometric direction). Burnaby [3] has extended discriminant analysis* to deal with this situation and Gower [7] has extended Burnaby's work to cover the case where the direction to be ignored is an estimated allometric direction.

REFERENCES

1. Anderson, T. W. (1963). *Ann. Math. Statist.*, **34**, 122–148.
2. Bartlett, M. S. (1949). *Biometrics*, **5**, 207–212.
3. Burnaby, T. P. (1966). *Biometrics*, **22**, 96–110.
4. Clarke, M. R. B. (1980). *Biometrika*, **67**, 441–446.
5. Corruccini, R. S. and Henderson, A. M. (1978). *Amer. J. Phys. Anthropol.*, **48**, 203–208.
6. Gould, S. J. (1966). *Biol. Rev.*, **41**, 587–640.
7. Gower, J. C. (1976). *Bull. Geol. Inst. Univ. Upps.* (N.S.), **7**, 1–10.
8. Hopkins, J. W. (1966). *Biometrics*, **22**, 747–760.
9. Huxley, J. S. (1924). *Nature (Lond.)*, **114**, 895–896.
10. Huxley, J. S. (1932). *Problems of Relative Growth*. Methuen, London.
11. Jolicoeur, P. (1963). *Growth*, **27**, 1–27.
12. Jolicoeur, P. (1968). *Biometrics*, **24**, 679–682.
13. Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*, Vol. 2., 2nd ed., Charles Griffin, London.
14. Kermack, K. A. and Haldane, J. B. S. (1950). *Biometrika*, **37**, 30–41.
15. Kshirsagar, A. M. (1961). *Biometrika*, **48**, 397–407.
16. Morrison, D. F. (1976). *Multivariate Statistical Methods* (2nd ed.), McGraw-Hill, New York.
17. Mosimann, J. E. (1970). *J. Amer. Statist. Ass.*, **65**, 930–945.
18. Mosimann, J. E. (1975). Statistical problems of size and shape, I, pp. 187–217. Statistical problems of size and shape, II, pp. 219–239. In *Statistical Distributions in Scientific Work*, Vol. 2, G. P. Patil, S. Kotz, and J. K. Ord, eds. D. Reidel, Dordrecht.
19. Mosimann, J. E. and James, F. C. (1979). *Evolution*, **33**, 444–459.
20. Pedley, T. J. (1977). *Scale Effects in Animal Locomotion*. Academic Press, New York.
21. Simpson, G. G., Roe, A., and Lewontin, R. C. (1960). *Quantitative Zoology* (rev. ed.). Harcourt Brace, New York.
22. Sprent, P. (1969). *Models in Regression and Related Topics*. Methuen, London.
23. Teissier, G. (1931). *Trav. Stn. Biol. Roscoff*, **9**, 227–238.
24. Teissier, G. (1948). *Biometrics*, **4**, 14–53.

FURTHER READING

D'Arcy Thompson's book *Growth and Form* (Cambridge University Press, Cambridge, 1917, 1942) provides the standard introduction to problems of scale in form and function. An account of some recent work is given in *Scale Effects in Animal Locomotion* by T. J. Pedley (Academic Press, New York, 1977). S. J. Gould's article "Allometry and size in ontogeny and phylogeny" (*Biol. Rev.*, **41**, 587–640) gives a comprehensive survey of the uses of allometry, and Sprent's 1972 article "The mathematics of size and shape" (*Biometrics*, **28**, 23–37) reviews the mathematical and statistical side. D. L. Pilbeam and S. J. Gould, in their 1974 contribution "Size and scaling in human evolution" (*Science*, **186**, 892–901), give an interesting account of the use of allometry to investigate evolutionary relationships. A variety of different allometric studies are reported in *Problems of Relative Growth* by J. S. Huxley (Methuen, London, 1932), and a good discussion of some of the problems with allometry is given in *Essays on Growth and Form Presented to D'Arcy Wentworth Thompson*, edited by W. E. Le Gros Clark and P. B. Medewar (Clarendon Press, Oxford, 1945).

See also Anthropology, Statistics in; Principal Component Analysis, Generalized; Regression (Various); and Size and Shape Analysis.

M. Hills

ALMOST CERTAIN CONVERGENCE.

See Convergence of Sequences of Random Variables

ALMOST-LACK-OF-MEMORY (ALM) DISTRIBUTIONS

A random variable X is defined to have an almost-lack-of-memory (ALM) distribution [1] if the Cauchy functional equation

$$\Pr(X - b < x | X \geqslant b) = \Pr(X < x) > 0,$$

here given in probabilistic form, holds for infinitely many values of $b, b > 0$. Continuous distributions with this property possess a density $f(\cdot)$, almost surely, that satisfies (see Ref. 1)

$$f_X(nc + x) = \alpha^n (1 - \alpha) f(x), \quad 0 < x < c, \quad (1)$$

$0 < \alpha < 1, n = 0, 1, 2 \ldots$, for any arbitrary positive integer c.

Let $S(\cdot)$ be a survival function defined in $[0, \infty)$, so that $S(x) = 1 - F(x)$ for a cumulative distribution function (cdf) $F(\cdot)$, and for c, a positive integer, the functional equation

$$S(nc + x) = S(nc)S(x), \quad x \geqslant 0, \quad (2)$$

$n = 1, 2, \ldots$, is a particular case of the ALM property (1). If (2) holds for a nonnegative random variable with cdf $F(x) = \Pr(X \leqslant x)$, $x > 0$, then [2] $F(\cdot)$ is of the form

$$F(x) = 1 - \alpha^{[x/c]} + \alpha^{[x/c]} F(x - [x/c]c), \quad x \geqslant c > 0, \quad (3)$$

where $\alpha = S(c) < 1$ and $[x]$ is the smallest integer less than or equal to x. Hence, each cdf $F(\cdot)$ satisfying Equation 1 is uniquely determined by the values of $F(x)$ for $0 \leqslant x < c$.

As an application, consider a queueing system with instantaneous repairs after any failure of a constant-lifetime server. Then each cdf satisfying Equation 3 is the cdf of the "blocking time," that is, the total time taken by a customer, and conversely [1,2].

REFERENCES

1. Chukova, S. and Dimitrov, B. (1992). On distributions having the almost-lack-of-memory property. *J. Appl. Probab.*, **29**, 691–698.
2. Lin, G. D. (1999). Letter to the editor. *J. Appl. Probab.*, **31**, 595–605.

See also Exponential Distribution and Queueing Theory.

α-LAPLACE DISTRIBUTION.

See Linnik Distribution

ALTERNATIVE HYPOTHESIS

A hypothesis that differs from the hypothesis being tested is an alternative hypothesis, usually one to which it is hoped that the test used will be sensitive. Alternative hypotheses should be chosen having regard to (1) what situation(s) are likely to arise if the hypothesis tested is not valid and (2) which ones, among these situations, it is of importance to detect.

Usually, a whole class of alternative hypotheses, rather than a single one, is used.

See also Critical Region; Hypothesis Testing; Level of Significance; Null Hypothesis; and Power.

AMERICAN JOURNAL OF HUMAN GENETICS

The American Journal of Human Genetics (AJHG) was founded in 1948, and is published monthly by the University of Chicago Press in two volumes per year, six issues per volume. The website for the journal is www.journals.uchicago.edu/AJHG

As stated on the website:

> "*AJHG* is [sm1]a record of research and review relating to heredity in humans and to the application of genetic principles in medicine, psychology, anthropology, and social services, as well as in related areas of molecular and cell biology. Topics explored by *AJHG* include behavioral genetics, biochemical genetics, clinical genetics, cytogenetics, dysmorphology, genetic counseling, immunogenetics, and population genetics and epidemiology".

AJHG is the official scientific publication of the American Society of Human Genetics.

AMERICAN SOCIETY FOR QUALITY (ASQ)

[This entry has been updated by the Editors.]

The United States is generally considered the country that founded modern quality control. The work of Walter Shewhart* of the American Telephone and Telegraph Company and his associates George Edwards, Harold Dodge, and Harry Romig forms the nucleus around which the movement grew. However, it was the crisis of World War II that gave impetus to the field. Business managers realized that government-mandated quality control programs for defense products had an equally important application in civilian products.

During World War II a series of short courses were conducted throughout the country on statistical quality control*. Those who attended were encouraged to get together to exchange ideas and to reinforce their new-found knowledge. A series of local societies and several regional ones were founded throughout the country. In 1946, they formed a confederation called the American Society for Quality Control, which acts as the primary professional society for the United States, Mexico, and Canada. The headquarters office, initially in New York, was later transferred to Milwaukee, Wisconsin, where it is now located.

In 1997 the society dropped 'Control' from its title, and, as the American Society for Quality (ASQ), adopted a new mission of promoting performance excellence worldwide.

The ASQ has encouraged community colleges and universities to offer courses in the field. Despite the demand, most institutions of higher education choose not to have separate curricula; rather, courses are woven into other curricula.

The society is organized in modified matrix fashion. There are in 2004 more than 250 local geographic sections throughout major industrial areas of the United States, Canada and Mexico. Some members also choose to join one or more of the 27 industrial or subdiscipline-oriented divisions. They usually hold one or more national conferences per year. The divisions are listed on the ASQ's website, www.asq.org.

By 1991 the membership of ASQ exceeded 100,000, and in 2001 membership extended to 122 countries. Members are classified as Members, Senior Members, or Fellows.

Honors are presented by the society at the national, divisional, and sectional levels. Best known are the Shewhart Medal, the Edwards Medal, the Grant Award, and the Brumbaugh Award.

The initial journal of ASQC was *Industrial Quality Control*, later renamed *Quality Progress*. The *Journal of Quality Technology** has been published since 1969. *Technometrics** is a journal published jointly with the American Statistical Association. In addition, the *Transactions* of the Annual Technical Conference have been found to be a useful source of practical information. *Quality Management Journal* and *Software Quality Professional* began publication in 1993 and 1998, respectively. Many other publications are published centrally and by the divisions, conference boards, and local sections.

In the early days of the ASQC the emphasis was primarily on the statistical basis of sampling schemes* for raw materials coming into a plant or warehouse. There was also interest in reducing the cost of final inspection. The early emphasis was on detection of nonconforming products. A bit later the emphasis changed to the prevention of defects. Shewhart's work in the economic control of quality* in manufacturing provided the basis for this work. Today, managerial, motivational, and engineering aspects get a more balanced hearing. The ASQ is the primary vehicle for teaching these concepts. It has an important educational arm which conducts courses in quality and reliability throughout North America. In addition, educational materials are available for persons wishing to conduct courses under local auspices. Examinations for certification as Certified Quality Engineer (CQE), Certified Reliability Engineer (CRE), Certified Quality Technician (CQT), Certified Quality Manager (CQM), and Certified Reliability Technician (CRT) are available throughout the world for both members and nonmembers.

The savings due to reliable quality control programs initiated since World War II approached $1 trillion worldwide.

In later years ASQ became involved in coordinating national quality standards on behalf of the American National Standards Institute. These have worldwide impact.

The society maintains close relations with societies associated with quality in various parts of the world. These include the European Organization for Quality, the Japanese Union of Scientists and Engineers, the New Zealand Organization for Quality Assurance, the Australian Organization for Quality Control, and a variety of others.

See also *Journal of Quality Technology*; Quality Control, Statistical; and *Technometrics*

W. A. Golomski

AMERICAN SOCIETY FOR QUALITY CONTROL (ASQC).
See American Society for Quality (ASQ)

AMERICAN STATISTICAL ASSOCIATION

[This entry has been updated by the Editors.]

The American Statistical Association, founded in 1839 as a nonprofit corporation, has as its purpose "to foster, in the broadest manner, statistics and its applications, to promote unity and effectiveness of effort among all concerned with statistical problems, and to increase the contribution of statistics to human welfare." It is a professional association whose membership is open to individuals with interest and background in the development and use of statistics in both methodology and application.

The Association is governed by an elected board of directors, which represents geographical areas, subject-matter areas, and the offices of President, Past President, President-elect, Secretary-treasurer, and three Vice-presidents. There is also a council made up of representation from each of the chapters of the association as well as the sections, as noted below. There are presently (in 2004) 78 chapters. Membership at the end of 2003 exceeded 17,000 in the U.S., Canada and other countries.

A central office is maintained at 1429 Duke Street, Alexandria, VA - 22314-3415; tel. (703) 684-1221. The staff includes an Executive Director, Director of Programs, Director of Operations, and support staff. Technical editing for some journals is maintained at this office.

The chapters of the ASA have individual programs throughout the year. These vary from one or two to as many as 20 meetings in a single year. Each chapter is autonomous in its program of activities. Chapters vary in size from 25 to 1800 members. A chapter may sponsor a seminar of one or more days or a short course in a subject-matter area or a set of methodological techniques. Chapters often collaborate with local or regional units of other professional associations.

Within ASA there are 21 sections. These represent different areas of activity and include Bayesian Statistical Science, Government Statistics, Health Policy Statistics, Nonparametric Statistics, Quality and Productivity, Risk Analysis, Statistical Consulting, Statistical Graphics, Statistics and the Environment, Statistics in Defense and National Security, Statistics in Epidemiology, Statistics in Sports, Business and Economics, Social Statistics, Statistical Education, Physical and Engineering Sciences, Biometrics, Statistical Computing, Survey Research Methods, Biopharmaceutics, Teaching of Statistics in the Health Sciences, and Statistics in Marketing. The sections develop various activities which are useful to statistical practitioners and researchers. These include (1) cosponsorship of regional meetings with other professional associations, (2) review of statistical computer packages, (3) development of visual-aids materials, (4) appraisal of surveys, and (5) recommendations of statistical curricula in health sciences, computer sciences, industrial statistics, etc. Sections cosponsor symposia, workshops, and special topic meetings. A number of the sections participate in advisory capacities in such areas as national and international standards, educational programs, and federal statistical activities.

One of the strengths of the association is its committees, which vary from advisory boards to government agencies to joint committees with other professional associations. There are approximately 90 committees within the association. Although some are the usual "in-house" committees and some are short-term ad hoc committees, 23 relate to activities and programs outside the statistical profession, as well as to key issues and concerns on national and international levels. For example, the Committee on Law and Justice Statistics reviews the programs and structure of statistical data collection and analysis in the U.S. Department of Justice. Advisory committees serve the U.S. Bureau of the Census* and the Energy Information Administration of the Department of Energy. Possibilities for similar committees are continually being studied. The areas of privacy and confidentiality, as well as the statistical components of legislative action, are serviced by active committees. For the latter, a consortium of 10 professional associations was formed. A joint committee of the American Statistical Association and the National Council of Teachers of Mathematics works on developing the teaching of Statistics and Probability at every level of school education from grades K-12. Some activities are in close communication with similar groups in other countries.

The association holds an annual meeting. Often, this is in cooperation with other statistical and related societies, e.g., the International Biometric Society*, Eastern and Western North American Regions, and the Institute of Mathematical Statistics*. This is usually preceded by one- or two-day short courses in statistical methodology. There is also joint sponsorship in national and regional activities with such professional organizations as the American Society for Quality (ASQ)*, the Society of Actuaries, etc.

The association engages in programs of research and development in various areas of statistics in the types of activities that would not ordinarily be considered appropriate for a particular university or research institution. At times, the association engages in national and international projects in an attempt to make statistics even more useful to the profession and to the public at large. An example of this was the international seminar on the Transfer of Methodology between Academic and Government Statisticians.

Some of the programs in which the ASA has been engaged have far-reaching consequences in the areas of social science, survey research methods, physical and engineering sciences, and health sciences, to state a few. For example, beginning in 1977, the association engaged in a joint venture with the U.S. Bureau of the Census, "Research to Improve the Social Science Data Base." This included research fellows and trainees in time series*, demography*, computer software, editing* of large data sets, and a number of related areas. Another research program concerns the quality of surveys.

Educational programs, often in conjunction with other professional associations, include the development of resource material and visual aids for the teaching of statistics*, a visiting lecturer program, development and testing of a statistical curriculum for secondary schools, and short courses. The association also sponsors research fellowships and traineeships in which an individual spends a major portion of a year at a federal statistical agency. In the area of continuing education, videotapes of short courses and special lectures are made available to a wide segment of professional statisticians, not only in ASA chapters, but at universities and in industrial organizations.

In 1976, in collaboration with ASA members in several Latin American countries, the association initiated a program of symposia and site visits to stimulate the development of statistics in these countries. The result has been increased professional activity and the development of new statistical associations in some of the host countries.

The American Statistical Association is affiliated with a number of national and international organizations. It is one of approximately 50 national statistical associations affiliated with the International Statistical Institute*. Furthermore, it maintains representation in sections of the American Association for the Advancement of Science (AAAS). Through its representatives the ASA is active in some AAAS programs, such as scientific freedom and human rights and the international consortium of professional associations. The ASA maintains representation on the councils of the American Federation of Information Processing Societies, the Conference Board of the Mathematical Sciences, the Social Science Research Council, and the National Bureau of Economic Research.

The Association publishes seven journals; *Journal of the American Statistical Association**, *The American Statistician**, *Technometrics** (jointly with the American Society for Quality Control), the *Journal of Agricultural, Biological and Environmental Statistics** (jointly with the International Biometric Society*), the *Journal of Business and Economic Statistics**, the *Journal of Computational and Graphical Statistics* (jointly with the Institute of Mathematical Statistics* and Interface Foundation of North America), and the *Journal of Educational and Behavioral Statistics** (jointly with the American Educational Research Association). ASA also publishes *Current Index to Statistics** (with the Institute of Mathematical Statistics*) and the three magazines *Amstat News, Chance* and *Stats*, the last-named reaching more than 3,000 student members of ASA. Beyond this, there are proceedings of its annual meetings by sections and various reports on conferences, symposia, and research programs.

The American Statistical Association is dedicated to be of service to its members abroad as well as its members within the United States and Canada, and has from time to time assisted other associations in the development of programs. Its services extend to the entire profession and to society in general. New programs are developed as needs become evident. Occasionally, the Association is called upon to provide an independent review or appraisal of statistical programs or the statistical content of a critical program in science or technology.

The website for ASA and its publications is www.amstat.org.

Fred C. Leone

AMERICAN STATISTICIAN, THE

[This entry has been updated by the Editors.]

The American Statistician is one of the principal publications of the American Statistical Association* (ASA). The *Journal of the American Statistical Association** (*JASA*) is devoted largely to new developments in statistical theory and its extensions to a wide variety of fields of application. *The American Statistician* emphasizes the professional development of ASA members by publishing articles that will keep persons apprised of new developments in statistical methodology through expository and tutorial papers dealing with subjects of widespread interest to statistical practitioners and teachers. It is published quarterly.

The origins of *The American Statistician* can be traced to the *American Statistical Association Bulletin*, which, from 1935 until 1947, served as the organ of information relating to ASA chapter activities, members, annual meetings, and employment opportunities for statisticians. At the end of World War II, the need was recognized for a more ambitious publication, providing expanded news coverage as well as serving the needs for professional development, and publication of *The American Statistician* was authorized. The first issue appeared in August 1947.

The contents of the first issue provide a preview of the major emphases that prevailed for many years. A major portion of the issue (11 of a total of 25 pages) was devoted to news items, including information about ASA programs, government statistics, other statistical societies, chapter activities, and news about members. The articles dealt with statistical applications in engineering and process control ("Statistical Engineering" by Tumbleson), electronic computing ("New High-Speed Computing Devices" by Alt), and the teaching of statistics ("A Well-Rounded Curriculum in Statistics" by Neiswanger and Allen). In addition, a note by Haemer on graphic presentation, the first in a series, appeared, and a "Questions and Answers" department to serve as a consulting forum, edited by Mosteller, was begun. The early emphasis on teaching of statistics continued throughout the period 1947–1979. Similarly, the early recognition of electronic computation was matched by continuing attention to statistical computing, with particular emphasis on this area found since the mid-1970s. The current "Statistical Computing and Graphics" section dates back to 1974.

Teaching of statistics has always been a major focus of *The American Statistician*. The section "The Teacher's Corner" was begun in 1962 and is still current in 2004. Other sections are "Statistical Practice", "General", "Reviews of Books and Teaching Materials", "Statistical Computing and Graphics", and "Letters".

The professional interests of statisticians were the focus of several early papers. In the 1970s, renewed interest in statistical consulting led to more frequent articles on this subject. It is interesting to note that an article "The Outlook for Women is Statistics," by Zapoleon, appeared in 1948.

Occasionally, special symposia papers have been published, such as a series of papers on "Reliability and Usability of Soviet Statistics" (1953) and the papers of a symposium on unemployment statistics (1955). During 1981, the "Proceedings of the Sixth Symposium on Statistics and the Environment" was published as a special issue of *The American Statistician*.

The year 1974 represented a major turning point for *The American Statistician*. Beginning in that year, all news items were moved to the new ASA publication *Amstat News*. In 1973, these items accounted for about one-third of the journal's pages, and a drop in the number of pages published from 1973 to 1974 corresponds largely to this shift. Since 1974, *The American Statistician* has devoted its pages exclusively to papers and notes on statistics.

The special departments of *The American Statistician* varied during the period 1947–1980. For example, the two initial departments "Questions and Answers" (a statistical consulting* forum) and "Hold That Line" (concerned with graphic presentation*) were terminated in 1953 and 1951 respectively. The department "Questions and Answers" was revived in 1954, but was modified from a consulting forum to a section containing essays on a variety of statistical subjects. The new "Questions and Answers" department was edited by Ernest Rubin from 1954 until 1973, when the department was terminated. Rubin wrote many of the essays published during this period.

In 1948, an editorial committee of six persons was organized to assist the editor. In 1954, this committee gave way to seven associate editors, a number that grew to 31 by the end of 2004.

The editorial review process for manuscripts is similar to that used by many other professional journals. All manuscripts that are potentially suitable for the journal are assigned to an associate editor for review.

The publication policy for *The American Statistician* is developed by the ASA Committee for Publications. The policy in effect during the late 1970s called for articles of general interest on *(i)* important current national and international statistical problems and programs, *(ii)* public policy matters of interest to the statistical profession, *(iii)* training of statisticians, *(iv)* statistical practice, *(v)* the history of statistics*, *(vi)* the teaching of statistics, and *(vii)* statistical computing. In addition, expository and tutorial papers on subjects of widespread interest to statistical practitioners and teachers are strongly encouraged.

The website for the journal can be accessed via that for the ASA, www.amstat.org

JOHN NETER

AMSTAT NEWS. See *American Statistician, The*

ANALYSIS OF COVARIANCE

The analysis of covariance is a special form of the analysis of variance* and mathematically need not be distinguished from it, although there are differences in utilization. (Any reader who is unfamiliar with the analysis of variance is advised to read the article on that topic before proceeding.) Using the analysis of covariance, an experiment or other investigation is planned and the analysis of variance is sketched out, based upon the model

$$\mathbf{y} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\eta},$$

where $\mathbf{y}$ is the vector of n data and $\boldsymbol{\eta}$ of n independent residuals*, $\boldsymbol{\theta}$ is a vector of p parameters, and $\mathbf{M}$ is an $n \times p$ matrix relating to the data to the parameters. It is then realized that there are q variates that could be measured which might explain some of the variation in $\mathbf{y}$, so the model is extended to read

$$\mathbf{y} = \mathbf{M}\boldsymbol{\theta} + \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta} = \left(\mathbf{M} \vdots \mathbf{D}\right) \begin{pmatrix} \boldsymbol{\theta} \\ \cdots \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\eta},$$

where $\mathbf{D}$ is an $n \times q$ matrix of supplementary data and $\boldsymbol{\beta}$ is a vector of q regression coefficients*, one appropriate to each variate. By this extension it is hoped to improve the estimate of $\boldsymbol{\theta}$. Following the nomenclature usual in correlation* and regression*, the values of y make up the dependent variate and those in the columns of $\mathbf{D}$ the independent variates.

As has been said, that is not different in its essentials from an ordinary analysis of variance. Thus there is nothing novel in introducing parameters that are not themselves under study but might serve to explain irrelevant variation. The blocks of an experimental design* will serve as an example. Further, the effect of blocks can be removed in either of two ways. If there were three of them, three block parameters could be introduced in θ, probably with some implied constraint to reduce them effectively to two. Alternatively, two independent variates, x_1 and x_2, could be introduced, such that x_1 was equal to $+1$ in block I, to -1 in block II, and to 0 in block III, while x_2 took the values $+1$, $+1$, and -2 in the three blocks, respectively. The outcome would be the same. Where a variate is thus derived from characteristics of the design rather than from measurement it is called a "pseudo-variate"*. The device is one that links the analyses of variance and covariance as a single technique.

Nevertheless, the user will continue to see them as different, usually thinking of the analysis of variance as the form visualized at the inception of the investigation and of the analysis of covariance as a means of coping with accidents and afterthoughts.

HISTORY AND DEVELOPMENT

The idea of allowing for an independent variate originated with Fisher* [4], who unfortunately did not appreciate that a covariance adjustment necessarily introduces nonorthogonality. The derivation of standard errors* of means is due to Wishart [6]. Bartlett [1] considerably extended the usefulness by introducing pseudo-variates for incomplete data. Later developments have tended to assimilate the method to the analysis of variance.

FORM OF THE CALCULATIONS

Nowadays, there are numerous computer packages able to carry out the necessary calculations. Nevertheless, some more detailed understanding of them can be helpful.

To take first the case of only one independent variate, the analysis of variance for its data, x, will give a sum of squared deviations* for error that has a quadratic form, i.e., it can be written as $\mathbf{x}'\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is some positive semidefinite matrix derived from the design. A corresponding quantity, $\mathbf{y}'\mathbf{H}\mathbf{y}$, exists for the dependent variate, $\mathbf{y}$. It will also be necessary to know $\mathbf{y}'\mathbf{H}\mathbf{x} = \mathbf{x}'\mathbf{H}\mathbf{y}$. Then, for a single independent variate, β, the regression coefficient of y on x, equals $\mathbf{y}'\mathbf{H}\mathbf{x}/\mathbf{x}'\mathbf{H}\mathbf{x}$. Also, the sum of squared deviations is reduced from $\mathbf{y}'\mathbf{H}\mathbf{y}$ with f degrees of freedom to $\mathbf{y}'\mathbf{H}\mathbf{y} - (\mathbf{y}'\mathbf{H}\mathbf{x})^2/(\mathbf{x}'\mathbf{H}\mathbf{x})$ with

$(f-1)$ when all values of y are adjusted to a standard value of x. This new mean-squared deviation* will be written as σ^2.

In the analysis of covariance the variation in x is regarded as a nuisance because it disturbs the values of y, the variate actually under study. Accordingly, any mean of y, e.g., a treatment mean*, is adjusted to a standard value of x. If the corresponding mean of x differs from this standard by d, the mean of y needs to be adjusted by βd. Similarly, if a difference of means of y is under study and the corresponding means of x differ by d, the same adjustment needs to be applied to make the y-means comparable.

An adjustment of βd will have a variance of $\sigma^2 d^2/(\mathbf{x}'\mathbf{H}\mathbf{x})$. If no adjustment had taken place, the variance of the y-mean (or difference of y-means) would have been, say, $A(\mathbf{y}'\mathbf{H}\mathbf{y})/f$, where A is a constant derived from the design. After adjustment the corresponding figure is $[A + d^2/(\mathbf{x}'\mathbf{H}\mathbf{x})]\sigma^2$, which is not necessarily a reduction, although sometimes the advantage will be considerable.

These results are readily generalized to cover p independent variates. Let $\mathbf{C}$ be a $(p+1)\times(p+1)$ matrix; the first row and the first column relate to the dependent variate and the others to the independent variates taken in some standard order. The element in the row for variate u and the column for variate v is $\mathbf{u}'\mathbf{H}\mathbf{v}$. Then writing $\mathbf{C}$ in partitioned form,

$$\mathbf{C} = \begin{pmatrix} \mathbf{Y} & \mathbf{P}' \\ \mathbf{P} & \mathbf{X} \end{pmatrix},$$

the new error sum of squared deviations is $Y - \mathbf{P}'\mathbf{X}^{-1}\mathbf{P}$ with $(f-p)$ degrees of freedom, thus giving σ^2, and the vector of regression coefficients, β, is $\mathbf{X}^{-1}\mathbf{P}$. If an adjustment of $\boldsymbol{\beta}'\mathbf{d}$ is applied to a mean, it will have a variance of $\mathbf{d}'\mathbf{X}^{-1}\mathbf{d}\sigma^2$.

Some special points need attention. For example, in the analysis of variance some lines, e.g., the treatment line in the analysis of data from an experiment in randomized blocks*, can be obtained directly without a second minimization. This is not so when covariance adjustments are introduced; it is necessary first to find $E = \mathbf{y}'\mathbf{H}\mathbf{y} - (\mathbf{x}'\mathbf{H}\mathbf{y})^2/\mathbf{x}'\mathbf{H}\mathbf{x}$ and then to ignore treatments in order to find $\mathbf{y}\mathbf{H}'_0\mathbf{y}$, $\mathbf{x}'\mathbf{H}_0\mathbf{y}$, and $\mathbf{x}'\mathbf{H}_0\mathbf{x}$, where $\mathbf{H}_0$ is some other matrix, and to attribute $(E_0 - E)$ to treatments, where $E_0 = \mathbf{y}'\mathbf{H}_0\mathbf{y} - (\mathbf{x}'\mathbf{H}_0\mathbf{Y})^2/\mathbf{x}'\mathbf{H}_0\mathbf{x}$; i.e., it is necessary to allow for the possibility that β_0, the regression coefficient when treatments are included with error, will be different from β, the regression coefficient* based on error alone. Some have argued against this complication on the grounds that the two cannot really be different, but reflection shows that they could be. In an agricultural field experiment, for example, the error may derive chiefly from differences in available nutrient from one part of the field to another. If treatments are the exaggeration of such differences by fertilizer applications, all may be well, but if they are something entirely different, such as pruning or a change of variety, it is not reasonable to expect that β_0 will be the same as β. Now that computation is so easy, the complication should be accepted at all times. It is, in any case, required by the mathematics.

Similar complications arise in split-plot* situations, because the two regression coefficients derived from the two error lines, one for main plots and the other from subplots, are often different. Since the two errors can be made up from quite different sources, i.e., the balance of different kinds of uncontrolled variation depends on the plot size, a comparison of the two regression coefficients can be illuminating. The difficulties are chiefly those of presentation and are much relieved if an intelligible explanation can be given for why adjustments, which may necessarily depend upon different regressions, behave as they do.

NUMERICAL EXAMPLE

The experiment described in the article on the analysis of variance has an available independent variate, namely x, the number of boxes of fruit, measured to the nearest tenth of a box, for the four seasons previous to the application of treatments. Full data are set out in Table 1. There need be no difficulty about the sums of squared deviations. The sums of products of deviations are here found simply by multiplying corresponding deviations from each plot and adding. In general, wherever in the calculation of sums of squared deviations a function of x or y

Table 1. Yields from a Soil Management Trial on Apple Trees

	Block[a]							
	I		II		III		IV	
Treatment	x	y	x	y	x	y	x	y
A	8.2	287	9.4	290	7.7	254	8.5	307
B	8.2	271	6.0	209	9.1	243	10.1	348
C	6.8	234	7.0	210	9.7	286	9.9	371
D	5.7	189	5.5	205	10.2	312	10.3	375
E	6.1	210	7.0	276	8.7	279	8.1	344
S	7.6	222	10.1	301	9.0	238	10.5	357

[a] x represents boxes of fruit per plot in the 4 years preceding the application of treatments; y, the crop weight in pounds during the 4 years following.

Source: These data were first presented by S. C. Pearce [5] and have been considered in some detail by D. R. Cox [3] and C. I. Bliss [2].

is squared, the sum of products of deviations is found by multiplying the function of x by the corresponding function of y. To cope with the plot feigned to be missing, it will be convenient to use a pseudo-variate, w, that has the value 1 for treatment A in block 1, and 0 elsewhere. Once that is done it does not matter what values for x and y are assigned to the missing plot*. Here, where it is intended to calculate the analysis with the plot first included and later excluded, it will be convenient to use the actual values throughout—8.2 and 287, respectively.

Allowing only for blocks, sums of squares and products are:

	y	w	x
y	24,182	51.50	710.1
w	51.50	0.8333	1.100
x	710.1	1.100	31.63

Allowing for treatments as well, they are:

	y	w	x
y	23,432	42.75	688.3
w	42.75	0.6250	0.958
x	688.3	0.958	24.23

Ignoring w and the missing plot for the moment, the sum of squared deviations allowing only for blocks is

$$24{,}182 - (710.1)^2/31.63 = 8240$$

with 19 degrees of freedom. Allowing for treatments also, it is

$$23{,}432 - (688.3)^2/24.23 = 3880$$

with 14 degrees of freedom. The analysis of variance now reads

Source	d.f.	Sum of Squares	Mean Square	F
Treatments	5	4360	872	3.15
Error	14	3880		
Treatments + error	19	8240		

The new picture is very different. The F-value* of 3.15 is significant ($P < 0.05$). Clearly, the independent variate has effected an improvement. If there is any doubt, an analysis of variance establishes it, namely,

Source	d.f.	Sum of Squares	Mean Square	F
Regression	1	19,552	19,552	70.58
Error	14	3,880	277	
Regression + error	15	23,432		

The regression coefficient is $688.3/24.23 = 28.41$. Adjusting treatment means for y to a standard value of $x = 8.308$ (its mean), that for A is

$$274.5 - 28.41(8.450 - 8.308) = 280.5$$

and for **S**

$$279.5 - 28.41(9.300 - 8.308) = 251.3.$$

The last figure shows a large effect of the adjustment. It appears that the randomization* had assigned treatment **S** to some heavily cropping trees, as the values of x show, and it had therefore shown up better than it should. To take the difference between the adjusted means of treatments **A** and **S**, i.e., $29.2 = 280.5 - 251.3$, the standard error is

$$\sqrt{277\left\{\frac{1}{4}+\frac{1}{4}+\frac{(9.300-8.450)^2}{24.23}\right\}}$$
$$= 12.1,$$

which suggests that A is, in fact, a more fruitful treatment than S. The other treatments can be investigated in the same way. The covariance adjustment has had two beneficial effects. It has markedly reduced the error variance and it has given better-based treatment means.

The case of the "missing plot" could have been dealt with in the same way. Adjusting y by w instead of x, the error sum of squared deviations would have been $23{,}432 - (42.75)^2/0.6250 = 20{,}508$ with 14 degrees of freedom, as before. For treatments and error together, the corresponding figure would have been $24{,}182 - (51.50)^2/0.8333 = 20{,}999$ with 19 degrees of freedom, which leaves 491 with 5 degrees of freedom for treatments. The error variance is now $1465 = (20{,}508/14)$. The method has the advantage of giving easily a figure for the standard error of the difference between treatment means. Thus, that for treatment A and any other treatment is

$$\sqrt{1465\left\{\frac{1}{4}+\frac{1}{4}+\frac{(0.2500-0.0000)^2}{0.6250}\right\}}$$
$$= 29.6$$

There is, however, no objection to using two independent variates, e.g., both w and x. In that case the new sum of squared deviations ignoring treatments is

$$24,182-(51.50\ 710.1)\begin{pmatrix}0.8333 & 1.100\\ 1.100 & 31.63\end{pmatrix}^{-1}$$
$$\times\begin{pmatrix}51.50\\ 710.1\end{pmatrix}$$
$$= 7295 \text{ with 18 degrees of freedom.}$$

For the error alone the corresponding figure is 3,470 with 13 degrees of freedom. That leads to the following analysis of variance:

Source	d.f.	Sum of Squares	Mean Square	F
Treatments	5	3825	765	2.87
Error	13	3470	267	
Treatments + error	18	7295		

CHOICE OF INDEPENDENT VARIATES

Modern computer packages permit the simultaneous use of several independent variates, and thus they extend the usefulness of the techniques. Nevertheless, the temptation to introduce every available independent variate is to be resisted.

The original use of covariance was to effect adjustments where some disturbing factor had not been controlled. That remains the most common application. For example, a suspicion may arise that a thermostat is not functioning properly. Pending its replacement, the investigator may introduce periodic measurements of temperature, which are then used as an independent variate in the analysis of data. Assuming that there is a straight-line relationship between the data, y, and the temperature, the outcome might be much the same apart from the loss of a degree of freedom from the error and an arbitrary loss of regularity in the standard errors*. Again, in an agricultural context it may be found that an experiment has been laid down on variable soil, a crisis that could be resolved by a covariance adjustment on measurements of soil texture or soil acidity. Sometimes, too, there are quantities that cannot be controlled precisely. It is not possible, for example, to grow test animals of exactly the same body weight or to find patients with identical blood counts. In these instances covariance adjustments may be used as a matter of course.

At one time the technique had a bad reputation, arising, it is said, from one much-publicized example. The story goes that a field experiment was conducted on the yield of barley, a covariance adjustment being made on germination rates. As a result the error sum of squared deviations* was gratifyingly reduced, but the treatment effects also disappeared. An obvious explanation is that the treatments had affected yield by way of the germination rates and in no other manner. If so, the conclusion could have been of some importance. The story is not a warning against using covariance adjustments at all; indeed, it shows their value in revealing mechanisms, but it does warn against facile interpretations, especially when the independent variate does not justify its name but is dependent upon the treatments.

If the covariate is measured before the application of treatments, which are then allocated at random, no question need arise. Nor does any arise when the aim is to find what the treatments do to the dependent variate apart from the obvious indirect effect through the independent. Sometimes, however, the user cannot be sure whether the treatments do or do not affect the independent variate*. It is wise in such cases to be very cautious. It is true that subjecting the figures to an analysis of variance may decide the matter. If, however, the proposed independent variate is an array of ones and zeros, indicating the presence or absence of some feature, the user is unlikely to obtain any useful guidance. Even a more amenable variate may give an inconclusive response.

CONSTANCY OF THE REGRESSION COEFFICIENT

It is by no means obvious that the regression coefficients will be independent of the treatments, and in some instances the assumption may verge on the absurd. Occasionally, the difficulty can be met by transformation of the variates. Where that is not feasible, an algebraic solution by least squares is usually not possible. Thus, with data from a designed experiment it is simple to fit separate regression coefficients if the design is completely randomized but not if there are blocks, as usually there will be. Although with the aid of a computer some kind of minimization can often be achieved, it is open to question whether a constant block effect regardless of treatments in conjunction with variable regression coefficients makes a convincing and realistic model.

ACCIDENTS AND MISHAPS

It has been recognized for a long time that the analysis of covariance provides a theoretically sound way of dealing with data from damaged experiments. For example, suppose that an experiment has been designed in such a way that a program is available for calculating an analysis of variance, but m plots (i.e., units) have been lost. (It must be reasonable to assume that the loss is not a result of the treatments that have been applied.) Each gap in the data is filled with a convenient value such as zero, the treatment mean, or even an arbitrary number. It is now required to estimate the deviation between the value for the missing plot given by the method of least squares and the value that has been assigned. That is done by writing down a pseudo-variate for each missing value. It equals zero for all plots except the one to which it refers, when it equals 1. A covariance adjustment on the pseudo-variates will give a correct analysis, the regression coefficient of the dependent variate on any pseudo-variate being minus the required deviation for the plot. The method has several advantages. For one thing, unlike many methods for dealing with incomplete data, it gives a correct F-value for any effect. For another, it gives correct standard errors for treatment contrasts*. Also, it obtains degrees of freedom without special adjustment.

A similar problem can arise when the data for any plot (or unit) is the sum of an unspecified number of observations, such as weighings. If someone makes a mistake, there can be doubt whether a certain observation belongs to this plot or that. The difficulty can sometimes be resolved by attributing it first to one and then to the other. If residuals look extraordinary in one case but are unremarkable in the other, the difficulty is over, but sometimes doubt will remain. A similar problem arises when samples are taken, say for chemical analyses, and some labels are lost. The samples can still be analyzed and a total found. In both examples it is possible to state the total for the two plots without knowing how to apportion it between them. It suffices to attribute the total to one plot and zero to the other and to adjust by a pseudo-variate equal to $+1$ and -1 for the two plots and to zero for all others. If three plots are involved in the muddle, two pseudo-variates are required. The total is attributed to one plot and zero to the others. The first pseudo-variate equals +2 for the plot with the total, −1 for the other two involved, and zero for the others. It therefore serves to apportion a correct amount to the plot credited with the total. A second pseudo-variate apportions the rest between the other two affected plots, being equal to +1 and −1 for those plots and to zero for all others. The method can be extended easily to more complicated cases.

Adjustments of this sort can be made in conjunction. Thus provided that the program can manage so many, it is permissible to have some independent variates for adjustment by related quantities, others to allow for missing values, and still others to apportion mixed-up values. All these adjustments can be made correctly and simultaneously.

TRANSFORMATION OF VARIATES

It should be noted that the independent variates, despite their name, play a role in the model analogous to that of **M** rather than that of y, i.e., no assumptions are involved about their distributions, which are perhaps known, but nothing depends upon them. Accordingly, there is no need to seek variance-stabilizing transformations* for them. It is, however, still necessary to consider if y needs one, since not only must the elements of $\boldsymbol{\eta}$ be distributed independently but also they should have equal variances. In the case of the independent variates, the need is for them to be linearly related to y (or to the transformation of y) and that may call for a transformation of a different kind. Alternatively, it may be desirable to introduce the same variate twice but in different forms. Thus, it has already been mentioned that a field experiment on variable land might be improved by an adjustment on soil acidity. However, unless the species is one that favors extremely acid or extremely alkaline soils, there will almost certainly be an optimal value somewhere in the middle of the range of acidity and it would be sensible to introduce both soil pH and its square to allow the fitting of a parabola.

The correct choice of transformation for the independent variate is especially important if the intention is to enquire how far treatments affect the dependent variate other than through the independent. It is then essential to fit the right relationship; an inept choice of transformations can do harm.

An additional variate does little harm but it is not wise to load an analysis with adjustments in the hope of something emerging. A further independent variate should be included only for good reason, but it should not be omitted if there are good reasons for regarding it as relevant. If it does nothing, there should be reserve about taking it out again. Clearly, if variates are included when they reduce the error variance and excluded if they do not, bias must result. Also, in presenting results it is more convincing to report that some quantity or characteristic was allowed for but had in fact made little difference than to ignore it.

LEVEL OF ADJUSTMENT OF AN INDEPENDENT VARIATE

Basically, the method consists of estimating the partial regression coefficients* of the dependent variate upon each of the independent variates and then adjusting the dependent variate to correspond to standard values of the independent variates. What these standard values should be requires some thought.

First, as long as only linear relations are in question, it does not much matter, because differences in adjusted means of the dependent variate will be unaffected, although not the means themselves. If, however, the first independent variate is some measured quantity, x, and the second is x^2 introduced to allow for curvature, a computer package, which does not know that the two are related, will adjust the first to $\overline{x}$, the mean of x, and the second to $(\overline{x^2})$, which will not be the same as $(\overline{x})^2$. Probably little harm will have been done, but the point needs to be noted.

Pseudo-variates do not usually cause much trouble with their standard values as long as only differences are in question. When they are used for missing plots, the adjustment should be to zero, corresponding to the presence of the plot and not to a mean value, if actual means of the dependent variate are to relate to those given by other methods.

INTERPRETATION OF AN ANALYSIS OF COVARIANCE

An analysis of covariance having been calculated, the first step usually is to look at the error-mean-squared deviation to see if it is as small as was hoped. If it is not, there are two possibilities. One is that, as with the analysis of variance, some important source of variation has been left in error; the other comes from the independent variates having had little effect. The latter case needs to be noted because a research team can go on using adjustments believing them to be a sovereign remedy, even though in fact they do no good. To test the matter formally, it is sufficient to carry out an F-test* using the two minimizations provided by the analysis of variance with and without the independent variates. Sometimes only one of the independent variates is in question; it may be helpful to repeat the calculations omitting that variate to see if it has really had a useful effect.

The testing of effects is the same as for the analysis of variance. They should be examined in logical order and tables prepared to show all effects of importance. Where there is any possibility of an independent variate having been affected by the treatments, it may be advisable to examine the position using the analysis of variance.

The standard error of a treatment mean depends party upon the design and the error variance, as in the analysis of variance, and partly on the magnitude of the adjustments that have been required. In one sense these modifications of standard errors are a help. Cases commonly dealt with using pseudovariates, missing plots, for example, require modifications that are not easy to make except by covariance and then they are made automatically. On the other hand, a measured independent variate will give arbitrary variation in the standard errors, which can be very awkward, for instance, in a multiple comparison test*. Incidentally, some computer packages disguise the situation by giving a common mean standard error for a set of quantities which, but for the adjustments, would all have been determined with the same precision. Although often convenient, the practice can also be misleading.

REFERENCES

1. Bartlett, M. S. (1937). *J. R. Statist. Soc. Suppl. 4*, 137–183.
2. Bliss, C. I. (1967). *Statistics in Biology*, Vol. 2. McGraw-Hill, New York, Chap. 20.
3. Cox, D. R. (1958). *Planning of Experiments*. Wiley, New York, Chap. 4.
4. Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh. (A passage was added to later editions of *Statistical Methods for Research Workers* prior to the appearance of *The Design of Experiments*.)
5. Pearce, S. C. (1953). *Field Experimentation with Fruit Trees and Other Perennial Plants*. Commonwealth Bureau of Horticulture and Plantation Crops, Farnham Royal, Slough, England, App. IV.
6. Wishart, J. (1936). *J. R. Statist. Soc. Suppl. 3*, 79–82.

FURTHER READING

The general literature on the subject is rather scanty. *Biometrics** devoted Part 2 of Volume 13 (1957) to a series of papers on the analysis of covariance. Later *Communications in Statistics** similarly devoted Volume A8, Part 8 (1979) to the topic. There is a valuable account of the method in C. I. Bliss's *Statistics in Biology*, Volume II, Chapter 20 (McGraw-Hill, New York). Also, there are short but illuminating descriptions by D. J. Finney in *An Introduction to Statistical Science in Agriculture* (Munksgaard, Copenhagen, 1962) and by D. R. Cox in Chapter 4 of *Planning of Experiments* (Wiley, New York, 1958).

See also AGRICULTURE, STATISTICS IN; ANALYSIS OF VARIANCE; DESIGN OF EXPERIMENTS: INDUSTRIAL AND SCIENTIFIC APPLICATIONS; FACTOR ANALYSIS; GENERAL LINEAR MODEL; and REGRESSION (Various Entries).

S. C. PEARCE

ANALYSIS OF MEANS FOR RANKS (ANOMR)

The analysis of means for ranks (ANOMR) procedure, attributed to Bakir [1], assumes k independent samples of sizes n_i, $i = 1, \ldots, k$. Observations

$$X_{ij}, i = 1, \ldots, k; j = 1, \ldots, n_i,$$

are selected from k continuous populations that may differ only in their location parameters $\mu_i, i = 1, \ldots, k$. We test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus the alternative that not all μ_is are equal.

Replacing X_{ij} by its rank R_{ij} in the combined sample of size $N = \sum_{i=1}^{K} n_{i-1}$, we calculate

$$\overline{R}_i = \sum_{j=1}^{n_i} (R_{ij}/n_i).$$

The null hypothesis is rejected if for any i

$$|\overline{R}_i - \overline{R}| \geqslant C,$$

where $\overline{R} = \frac{1}{2}(N + 1)$ is the grand mean of all the ranks, and where the critical value C is a function of the size α of the test, of k, and of $n_1, \ldots, n_k$. Tables of values of C for selected combinations for $\alpha, k = 3, 4$, and for small values of n_i have been provided [1]. For equal sample sizes ($n_i = n, i = 1, \ldots, k$), a Bonferroni* approximation based on the Wilcoxon rank sum* statistic is available [1], which turns out to be satisfactory when α is in the vicinity of 0.072.

An asymptotically normal procedure is suggested [1]. Specifically, under H_0 and when the sample sizes are equal, each R_i has the expected value $(N + 1)/2$ and variance

$$\begin{aligned} \sigma^2 &= \frac{\mathrm{Var}(R_{ij})}{n} \cdot \left(\frac{N - n}{N - 1}\right) \\ &= \frac{(N - n)(N + 1)}{12n} \\ &= \frac{(k - 1)(kn + 1)}{12}, \end{aligned}$$

since $\mathrm{Var}(R_{ij}) = (N^2 - 1)/12$. Note that H_0 here corresponds to the classical Kruskal–Wallis test* that, however, involves more general alternatives than does the ANOMR procedure here.

The standard differences $W_i = |\overline{R}_i - \overline{R}|/\sigma$ define the random vector $\boldsymbol{W} = (W_1, \ldots, W_k)'$. As $N \to \infty$, the limiting distribution of $\boldsymbol{W}$ is a (singular) multivariate distribution with equicorrelated structure, given by

$$\mathrm{Corr}(W_i, W_j) = 1/(k - 1).$$

The large-sample ANOMR test is based on the following rule: reject H_0 if for any i

$$|W_i| \geqslant w,$$

where the critical value w is obtained from the analysis of means (ANOM) table of Nelson [2] with appropriate values of k and infinite degrees of freedom. Further details are provided in Reference 3.

REFERENCES

1. Bakir, S. T. (1989). Analysis of means using ranks. *Commun. Stat. Simul. Comp.*, **18**, 757–775.
2. Nelson, L. S. (1983). Exact critical values for the analysis of means. *J. Qual. Tech.*, **15**, 40–44.
3. Wludka, P. S. and Nelson, P. R. (1999). Two non-parametric, analysis-of-means-type tests for homogeneity of variances. *J. Appl. Stat.*, **26**, 243–256.

See also KRUSKAL–WALLIS TEST.

ANALYSIS OF VARIANCE

The analysis of variance is best seen as a way of writing down calculations rather than as a technique in its own right. For example, a significance of a correlation coefficient*, r, based on n pairs of observations, can be tested in several ways. Using the analysis of variance the calculations are set out thus:

Source	Degrees of Freedom	Sum of Squared Deviations	Mean Squared Deviation
Correlation	1	Sr^2	Sr^2
Error	$n - 2$	$S(1 - r^2)$	$S(1 - r^2)/(n - 2)$
Total	$n - 1$	S	

The significance is then judged by an F-test*, i.e., from the ratio of the two mean squared deviations, which is $(n-2)r^2/(1-r^2)$ with one and $(n-2)$ degrees of freedom. By using a standard format, standard tests are available, thus obviating the need for numerous formulas and a range of tables, but in general the same conclusions will be reached however the calculations are set out.

HISTORY AND DEVELOPMENT

In its origins the method was devised by Fisher [1] for the study of data from agricultural experiments. At first the only designs, i.e., randomized blocks* and Latin squares*, were orthogonal, but later F. Yates [4] showed how to deal with those that were nonorthogonal, like balanced incomplete blocks*, and those with a factorial structure of treatments [5]. From this point development was rapid, so that the analysis of variance was being used for a wide range of problems that involved the studying of a linear hypothesis*. The original problem of analyzing data from block designs acquired a new dimension from the use of matrices, due to Tocher [3], and the solution of the normal equations in terms of generalized inverses*. *See* General Linear Model.

THE METHOD DESCRIBED

In essentials the method depends upon the partition of both degrees of freedom and the sums of squared deviations between a component called "error" and another, which may be termed the "effect," although generally it will have a more specific name. Thus, in the example above the effect was the correlation. The nomenclature should not mislead. The sum of squared deviations for the effect is influenced by error also, which is thought of as an all-pervading uncertainty or noise* distributed so that, in the absence of the effect i.e., on the null hypothesis*, the expectation of the two sums of squared deviations will be in the ratio of their respective degrees of freedom. Hence the mean squared deviations, i.e., the sums of squares divided by their degrees of freedom, should have similar expectations. If, however, the effect does exist, it will inflate its own mean squared deviation but not that of the error, and if large enough, will lead to significance being shown by the F-test. In this context, F equals the ratio of the mean squared deviations for the effect and for error.

In practice, analyses of variance are usually more complicated. In the example, the total line was obtained by minimizing the sum of squared residuals*, $\sum_i \eta_i^2$, in

$$y_i = \alpha + \eta_i,$$

whereas that for error came from minimizing a similar quantity in

$$y_i = \alpha + \beta x_i + \eta_i.$$

In short, the test really investigated the existence or non-existence of β. In the example α is common to both minimizations, representing as it does a quantity needed to complete the model but not itself under test, whereas β was in question. That is the general pattern. Thus in a randomized block design the blocks form a convenient "garbage can" where an ingenious experimenter can dispose of unwanted effects such as spatial position, different observers, sources of material, and much else that would disturb the experiment if not controlled. Consequently, they must be allowed for, although no one is studying the contents of garbage cans. There will also be parameters for treatments*, which are under study. Minimization with respect to the block parameters alone gives a measure of the remaining variation, i.e., that due to treatments and uncontrollable error. A further minimization on block and treatments parameters together gives the error line, that for treatments being found by difference. (The fact that it can be found more easily by direct calculation obscures its real origins.) The block line, relating as it does to variation that has been eliminated, is not really relevant but is ordinarily included.

Such an analysis is called "intrablock"*, studying as it does variation within blocks and discarding any between them. In some instances it is possible to derive an "interblock"* analysis, in which the block parameters are regarded as random variables. The procedure then is to minimize the sum of their squares, both when the treatment parameters are included and when they are excluded.

The additional information can be worthwhile, but not necessarily so. For example, if each block is made up in the same way with respect to treatments, a study of the differences between blocks can provide no information about the effects of treatments. Also, unless the number of blocks appreciably exceeds that of treatments, there will not be enough degrees of freedom to determine the interblock error properly. Not least, if good use has been made of blocks as garbage cans, the distribution of their parameters must be regarded as arbitrary.

Complications arise when there are several effects. Here it is advisable to form the error allowing for them all, although that will be considered in more detail below. The problems arise rather in deciding the order of testing, but that is often a matter of logic rather than statistics. For example, with a factorial design* of treatments, if it appears that there is an interaction* of factors A and B, the conclusion should be that the response to the various levels of A depends on the level of B, and *vice versa*. If that is so, there is no point in examining the main effects of factors A and B, since each relates to the response to the levels of one factor when it has been averaged over levels of the other. The only true interpretation must rest upon a two-way table* of means. Again, if the example is extended to cover parabolic effects, i.e.,

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \eta_i,$$

and if it appears that γ should be included in the model, i.e., the relationship of x_i and y_i is not a straight line, there need be no detailed study of β, since it has no meaning except as the slope of such a line. However, it is always necessary to consider what really is under test. For example, it could be that

$$y_i = \alpha + \gamma x_i^2 + \eta_i$$

was the expected relationship and doubts had arisen whether βx_i was not needed as well. The analysis of variance is an approach of wonderful subtlety, capable of adaptation to a wide range of problems. It is used to best advantage when it reflects the thinking and questioning that led to the inception of the investigation in the first place. Consequently, each analysis should be individual and should study each question in logical order. *There is no place for automated procedures, as if all research programs raised the same questions*. Also, although the analysis of variance had its origins in the testing of hypotheses*, there is no reason for leaving it there. It can shed light on the sources of experimental error*; it can suggest confidence limits* for means and differences of means and much else. In the hands of a thoughtful user it has unimagined potentiality; as an unthinking process it leads to few rewards.

Table 1. Yields in Pounds per Plot over Four Seasons from an Experiment on Soil Management with Apple Trees

	Blocks				
Treatment	I	II	III	IV	Totals
A	287	290	254	307	1138
B	271	209	243	348	1071
C	234	210	286	371	1101
D	189	205	312	375	1081
E	210	276	279	344	1109
S	222	301	238	357	1118
Totals	1413	1491	1612	2102	6618

NUMERICAL EXAMPLE

The data in Table 1 [2] represent yields per plot from an apple experiment in four randomized blocks, I through IV. There were six treatments. One of them, S, was the standard practice in English apple orchards of keeping the land clean during the summer, letting the weeds grow up in the fall, and turning them in for green manure in the spring. The rest, A through E, represented alternative methods in which the ground was kept covered with a permanent crop. The interest then lay in finding out if any of the other methods showed any improvement over S.

It is first necessary to find the sum of squared deviations ignoring treatments and considering only blocks. The estimated value for each plot, i, is

$$y_i = \beta_j + \eta_i,$$

where β_j is the parameter for the block, j, in which it finds itself. It will quickly appear that the sum of $\eta_i^2 = (y_j - \beta_j)^2$ is minimized where β_i is taken to be the appropriate block mean, i.e.,

$$\beta_1 = 235.50, \quad \beta_2 = 248.50,$$
$$\beta_3 = 268.67, \quad \beta_4 = 350.33.$$

With these values known it is possible to write down η_i for each plot; e.g., those for blocks I and II and treatments A and B are

$$\begin{array}{rr}
51.50 & 41.50 \ \ldots \\
35.50 & -39.50 \ \ldots \\
\vdots & \vdots
\end{array}$$

The sum of these quantities squared comes to 24,182; it has 20(= 24 − 4) degrees of freedom, since there are 24 data to which four independent parameters have been fitted. It is now required to fit treatment parameters as well, i.e., to write

$$y_i = \beta_j + \gamma_k + \eta_i.$$

In a randomized block design* in which all treatment totals are made up in the same way with respect to block parameters, i.e., the design is orthogonal*, it is sufficient to estimate a treatment parameter, γ_k, as the difference of the treatment mean from the general mean. The table of deviations, η_i, now starts

$$\begin{array}{rr}
42.75 & 32.75 \ \ldots \\
43.50 & -31.50 \ \ldots \\
\vdots & \vdots
\end{array}$$

The sum of these squares is now 23,432 with 15 degrees of freedom, because an additional five degrees of freedom have been used to estimate how the six treatment means diverge from the general mean. (Note that only five such quantities are independent. When five have been found, the sixth is known.) The analysis of variance is therefore

Source	d.f.	Sum of Squares	Mean Square
Treatment	5	750	150
Error	15	23,432	1562
Treatments + error	20	24,182	

There is no suggestion that the treatment mean square has been inflated relative to the error, and therefore no evidence that the treatments in general have had any effect. However, the study really relates to comparisons between A through E and S. In view of the orthogonality of the design and the fact that each treatment mean is based on four data, the variance of a difference of two means is $(\frac{1}{4} + \frac{1}{4})1562 = 781$, the standard error being the square root of that quantity, i.e., 27.9. There is no question of any other treatment being an improvement on the standard because all except A give smaller means. (However, the situation can be changed; *see* ANALYSIS OF COVARIANCE.)

The analysis above was for an orthogonal design. If, however, the datum for treatment A in block 1 had been missing, a more complicated situation would have arisen. (It is true that in practice a missing plot value would be fitted, but it is possible to carry out a valid analysis of variance without doing that.) The deviations allowing only for blocks start

$$\begin{array}{rr}
— & 32.75 \ \ldots \\
45.80 & -47.50 \ \ldots \\
\vdots & \vdots
\end{array}$$

the mean for block I being now 225.2. The sum of squared deviations is 20,999 with 19 degrees of freedom. The so-called normal equations*, derived from the block and treatment totals, are

$$\begin{array}{ll}
1126 = 5\beta_1 + \sum\gamma - \gamma_1 & 1138 = \sum\beta - \beta_1 + 3\gamma_1 \\
1491 = 6\beta_2 + \sum\gamma & 1071 = \sum\beta \qquad + 4\gamma_2 \\
1612 = 6\beta_3 + \sum\gamma & 1101 = \sum\beta \qquad + 4\gamma_3 \\
2102 = 6\beta_4 + \sum\gamma & 1081 = \sum\beta \qquad + 4\gamma_4 \\
 & 1109 = \sum\beta \qquad + 4\gamma_5 \\
 & 1118 = \sum\beta \qquad + 4\gamma_6,
\end{array}$$

where $\sum\beta = \beta_1 + \beta_2 + \beta_3 + \beta_4$ and $\sum\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6$. At this point it is necessary to say that no one in practice solves the normal equations as they stand because better methods are available, e.g., the Kuiper–Corsten iteration* or the use of generalized inverses*. However, there is no objection in principle to a direct solution, the difficulty being that there are not enough equations for the parameters. An equation of constraint, e.g.,

$$3\gamma_1 + 4\gamma_2 + 4\gamma_3 + 4\gamma_4 + 4\gamma_5 + 4\gamma_6 = 0,$$

can always be used. Further, the one just suggested, which makes the treatment parameters sum to zero over the whole experiment, is very convenient. It follows that

$$\begin{array}{ll} \beta_1 = 224.34 & \gamma_1 = -5.74 \\ \beta_2 = 248.74 & \gamma_2 = -5.39 \\ \beta_3 = 268.91 & \gamma_3 = 2.11 \\ \beta_4 = 350.57 & \gamma_4 = -2.89 \\ & \gamma_5 = 4.11 \\ & \gamma_6 = 6.36. \end{array}$$

Hence subtracting the appropriate parameters from each datum, the values of η_i are

$$\begin{array}{ccc} — & 47.00 & \ldots \\ 52.05 & -34.35 & \ldots \\ \vdots & \vdots & \end{array}$$

The sum of their squares is now 20,508 with 14 degrees of freedom, yielding the following analysis of variance:

Source	d.f.	Sum of Squares	Mean Square
Treatments	5	491	98
Error	14	20,508	1465
Treatments + error	19	20,999	

MULTIPLE COMPARISONS*

Over the years there has grown up an alternative approach to testing in the analysis of variance. As long as there are only two treatments or two levels of a factor, the *F*-test has a clear meaning but if there are three or more, questions arise as to where the differences are. Some care is needed here, because background knowledge is called for. If, for example, it had appeared that different varieties of wheat gave different percentages of a vegetable protein in their grain, the result would surprise no one and would merely indicate a need to assess each variety separately. At the other extreme, if the treatments formed a highly structured set, it might be quite obvious what should be investigated next. Thus if it had appeared that the outcome of a chemical reaction depended upon the particular salt used to introduce a metallic element, the science of chemistry is so developed that a number of lines of advance could probably be suggested immediately, some of which might receive preliminary study from further partition of the treatment line. (Of course, no body of data can confirm a hypothesis suggested by itself.) Sometimes, however, the experimenter is in the position of suspecting that there could be a structure but he does not know what it is. It is then that multiple comparison tests have their place. Like much else in the analysis of variance, their unthinking use presents a danger, but that is not to deny them a useful role. Also, an experimenter confronted with a jumble of treatment means that make no sense could well adopt, at least provisionally, the treatment that gave the highest mean, but would still want to know how it stood in relation to its nearest rivals. It might be so much better that the others could be discarded, or it might be so little better that further study could show that it was not really to be preferred. Such a test can be useful. Like others, it can be abused.

COMPOSITION OF ERROR

In its original form the analysis of variance was applied to data from agricultural field experiments designed in randomized blocks. The error was then clearly identified as the interaction of treatments and blocks, i.e., the *F*-test investigated the extent to which treatment differences were consistent from block to block. If a difference of means was much the same in each block, it could clearly be relied upon; if it had varying values, the significance was less well established. In other instances the error may be the interaction of treatments and some other factor such as occasion or operator. In all such cases the purport of the test is clear.

Sometimes, however, the error is less open to interpretation, being little more than a measure of deviations from a hypothesis that is itself arbitrary. Thus, if the agricultural field experiment had been designed in a Latin square*, the error would have been made up of deviations from parameters for treatments added to those for an underlying fertility pattern, assumed itself to derive from additive effects of rows and columns. If, as is usually the case, there is no prior reason why the fertility pattern should have that form, there is a danger of the error sum of squared deviations being inflated by sources that will not affect the treatment line, thus reversing the usual position. In fact, it is not unknown for the error mean square deviation to be the larger, and sometimes there is good reason for it.

The subject of error is sometimes approached by distinguishing fixed effects* from random*. In the first, the levels are determinate and reproducible, like pressure or temperature, whereas in the latter they are to be regarded as a random selection of possible values, like the weather on the successive days of an investigation. For many purposes the distinction is helpful. The ideal, as has been suggested, is an error that represents the interaction between a fixed effect and a random effect of the conditions over which generalization is required, but other possibilities can be recognized. For example, an error made up of interactions between fixed effects is nearly useless unless it can be assumed that there will be no real interaction.

Further questions arise when confounding* is introduced. The experimenter may have decided that there can be no interaction of factors A, B, and C and would then be ready to have it confounded. The experimenter may get rid of it in that way if possible, but if it remains in the analysis it should be as part of error. However, if the experimenter believes that A × B × C never exists, he or she cannot believe that its value depends upon the level of D, so A × B × C × D also should either be confounded or included in error. Such decisions can be made readily enough before analysis begins; they are more difficult afterward. Plainly, it would be wrong to start with an error that was acceptable and add to it those high order interactions* that were small and to exclude from it those that were large. The error that finally resulted would obviously be biased*. On the other hand, there are occasions when sight of the data convinces the experimenter that his or her preconceptions were wrong. In that case the experimenter usually does well to confine the analysis to an indubitable error, e.g., the interaction of blocks and treatments, and to regard all else as subject to testing. As with the Latin square*, mistaken assumptions about the composition of error can lead to the inflation of its sum of squared deviations.

The question of what is and what is not error depends to some extent upon the randomization*. Let there be b blocks of xy treatments, made up factorially by x levels of factor X and y levels of factor Y. The obvious partition of the $(bxy - 1)$ degrees of freedom between the data is

Blocks	$(b-1)$
X	$(x-1)$
Y	$(y-1)$
$X \times Y$	$(x-1)(y-1)$
Blocks × $X(I)$	$(b-1)(x-1)$
Blocks × $Y(II)$	$(b-1)(y-1)$
Blocks × $X \times Y(III)$	$(b-1)(x-1)(y-1)$.

No one need object if each of the effects X, Y and X × Y is compared with its own interaction with blocks, i.e., with I, II, and III, respectively. If, however, all treatment combinations have been subject to the same randomization procedure, the components I, II, and III can be merged to give a common error with $(b-1)(xy-1)$ degrees of freedom. In a split-plot* design, where X is allocated at random to the main plots, II and III can together form a subplot error, leaving I to form that for main plots. Given a strip-plot* (or crisscross) design, on the other hand, each component of error must be kept separate. The example illustrates the old adage: “As the randomization is, so is the analysis.” Sometimes there are practical difficulties about randomization and they can raise problems in the analysis of data, but they are not necessarily insuperable ones. For example, in the split-plot case it may not be feasible to randomize the main plot factor, X, in which case it is vitiated and so is its interaction with blocks (component I), but it might still be permissible to use the subplot analysis*. Again, if the subplot factor, Y, has to be applied systematically (it might be different occasions on which the main plot was measured), component II may be vitiated, but X × Y can still be compared with component III.

SUITABILITY OF DATA

With so many computer packages available it is easy to calculate an analysis of variance that is little better than nonsense. Some thought therefore needs to be given to the data before entering the package.

Strictly, the data should be continuous. In fact, it is usually good enough that they spread themselves over 10 or more points of a scale. When they do not, e.g., as with a body of data that consists mostly of ones and zeros, the sum of squared deviations for error is inflated relative to those for effects with consequent loss of sensitivity. Discrete data often come from Poisson* or binomial* distributions and may call for one of the variance-stabilizing transformations* to be considered below.

It is also required that the residuals* (the η_i in the examples) should be distributed independently and with equal variance. Independence is most important. Where the data come from units with a spatial or temporal relationship, independence is commonly achieved by a randomization of the treatments. If that is impracticable, the analysis of variance in its usual forms is better avoided. Equality of variance is more problematical. Where treatments are equally replicated*, the F-test is fairly robust* against some treatments giving a higher variance than others. If, however, attention is directed to a subset of treatments, as happens with multiple comparisons* and can happen when the treatment line is partitioned, it is necessary to have a variance estimated specifically for that subset or serious bias could result. Here two main cases need to be distinguished. Some treatments may involve more operations than others and may therefore give rise to larger errors. For example, the injection of a plant or animal can itself give rise to variation that is absent when other methods of administration are adopted. In the other case, the variance for any treatment bears a functional relationship to the mean value for that treatment.

To take the first case, if the error is the interaction of treatment and blocks, it is an easy matter to partition the treatment line into three parts: (1) between the groups, one with higher and one with lower variance; (2) within one group; and (3) within the other. The error can then be partitioned accordingly. Depending upon circumstances, it may be easier to regard each component of treatments as having its own error or to concoct a variance for each contrast of interest between the treatment means. Alternatively, if the treatments of high and low variance are associated with two levels of a factor, e.g., administration by injection as compared with some other method, a better way and one that places fewer constraints on the nature of error may be to group the data into pairs according to the two levels, a pair being the same with respect to other factors, and to analyze separately the sum and difference of data from the pair. In effect, that is virtually the same as regarding the design as if it were in split plots, the factor associated with the divergent variances being the one in subplots.

In the other case, where the variance of any observation depends upon the mean, the usual solution is to find a variance-stabilizing transformation*. Thus, if the variance is proportionate to the mean, as in a Poisson distribution*, it may be better to analyze the square roots of the data rather than the data themselves. Such transformations can be useful, especially when they direct attention to some quantity more fundamental than that measured. Thus given the end product of a growth process, it is often more profitable to study the mean growth rate than the final size, because that is what the treatments have been affecting, and similarly the error has arisen because growth rates have not been completely determined. The approach can, however, cause problems when the transformation is no more than a statistical expedient, especially when it comes to the interpretation of interactions. Thus suppose that an untreated control has led to 90% of insects surviving, applications of insecticide A had given a survival rate of 60% while B had given 45%. A declaration that the two insecticides had not interacted would lead most people to suppose that A and B in combination had given a survival rate of 30% ($30 = 60 \times 45/90$). If the data are analyzed without transformation, a zero interaction would imply 15% for the combination ($15 = 60 + 45 - 90$). Using the appropriate angular transformation, the figures become (0) 71.6, (A) 50.8, (B) 42.1, leading to an expectation for (AB) of 21.3 for zero interaction. This last figure corresponds to 13.2%, which is even further from the ordinary meaning of independence*.

INTERPRETATION OF AN ANALYSIS OF VARIANCE

As has been said, the analysis of variance is a subtle approach that can be molded to many ways of thought. It is at its best when the questions implicit in the investigation are answered systematically and objectively. It is not a method of "data snooping," the treatment line being partitioned and repartitioned until something can be declared "significant." Nor is it rigid, as if there were some royal way to be followed that would ensure success. Its function is to assist a line of thought, so any unthinking procedure will be unavailing.

The first step is usually to look at the mean squared deviation for error, which sums up the uncontrolled sources of variation. An experienced person may note that its value is lower than usual, which could suggest that improved methods have paid off, or it could be so large as to show that something had gone wrong. If that is the case, an examination of residuals may show which observations are suspect and provide a clue for another time. It may even lead to a positive identification of the fault and the exclusion of some data. The fault, however, must be beyond doubt; *little credence attaches to conclusions based on data selected by the investigator to suit his own purposes*.

Next, it is wise to look at the line for blocks or whatever else corresponds to the control of extraneous variation. It is possible for a research team to fill their garbage cans again and again with things that need not have been discarded. If the block line rarely shows any sign of having been inflated, it could well be that a lot of trouble is being taken to control sources of variation that are of little importance anyway. Also, if the sources of variation are indeed so little understood, it could be that important components are being left in error.

All this is a preliminary to the comparison of treatment effects* and error. Here some thought is needed. First, the partition of the treatment line into individual effects may be conventional but irrelevant to immediate needs. For example, in a $2 \times p$ factorial set of treatments, there may be no interest in the main effects* and interactions*, the intention being to study the response of the factor with two levels in p different conditions. Even if the partition is suitable, the order in which the effects should be studied needs consideration. As has been explained, a positive response to one test may render others unnecessary.

However, the need may not be for testing at all. The data may give a set of means, and it is only necessary to know how well they have been estimated. Even if the need is for testing, there are occasions when an approach by multiple comparisons* is called for rather than by F-test. Also, there are occasions when a significance test* has proved negative, but interest centers on its power*; that is, the enquiry concerns the probability of the data having missed a difference of specified size, supposing that it does exist.

A difficult situation arises when a high-order interaction appears to be significant but without any support from the lower-order interactions contained in it. Thus if A × B × C gives a large value of F while B × C, A × C, A × B, A, B, and C all seem unimportant, it is always possible that the 1 : 20 chance, or whatever it may be, has come off. Before dismissing awkward effects, however, it is as well to look more closely at the data. The whole interaction could depend upon one observation that is obviously wrong. If, however, all data for a particular treatment combination go the same way, whether up or down, there is the possibility of some complicated and unsuspected phenomenon that requires further study.

Anyone who interprets an analysis of variance should watch out for the inflated error. If there is only one treatment effect and that is small, little can be inferred, but if several F-values for treatment effects are well below expectation, it is at least possible that the error has been badly conceived. For example, blocks may have been chosen so ineptly that, instead of bringing together similar plots or units, each is composed of dissimilar ones. Again, the distribution of the residuals may be far from normal. Randomization* may have been inadequate, or the error may represent deviations from an unlikely model. The matter should not be left. A valuable pointer might be obtained to the design of better investigations in the future.

REFERENCES

1. Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.
2. Pearce, S. C. (1953). *Field Experimentation with Fruit Trees and Other Perennial Plants. Tech. Commun. Bur. Hort. Plantation Crops*, Farnham Royal, Slough, England, **23**. App IV.
3. Tocher, K. D. (1952). *J. R. Statist. Soc. B*, **14**, 45–100.
4. Yates, F. (1933). *J. Agric. Sci.*, **23**, 108–145.
5. Yates, F. (1937). The Design and Analysis of Factorial Experiments. *Tech. Commun. Bur. Soil Sci. Rothamsted*, **35**.

FURTHER READING

A standard work is H. Scheffé's book *The Analysis of Variance* (Wiley, New York, 1959). Other useful books are *Statistical and Experimental Design in Engineering and the Physical Sciences* by N. L. Johnson and F. C. Leone (Wiley, New York, 1966), especially Volume 2, and *The Linear Hypothesis: A General Theory* by G. A. F. Seber (Charles Griffin, London, 1966). In *Experiments: Design and Analysis* (Charles Griffin, London, 1977), J. A. John and M. H. Quenouille discuss the analysis of variance for many standard designs, and G. B. Wetherill's *Intermediate Statistical Methods* (Chapman & Hall, London, 1980) looks with some care at various models. In Chapter 9 of their book, *Applied Regression Analysis* (Wiley, New York, 1966), N. R. Draper and H. Smith show the relationship of the analysis of variance to regression methods. Other useful references are Chapter 3 of C. R. Rao's *Advanced Statistical Methods in Biometric Research*, and Chapter 4 of G. A. F. Seber's *Linear Regression Analysis*, both published by Wiley, New York, the first in 1952 and the second in 1976.

See also Agriculture, Statistics in; Analysis of Covariance; Confounding; Design of Experiments; *F*-Tests; General Linear Model; Multiple Comparisons—I; and Regression (Various).

S. C. Pearce

ANALYSIS OF VARIANCE, WILK–KEMPTHORNE FORMULAS

For ANOVA* of data from cross-classified experiment designs with unequal but proportionate class frequencies, Wilk and Kempthorne [3] have developed general formulas for the expected values of mean squares (EMSs) in the ANOVA table. We will give appropriate formulas for an $a \times b$ two-way cross-classification.

If the number of observations for the factor level combination (i, j) is n_{ij}, then proportionate class frequencies require

$$n_{ij} = nu_iv_j, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b.$$

In this case the Wilk and Kempthorne formulas, as presented by Snedecor and Cochran [2] for a variance components* model, are: For main effect* factor A,

$$\text{EMS } = \sigma^2 + \frac{nuv(1-u^*)}{a-1} \times \{(v^* - B^{-1})\sigma^2_{AB} + \sigma^2_A\};$$

for main effect of factor B,

$$\text{EMS } = \sigma^2 + \frac{nuv(1-v^*)}{b-1} \times \{(u^* - A^{-1})\sigma^2_{AB} + \sigma^2_B\};$$

for main interaction* $A \times B$,

$$\text{EMS} = \sigma^2 + \frac{nuv(1-u^*)(1-v^*)}{(a-1)(b-1)}\sigma^2_{AB},$$

where

$$u = \sum_{i=1}^{a} u_i, \quad u^* = \left(\sum_{i=1}^{a} u_i^2\right) \Big/ u^2,$$

$$v = \sum_{j=1}^{b} v_j, \quad v^* = \left(\sum_{j=1}^{b} v_j^2\right) \Big/ v^2,$$

$\sigma_A^2, \sigma_B^2, \sigma_{AB}^2$ are the variances of the (random) terms representing main effects of A and B and the $A \times B$ interactions in the model, respectively, and σ^2 is the (common) variance of the residual terms.

Detailed numerical examples and discussions are available, for example in Bancroft [1].

REFERENCES

1. Bancroft, T. A. (1968). *Topics in Intermediate Statistical Methods*, Vol. 1. Iowa State University Press, Ames, IA.
2. Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 7th ed. Iowa State University Press, Ames, IA.
3. Wilk, M. B. and Kempthorne, O. (1955). *J. Amer. Statist. Ass.*, **50**, 1144–1167.

See also ANALYSIS OF VARIANCE; FACTORIAL EXPERIMENTS; INTERACTION; MAIN EFFECTS; and VARIANCE COMPONENTS.

ANCILLARY STATISTICS—I

Let X be an observable vector of random variables with probability density function* (PDF) $f_X(x; \theta)$, where θ is an unknown parameter taking values over a space Λ. If the distribution of the statistic, or vector of statistics, $A = a(X)$ is not dependent upon θ, then A is said to be ancillary for (the estimation of) θ. Suppose now that X is transformed in a $1:1$ manner to the pair (S, A). The joint PDF of S, A can be written

$$f_{S|A}(s; \theta|a) f_A(a), \qquad (1)$$

where the second term is free of θ. As the examples below will make apparent, an ancillary statistic often indexes the precision of an experiment; certain values of A indicate that the experiment has been relatively informative about θ; others indicate a less informative result. For this reason, it is often argued that procedures of inference should be based on the conditional distribution* of the data given the observed value of the ancillary. For example, estimation* and hypothesis-testing* procedures are based on the first term of (1), and their frequency properties are evaluated over the reference set in which $A = a$ is held fixed at its observed value.

In the simplest context, ancillary statistics arise in experiments with random sample size.

Example 1. Contingent upon the outcome of a toss of an unbiased coin, either 1 or 10^4 observations are taken on a random variable Y which has a $N(\theta, 1)$ distribution. The sample size N is ancillary and, in estimating or testing hypotheses about θ, it seems imperative that N be regarded as fixed at its observed value. For example, a size α test of $\theta = 0$ versus the one-sided alternative $\theta > 0$ has critical region* $\overline{Y} > z_\alpha$ if $N = 1$ or $\overline{Y} > 10^{-2} z_\alpha$ if $N = 10^4$, where $\overline{Y}$ is the sample mean and z_α is the upper α point of a $N(0, 1)$ distribution. Conditional on the observed value of N, this test is uniformly most powerful* (UMPT), and since this test is conditionally of size α for each N, it is also unconditionally of size α. It is discomforting, however, that power considerations when applied to the unconditional experiment do not lead to this test [5].

A second example illustrates further the role of ancillary statistics.

Example 2. Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics* of a random sample from the density $f(x - \theta)$, where f is of specified form and $-\infty < \theta < \infty$. It is easily seen from considerations of the group of location transformations that the statistics $A_i = X_{i:n} - X_{1:n}$, $i = 2, \ldots, n$, are jointly ancillary for θ. R. A. Fisher* [7] describes these statistics as giving the "configuration" of the sample and their observed values can be viewed as being descriptive of the observed likelihood* function. For example, if $n = 2$ and $f(z) = \pi^{-1}(1 + z^2)^{-1}$, $-\infty < z < \infty$, the likelihood is unimodal if $A_2 \leqslant 1$ and bimodal if $A_2 > 1$, while $\overline{X} = (X_1 + X_2)/2$ is the maximum likelihood estimator* in the first case and a local minimum in the second. Here again, a conditional approach to the inference problem is suggested.

By an easy computation, the conditional density of $M = X_{1:n}$ given $A_2 = a_2, \ldots, A_n = a_n$ is

$$c \prod_{1}^{n} f(m + a_i - \theta), \qquad (2)$$

where $a_1 = 0$ and c is a constant of integration. The choice of the minimum M is arbitrary; the maximum likelihood estimator T or any other statistic that measures location may be used in its place. All such choices will lead to equivalent conditional inferences*. The essential point is that inferences can be based on the conditional distribution* (2).

Both of the foregoing examples suggest that, at least in certain problems, a conditionality principle is needed to supplement other statistical principles. The approach then seems clear: In evaluating the statistical evidence, repeated sampling criteria should be applied to the conditional experiment defined by setting ancillary statistics at their observed values. As will be discussed in the section "Nonuniqueness and Other Problems," however, there are some difficulties in applying this directive.

It should be noted that some authors require that an ancillary A be a function of the minimal sufficient statistic* for θ. This is discussed further in the section just mentioned and in "Conditionality and the Likelihood Principle."

RECOVERY OF INFORMATION

Ancillary statistics were first defined and discussed by Fisher [7], who viewed their recognition and use as a step toward the completion of his theory of exhaustive estimation [8, pp. 158 ff.]. For simplicity, we assume that θ is a scalar parameter and that the usual regularity conditions apply so that the Fisher information* may be written

$$\begin{aligned} I_X(\theta) &= E[\partial \log f_X(x;\theta)/\partial\theta]^2 \\ &= -E[\partial^2 \log f_X(x;\theta)/\partial\theta^2]. \end{aligned}$$

If the maximum likelihood estimator $T = \hat{\theta}(X)$ is sufficient* for θ, then $I_T(\theta) = I_X(\theta)$ for all $\theta \in \Lambda$. Fisher calls T exhaustive because all information is retained in reducing the data to this scalar summary.

It often happens, however, that T is not sufficient and that its sole use for the estimation of θ entails an information loss measured by $I_X(\theta) - I_T(\theta)$, which is nonnegative for all θ and positive for some θ. Suppose, however, that T can be supplemented with a set of ancillary statistics A such that (T, A) are jointly sufficient* for θ. The conditional information in T given $A = a$ is defined as

$$\begin{aligned} I_{T|A} &= a^{(\theta)} \\ &= -E[\partial^2 \log f_{T|A}(t;\theta|a)/\partial\theta^2|A = a] \\ &= -E[\partial^2 \log f_{T,A}(t,a;\theta)/\partial\theta^2|A = a], \end{aligned}$$

since $f_A(a)$ is free of θ. Thus since T, A are jointly sufficient,

$$E[I_{T|A}(\theta)] = I_{T,A}(\theta) = I_X(\theta).$$

The average information in the conditional distribution of T given A is the whole of the information in the sample. The use of the ancillary A has allowed for the total recovery of the information on θ. Depending on the particular observed outcome $A = a$, however, the conditional information $I_{T|A} = a^{(\theta)}$ may be greater or smaller than the expected information $I_X(\theta)$.

Viewed in this way, an ancillary statistic A quite generally specifies the informativeness of the particular outcome actually observed. To some extent, the usefulness of an ancillary is measured by the variation in $I_{T|A}(\theta)$.

Although only a scalar parameter θ has been considered above, the same general results hold also for vector parameters. In this case, $I_X(\theta)$ is the Fisher information matrix* and $I_X(\theta) - I_T(\theta)$ is nonnegative definite. If T is the vector of maximum likelihood estimators, A is ancillary and T, A are jointly sufficient, the conditional information matrix, $I_{T|A}(\theta)$, has expectation $I_X(\theta)$ as above.

NONUNIQUENESS AND OTHER PROBLEMS

Several difficulties arise in attempting to apply the directive to condition on the observed values of ancillary statistics.

1. There are no general constructive techniques for determining ancillary statistics.
2. Ancillaries sometimes exist which are not functions of the minimal sufficient statistic*, and conditioning upon their observed values can lead to procedures

that are incompatible with the sufficiency principle.

3. There is, in general, no maximal ancillary.

In this section we look at problems 2 and 3. It should be noted that in certain problems (e.g., Example 2), group invariance* arguments provide a partial solution to problem 1. Even in such problems, however, there can be ancillaries present which are not invariant. An interesting example is given by Padmanabhan [10]. In the context of Example 2 with $f(\cdot)$ a standard normal density* and $n = 2$, he defines the statistic

$$B = \begin{cases} X_1 - X_2 & \text{if } X_1 + X_2 \geqslant 1, \\ X_2 - X_1 & \text{if } X_1 + X_2 < 1, \end{cases}$$

and shows that B is ancillary but not invariant.

Basu [2] has given several examples of nonunique ancillary statistics. The first of these concerns independent bivariate normal* variates (X_i, Y_i), $i = 1, \ldots, n$, with means 0, variances 1, and correlation ρ. In this example, $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are each ancillary and conditional inference would clearly lead to different inferences on ρ. this is an example of problem 2 above, and to avoid this difficulty many authors require that the ancillary be a function of the minimal sufficient statistic (e.g., ref. 5). If an initial reduction to the minimal sufficient set, $\sum X_i^2 + \sum Y_i^2$ and $\sum X_i Y_i$, is made, there appears to be no ancillary present.

Not all examples of nonuniqueness are resolved by this requirement. Cox [4] gives the following example, which derives from another example of Basu.

Example 3. Consider a multinomial distribution* on four cells with respective probabilities $(1-\theta)/6$, $(1+\theta)/6$, $(2-\theta)/6$, and $(2+\theta)/6$, where $|\theta| < 1$. Let $X_1, \ldots, X_4$ represent the frequencies in a sample of size n. Each of the statistics

$$A_1 = X_1 + X_2 \qquad A_2 = X_1 + X_4$$

is ancillary for θ, but they are not jointly ancillary.

If conditional inference* is to be useful in such problems, methods for selecting from among competing ancillaries are needed. Cox [4] notes that the usefulness of an ancillary is related to the variation in $I_{T|A}(\theta)$ (see the preceding section) and suggests (again with scalar θ) that the ancillary be chosen to maximize

$$\text{var}\{I_{T|A}(\theta)\}.$$

In general, this choice may depend on θ. In the example above, however, Cox shows that A_1 is preferable to A_2 for all θ. The choice of variance as a measure of variation is, of course, arbitrary.

Barnard and Sprott [1] argue that the ancillary's role is to define the shape of the likelihood function, and that, in some problems, invariance* considerations lead to a straightforward selection between competing ancillaries. In the example above, the estimation problem is invariant under reflections with $\theta \leftrightarrow -\theta$ and $X_1 \leftrightarrow X_2$, $X_3 \leftrightarrow X_4$. Under this transformation, the ancillary A_1 is invariant while $A_2 \leftrightarrow n - A_2$. Thus under a natural group of transformations, the ancillary A_1 and not A_2 is indicated. This type of argument suggests that *invariance**, and not ancillarity, is the key concept.

CONDITIONALITY AND THE LIKELIHOOD PRINCIPLE*

In a fundamental paper, Birnbaum [3] formulates principles of sufficiency, conditionality, and likelihood and shows that the sufficiency* and conditionality principles are jointly equivalent to the likelihood principle. In this section we outline Birnbaum's arguments and some of the subsequent work in this area.

Birnbaum introduces the concept of the "total evidential meaning" (about θ) of an experiment E with outcome x and writes $\mathscr{E} \downarrow (E, x)$. Total evidential meaning is left undefined but the principles are formulated with reference to it, as follows.

The Sufficiency Principle (S). Let E be an experiment with outcomes x, y and t be a sufficient statistic. If $t(x) = t(y)$, then

$$\mathscr{E} \downarrow (E, x) = \mathscr{E} \downarrow (E, y).$$

This principle (S) is almost universally accepted by statisticians, although some would limit the types of experiments E to which the principle applies.

Let $L(\theta; x, E)$ denote the likelihood* function of θ on the data x from experiment E.

The Likelihood Principle* (L). Let E_1 and E_2 be experiments with outcomes x_1 and x_2, respectively. Suppose further that

$$L(\theta; x_1, E_1) \propto L(\theta; x_2, E_2).$$

Then $\mathscr{E} \downarrow (E_1, x_1) = \mathscr{E} \downarrow (E_2, x_2)$.

This principle (L) (sometimes called the strong likelihood principle) asserts that only the observed likelihood is relevant in assessing the evidence. It is in conflict with methods of significance testing*, confidence interval procedures, or indeed any methods that are based on repeated sampling*. The sufficiency principle is sometimes called the weak likelihood principle since, by the sufficiency of the likelihood function, it is equivalent to the application of the likelihood principle to a single experiment.

An experiment E is said to be a mixture experiment* with components E_h if, after relabeling of the sample points, E may be thought of as arising in two stages. First, an observation h is made on a random variable H with known distribution, and then x_h is observed from the component experiment E_h. The statistic H is an ancillary statistic.

The Conditionality Principle* (C). Let E be a mixture experiment with components E_h. Then

$$\mathscr{E} \downarrow (E, (h, x_h)) = \mathscr{E} \downarrow (E_h, x_h).$$

This principle asserts that the inference we should draw in the mixture experiment with outcome (h, x_h) should be the same as that drawn from the simpler component experiment E_h when x_h is observed.

Birnbaum's Theorem. (S) + (C) $\Leftrightarrow$ (L).

Proof. It follows immediately that (L) implies (C). Also (L) $\Rightarrow$ (S) since, by the factorization theorem*, two outcomes giving the same value of a sufficient statistic yield proportional likelihood functions. To show that (C) and (S) together imply (L), let E_1 and E_2 be two experiments with outcomes x_1 and x_2, respectively, such that $\mathrm{L}(\theta; x_1, E_1) \propto \mathrm{L}(\theta; x_2, E_2)$. Let $\Pr[H = 1] = 1 - \Pr[\mathrm{H} = 2] = p$ be a specified nonzero probability. In the mixture experiment E with components E_h, $h = 1, 2$, the outcomes $(H = 1, x_1)$ and $(H = 2, x_2)$ give rise to proportional likelihoods. Since the likelihood function is itself minimally sufficient, (S) implies that

$$\mathscr{E} \downarrow (E, (H = 1, x_1)) = \mathscr{E} \downarrow (E, (H = 2, x_2)).$$

On the other hand, (C) implies that

$$\mathscr{E} \downarrow (E, (H = h, x_h)) = \mathscr{E} \downarrow (E_h, x_h), \quad h = 1, 2,$$

and hence $\mathscr{E} \downarrow (E_1, x_1) = \mathscr{E} \downarrow (E_2, x_2)$. Thus (S) + (C) $\Rightarrow$ (L).

Since almost all frequency-based methods of inference contradict the likelihood principle, this result would seem to suggest that sufficiency and conditionality procedures are jointly incompatible with a frequency theory. It should be noted, however, that the foregoing argument pertains to the symmetric use of sufficiency and conditionality arguments. The likelihood principle would appear to follow only under such conditions.

Durbin [6] restricted the conditionality principle to apply only after an initial reduction to the minimal sufficient statistic. He defined a reduced experiment E' in which only T is observed and considered a revised conditionality principle (C') which applied to this reduced experiment. In essence, this restricts attention to ancillary statistics that are functions of the minimal sufficient statistic. This change apparently obviates the possibility of deducing (L).

A second approach [9] classifies ancillaries as being experimental or mathematical. The former are ancillary by virtue of the experimental design and the purpose of the investigation. They are ancillary statistics regardless of the parametric model chosen for the chance setup being investigated. Mathematical ancillaries, on the other hand, are ancillary because of the particular parametric model assumed. In the examples given above, N is an experimental ancillary in Example 1 while the ancillaries A_1 and A_2 in Example 3 are mathematical. Example 2 may be interpreted in two ways. If this is a measurement model* whereby the response

X arises as a sum, $X = \theta + e$, e being a real physical entity with known distribution, then the ancillaries $A_2, \ldots, A_n$ are experimental. (The purpose of the experiment is to determine the physical constant θ.) More usually, an experiment with data of this type is designed to determine the distribution of X and the model $f(x - \theta)$ is a preliminary specification. In this case, the ancillaries are mathematical. The primary principle is taken to be the experimental conditionality principle and other principles (e.g., sufficiency) are applied only after conditioning on any experimental ancillaries present.

REFERENCES

1. Barnard, G. A. and Sprott, D. A. (1971). In *Waterloo Symposium on Foundations of Statistical Inference.* Holt, Rinehart and Winston, Toronto, pp. 176–196. (Investigates relationships between ancillaries and the likelihood function.)
2. Basu, D. (1964). *Sankhyā A*, **26**, 3–16. (Discusses the problem of nonuniqueness and gives several interesting examples.)
3. Birnbaum, A. (1962). *J. Amer. Statist. Ass.*, **57**, 269–306. (A fundamental paper dealing with the formalization of inference procedures and principles.)
4. Cox, D. R. (1971). *J. R. Statist. Soc. B*, **33**, 251–255.
5. Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman & Hall, London. (Probably the most complete text reference on ancillary statistics.)
6. Durbin, J. (1970). *J. Amer. Statist. Ass.*, **65**, 395–398.
7. Fisher, R. A. (1936). *Proc. Amer. Acad. Arts Sci.*, **71**, 245–258. (First discussion of the existence and use of ancillary statistics.)
8. Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed. Oliver & Boyd, Edinburgh. (A discussion of the recovery of information and the use of ancillaries.)
9. Kalbfleisch, J. D. (1975). *Biometrika*, **62**, 251–268.
10. Padmanabhan, A. R. (1977). *Amer. Statist.*, **31**, 124.

See also ANCILLARY STATISTICS, FIRST DERIVATIVE; BASU THEOREMS; CONDITIONAL INFERENCE; FIDUCIAL INFERENCE; INFERENCE, STATISTICAL; LIKELIHOOD; and SUFFICIENT STATISTICS.

JOHN D. KALBFLEISCH

ANCILLARY STATISTICS—II

Since the original entry "Ancillary Statistics" appeared in *ESS* Vol. 1 in 1982, further study has been devoted to the properties and applications of ancillary statistics. This entry will emphasize some of the practical developments, such as methods of approximating conditional distributions and of construction of ancillaries, as well as further problems of nonuniqueness. We also address the role of ancillaries when a nuisance parameter is present, an area not covered in the original article.

ANCILLARIES WITH NO NUISANCE PARAMETERS

The first mention of the term "ancillary" occurs in Fisher [12], where it is pointed out that, when the maximum-likelihood estimator is not fully efficient, the information it provides may be enhanced by the adjunction as an ancillary statistic of the second derivative of the likelihood function, giving the weight to be attached to the value of the estimator. In Fisher [13], he also requires that the distribution of an ancillary statistic be free of the unknown parameter. This requirement is now generally recognized as the characteristic property. In fact the second derivative of the likelihood function, more commonly called the observed information, is not generally ancillary; however, Efron and Hinkley [11] showed that, under certain conditions, it is a better estimator of the conditional variance of the estimator than is the expected Fisher information*.

Useful surveys of the properties of ancillary statistics have been given by Buehler [6] and by Lehmann and Scholz [15]. These surveys discuss, among other things, the sense in which one type of ancillary may function as an index of precision of the sample. There is an important distinction to be made between those ancillary statistics that are part of the minimal sufficient statistic (for the parameter of interest) and those that are not. Fisher's applications refer to the former cases of which he states: "The function of the ancillary statistic is analogous to providing a true, in place of an approximate, weight for the value of the estimate." Ancillaries that are a function of the minimal sufficient statistic we term *internal*, and those that are not we term *external*.

External ancillaries are of interest primarily when nuisance parameters are present, a discussion of which forms the second part of our contribution. Internal ancillaries are available for location and scale families of distributions, and indeed to all transformation models (see below). For such families Fisher [13] provided the initial analysis, which was developed and clarified by Pitman [22]. Pitman estimators have optimal properties within the conditional framework of evaluation. These were covered in the original article.

Approximate Ancillaries

In the first section of our discussion we will understand the term ancillary to mean internal ancillary. For many statistical models ancillary statistics do not exist. Indeed, formulation of precise conditions under which an ancillary does exist remains an open problem. Certainly, a model which does admit an ancillary statistic may not admit one upon minor modification. If such minor modifications are not to make a radical difference to the mode of inference, then we are driven to condition on statistics that are approximately ancillary in an appropriate sense.

This problem was pursued by Cox [7]. Concentrating on the problem of testing $\theta = \theta_0$, we argue that, whatever test statistic is to be used, its distribution should be computed conditional on a statistic A, possibly depending on θ_0, chosen so that its distribution depends on θ as little as possible, locally near θ_0. Assuming appropriate regularity, consider the Taylor expansion

$$E(A;\theta) \approx E(A;\theta_0) + c_1(\theta-\theta_0) + c_2(\theta-\theta_0)^2$$

of the mean of A about θ_0. The statistic A is *first-order local ancillary* if the coefficient $c_1 = 0$. In this case, for values of θ close to θ_0, the mean of A depends little on θ. For instance, at $\theta = \theta_0 + \delta/\sqrt{n}$, close to θ_0, where n is the sample size, the error in the above approximation is $O(n^{-3/2})$ and dependence on θ occurs in the $O(n^{-1})$ term, assuming that A is scaled to have asymptotic unit variance. The statistic A is *second-order local ancillary* if both coefficients c_1 and c_2 are zero as well as the first coefficient in an expansion of the second moment. In many cases, such as for curved exponential families*, local ancillaries are simply constructed from linear or quadratic forms in the canonical sufficient statistics. Barndorff-Nielsen [2] uses a statistic A which, for fixed $\hat{\theta}$, is an affine (i.e., linear) function of the canonical sufficient statistics and is approximately free of θ in its mean and variance.

Example. Let *(X, Y)* be bivariate normal with standard mean and variance, and correlation θ. From an identical independent sample, the sufficient statistic for θ is

$$S = \sum_{i=1}^{n}(X_iY_i, X_i^2 + Y_i^2).$$

In this case S_2 has mean $2n$, entirely free of θ, but has variance $4n(1+\theta^2)$. Then

$$A = \frac{S_2 - 2n}{\sqrt{4n(1+\theta_0^2)}}$$

is first-order local ancillary for θ near θ_0. A locally conditional confidence region for θ can be given approximately by inverting an Edge-worth expansion for the conditional distribution of S_1 given A; a general expression is given by Cox [7].

Information Recovery

When the sufficient statistic S is expressed as (T, A), then, using standard notation,

$$E\{I_{T|A}(\theta)\} = I_S(\theta) - I_A(\theta).$$

Thus, conditioning on approximate ancillaries involves loss of information, on average equal to the amount of information in the conditioning variable. On the other hand, the conditional analysis delivers a more relevant evaluation of the precision of estimation, as discussed earlier.

An alternative requirement of an approximate ancillary A is hence that the Fisher information* in it should be small. Statistics typically have information of order n in the sample size. A first-order approximate ancillary has information $O(1)$. This can be achieved by requiring that the derivative of *E(A)* with respect to θ should be $O(n^{-1/2})$ (Lloyd [17]). By arranging a special condition on the variance and skewness of A and their derivatives, the information can be reduced to $O(n^{-1})$. Where a statistic is first-order ancillary not only locally but globally, it will be first-order approximate ancillary. Conversely, a statistic that is first-order approximate ancillary locally will be first-order local ancillary.

Balancing Information Loss Against Relevance

Cox's criterion judges the effectiveness of A as a conditioning variable by how large $\mathrm{Var}(I_{T|A})$ is. On the other hand, information I_A is lost on average. An exact ancillary may be quite ineffective while another statistic, which is only approximately ancillary, is much more effective. The statistic most "θ-free" in its distribution is not necessarily the best conditioning variable.

It is not clear how to balance the effectiveness against the information loss. Some progress has been made (Lloyd [18]); however, at present there seems to be no entirely objective way of making this tradeoff.

Nonuniqueness of Conditional Inference*

McCullagh [20] has given an example where the minimal sufficient statistic* may be expressed either as $(\hat{\theta}, A)$ or as $(\hat{\theta}, A^*)$, where A and A^* have exactly the same distributions and where $\mathrm{var}(I_{T|A})$ is identical for both A and A^*. Thus there is no hope of deciding which to condition on from consideration of marginal or conditional distributions. Further, (A, A^*) together are not ancillary.

The example is beautifully simple. An independent sample is taken from the Cauchy distribution with location θ_1 and scale θ_2. The entire parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is of interest. The configuration ancillary A is the n-vector with ith component $(x_i - \hat{\theta}_1)/\hat{\theta}_2$. Fisher [13] and Pitman [22] recommend using the distribution of $\hat{\boldsymbol{\theta}}$ conditional on the configuration A. Now for any real numbers a, b, c, d the transformed data

$$X_i^* = \frac{aX_i + b}{cX_i + d}$$

is also Cauchy in distribution with parameters $\boldsymbol{\theta}^*$, a simple transformation of $\boldsymbol{\theta}$. The X_i^* data are equivalent to the original data; it makes no difference which we use. However the configuration A^* of the X_i^* data is not the same as A. Which density do we use to assess $\hat{\boldsymbol{\theta}}$, the one conditional on A or on A^*?

McCullagh shows that for large deviations, i.e., when $\hat{\theta} - \theta$ is $O(1)$, the two conditional densities differs relatively by $O(n^{-1/2})$, which is of the same magnitude as the difference between conditional and unconditional densities. When the observed value of the configuration statistic is atypically extreme, this difference is even larger. The conclusion is that the general recommendation to condition on exact ancillaries is ambiguous.

The Likelihood-Ratio Statistic

Let $L(\theta)$ be the likelihood function maximized at $\hat{\theta}$. Then the likelihood-ratio statistic is

$$w(\theta_0) = 2 \log\{L(\hat{\theta})/L(\theta_0)\}.$$

Confidence intervals are set by collecting values of θ_0 for which $w(\theta_0)$ is less than a quantile of its sampling distribution which is approximately χ_p^2, where p is the dimension of θ. The approximation improves as cumulants of derivatives of the log-likelihood (usually proportional to sample size n) diverge. A slight adjustment to $w(\theta_0)$, called the Bartlett correction (*see* BARTLETT ADJUSTMENT—I), reduces the error of this distributional approximation to $O(n^{-2})$.

Such intervals are attractive because they produce sensible confidence sets even for anomalous likelihood functions, such as those with multiple maxima, divergent maximum, or regions of zero likelihood. Classical confidence regions based on the estimator and standard error are, in contrast, always elliptical in shape. Thus, likelihood-based intervals seem, in an informal sense, to better summarize what the data say about the parameter. Making inference more relevant to the particular data set is a fundamental reason for conditioning.

It has been shown that to high order, $w(\theta_0)$ is independent of the local ancillary $A(\theta_0)$, and when there are several choices for $A(\theta_0)$ it is approximately independent of all of them; see Efron and Hinkley [11], Cox [7], McCullagh [19]. Intervals based directly on the likelihood-ratio statistic are therefore automatically approximately conditional on the local ancillary, whereas alternative approaches via the score or Wald statistics are not. Seen another way, conditioning on the local ancillary leads to inference that agrees qualitatively with the likelihood function.

Approximate Conditional Distributions

Once an appropriate conditioning statistic is identified, there remains the problem of computing the conditional distribution of the information-carrying statistic. A remarkable amount such conditional distributions in more recent years. Among several remarkable properties of these approximations, the most important is that the conditioning statistic need not be explicitly identified.

Let A be an approximately ancillary statistic, as yet unspecified explicitly. Barndorff-Nielsen [2,3] gives the approximation

$$P^{*}_{\hat{\theta}|A=a}(t) = c|\hat{J}|^{1/2}\frac{L(t)}{L(\hat{\theta})} \qquad (1)$$

to the density function of $\hat{\theta}$ given $A = a$, where $\hat{J}$ is minus the second derivative matrix of the log-likelihood function evaluated at the maximum, and $|\hat{J}|$ denotes the determinant of this nonnegative definite matrix. The constant *c(a, t)* is a norming constant, although in important cases it does not depend on the argument t. The norming constant possibly excepted, the formula is well suited to practical use, involving only a knowledge of the likelihood function.

The p^*-formula* is a synthesis and extension of the results of Fisher [13] and Pitman [22] for location and scale parameters. They showed that the distribution of the estimator conditional on the configuration ancillary was essentially the likelihood function itself, renormalized to integrate to 1. The p^*-formula, while approximate in general, establishes a similar link between conditional inference* and the shape of the likelihood function. The presence of $|\hat{J}|$ allows for the particular parametrization.

Transformation models are models that are closed under the action of a group of transformations, and that induce the same transformation on the parameter. An example is a location model where translation of the data translates the mean by the same amount. For transformation models an exact analogue of the configuration ancillary exists and the p^*-formula gives the exact conditional distribution. A curved exponential family is a multidimensional exponential family with a smooth relation between the canonical parameters. In this case the p^*-formula gives the distribution of the maximum-likelihood estimator conditional on the affine ancillary with accuracy $O(n^{-3/2})$ and the norming constant is a simple power of 2π. This has been established by relating the formula to the so-called saddle-point approximation*.

ANCILLARIES WITH NUISANCE PARAMETERS*

Let λ now represent the unknown parameter vector for a set of data $\{x\}$ with density function $f(x; \lambda)$. Only rarely are all components of λ of equal interest; we are typically interested in separately summarizing what the data say about specific components of interest.

Consider the case where $\lambda = (\theta, \phi)$. We take θ to be the parameter of interest and ϕ a nuisance (sometimes also called accessory) parameter. An inferential statement about θ is desired without regard to the value of ϕ. For instance we require a confidence region for θ whose coverage is at least close to nominal not only for all θ but for all ϕ.

A host of methods exist for achieving this under specific circumstances and a wide ranging survey was given by Basu [5]. The use of ancillary statistics, defined in an extended and appropriate sense, is central to many of these methods.

Pivotals and External Ancillaries

If R is a $100(1-\alpha)\%$ confidence region for θ, then the indicator function of the event $\theta \in R$ is a binary random variable taking value 1 with probability $1 - \alpha$. It thus has a distribution free of both θ and ϕ. A function $P(\{x\}, \theta)$ which has a distribution free of λ is called a *pivotal*. When θ is known, P is an ancillary statistic in the model $f(x; \theta, \phi)$ with θ known, i.e. ancillary with respect to ϕ. Thus a necessary and sufficient condition for construction of a confidence interval for θ is that for any θ one can construct an ancillary with respect to ϕ.

Note that the pivotal is used directly for inference on θ, for instance by collecting all those values of θ for which $P(\{x\}, \theta)$ lies within the central part of its known distribution. There is no suggestion that it be conditioned on, and the rationale for its construction is quite distinct from arguments given earlier.

While such an ancillary should be a function of the minimal sufficient statistic for λ, the sufficiency principle would not require it to be a function of the minimal sufficient statistic for ϕ. External ancillary statistics therefore have some relevance to the elimination of nuisance parameters from inference. In a wide class of cases, external ancillaries will actually be independent of the sufficient statistic for ϕ; see Basu [4].

For continuous distributions admitting a complete sufficient statistic, an external ancillary may always be constructed as the distribution function of the data, conditional on the sufficient statistic. This construction is discussed by Rosenblatt [23] and extends to the construction of a pivotal by considering the interest parameter θ known. The pivotal is the distribution function of the data conditional on a complete sufficient statistic for ϕ. This “statistic” may also depend on θ; see Lloyd [16] for theory and for the following example.

Example. Let X, Y be independently distributed exponentially with means ϕ^{-1} and $(\theta + \phi)^{-1}$ respectively. From a sample of n independent pairs, inferences about θ are required, free of the nuisance parameter ϕ. The sufficient statistic for ϕ, whose distribution also involves θ, is $S = \Sigma(X_i + Y_i)$. (In other examples, S may even involve θ in its definition.) The conditional distribution function of the remaining element of the minimal sufficient statistics $T = \Sigma X_i$ is

$$\int_{s-x}^{s} g(u, \theta)\, du \Big/ \int_{0}^{s} g(u, \theta)\, du,$$

where $g(u, \theta) = [u(s-u)]^{n-1} e^{-\theta u}$. For $n = 1$ the expression simplifies to

$$P(X, \theta) = \frac{1 - e^{-\theta X}}{1 - e^{-\theta(X+Y)}},$$

a probability that gives directly the significance of any proposed value of θ.

For the normal and gamma distributions, which admit complete sufficient statistics for location and scale respectively, the independence of external ancillaries (location- and scale-free statistics, respectively) are particularly important and useful, as the following examples show.

Example. A sample of size n is drawn from a population of unknown mean μ and standard deviation σ and is to be tested for normality. The mean $\overline{X}$ and standard deviation S of the sample are calculated. Then a sample of n from a standard normal population is drawn, and its mean $\overline{X}^*$ and standard deviation S^* calculated. The standardized members of the original sample are

$$\frac{X_i - \mu}{\sigma} = \frac{X_i - \overline{X}}{\sigma} + \frac{\overline{X} - \mu}{\sigma}.$$

Now $\overline{X}$ is complete sufficient for μ, and the external ancillary comprises the statistics $X_i - \overline{X}$, which are independent of $\overline{X}$ and therefore of the last term above. Thus the distribution is unchanged if we replace it with the equidistributed $\overline{X}^*$. Further,

$$\frac{X_i - \overline{X}}{\sigma} = \frac{X_i - \overline{X}}{S} \frac{S}{\sigma}.$$

The statistics $A_i = (X_i - \overline{X})/S$ are external ancillary for (μ, σ) and independent of both $\overline{X}$ and S, and so we may replace S/σ by the equidistributed S^*. Hence finally we have

$$\frac{X_i - \mu}{\sigma} \overset{d}{=} A_i S^* + \overline{X}^*,$$

and these statistics are available for testing normality in the absence of knowledge of the unknown parameters. This device is due to Durbin [10], who used a different method of proof.

Example. In some sampling situations the probability of an item appearing in the sample is inversely proportional to its measured values (such as lifetime), and a weighted mean $W = S_2/S_1$, where $S_1 = \Sigma X_i$ and $S_2 = \Sigma X_i^2$, is a reasonable estimator. The expectation and other moments of this estimator are readily determined if the X_i are $\Gamma(m)$-distributed, where the scale parameter is suppressed. The effect of the method of sampling is to reduce the shape parameter to $m - 1$. For a sample of n, the total S_1 is $\Gamma(n(m-1))$-distributed, is sufficient for the scale parameter, and is independent of any scale-free statistic, such as $W/S_1 = S_2/S_1^2$. Hence $E(W) = E(W/S_1)E(S_1)$ and $E(S_2) = E(W/S_1)E(S_1^2)$, so that, in terms of the moments,

$$E(W) = \frac{E(S_1)S(S_2)}{E(S_1^2)} = \frac{nm(m-1)}{nm - n + 1}.$$

Thus W has a small negative bias, $O(n^{-1})$.

Ancillarity in the Presence of a Nuisance Parameter

Suppose we have the following factorization of the density function:

$$f(x; \lambda) = f(x; a, \theta) f(a; \theta; \phi)$$

for some statistic A. The first factor is the density conditional on $A = a$, and the second the marginal distribution of A; the salient feature is that the first factor depends on λ only through θ.

It is generally agreed that we must have this feature present in the factorization of $f(x; \lambda)$ before we can say that A is ancillary for θ in any sense (Barndorff-Nielsen [1]). An alternative description is that A is "sufficient" for ϕ. However, the requirement that the marginal distribution of A be completely free of θ seems too strong. Even when the distribution of A depends on θ, A may be incapable of providing information (in the common usage of the term) about θ by itself, because information concerning θ cannot be separated from ϕ in the experiment described by $f(a; \theta, \phi)$. One definition that makes this notion precise requires

1. that for every pair θ_1, θ_2 and for every $a, f(a; \theta_1, \phi)/f(a; \theta_2, \phi)$ run through all positive values as ϕ varies,
2. that given possible values θ_1, θ_2, ϕ, and a, there exist possible values θ, ϕ_1, ϕ_2 such that

$$\frac{f(a; \theta_1, \phi)}{f(a; \theta_2, \phi)} = \frac{f(a; \theta, \phi_1)}{f(a; \theta, \phi_2)}.$$

In this case we say that A is ancillary for θ and inference should be carried out using the distribution conditional on $A = a$ (see ref. 1). Under slightly different conditions, Godambe [14] has shown that the conditional likelihood leads to an estimating function* that is optimal amongst those not depending on ϕ. Yip [24,25] provides examples of this type where the information loss is negligible.

Approximate Conditional Likelihood

Maximum-likelihood estimates are not always consistent. The first example was given by Neyman and Scott [21]. Another simple example involves pairs of twins, one of each pair being given a treatment that increases the odds of a binary response by θ. The underlying rate of response ϕ_1 is specific to each pair. As more twins are collected, $\hat{\theta}$ converges to θ^2 rather than θ. This example is not artificial. Moreover, maximum-likelihood estimation is generally poor whenever a model involves a large number of nuisance parameters.

Let $L(\theta, \phi)$ be the likelihood function maximized at $(\hat{\theta}, \hat{\phi})$, and let $\hat{\phi}_0$ partially maximize L with respect to ϕ for fixed θ. The *profile likelihood* function

$$L_p(\theta) = L(\theta, \hat{\phi}_\theta)$$

shares many logical properties of likelihood and is maximized at $\theta = \hat{\theta}$. However, we have seen that it may perform very poorly. When a sufficient statistic A for ϕ exists, the density function $f(x; a, \theta)$ conditional on $A = a$ is free of ϕ. The *conditional likelihood* $L_c(\theta)$ is just this conditional density considered as a function of θ. The conditional likelihood avoids the bias and consistency problems of the unconditional likelihood. Thus, in the presence of nuisance parameters, another role of conditioning on the ancillary A is to reduce the bias of estimation. To perform this alternative function A need not be "free of θ" in any sense.

When the model $f(x; \lambda)$ is an exponential family with canonical parameter θ, such a conditional likelihood is always obtained simply by conditioning on $\hat{\phi}$. The resulting conditional likelihood can be treated much as an ordinary likelihood, and standard asymptotic theory often applies, conditional on any value $\hat{\phi}$. Davison [9] has applied these ideas to generalized linear models* as well as the approximation to be described below. Appropriately conditional inferences are easily computed within a standard computer package.

More generally, when a sufficient statistic $A(\theta_0)$ for θ exists for fixed θ_0, the conditional likelihood may be calculated. Barndorff-Nielsen [3] gives an approximation, called *modified profile likelihood*, which is

$$L_M(\theta) \approx \left| \frac{\partial \hat{\phi}}{\partial \hat{\phi}_\theta} \right| |\hat{J}_\theta|^{-1/2} L_p(\theta),$$

where $\hat{J}_\theta$ is the observed information for ϕ when θ is known. The only difficult factor to compute is the first, which can be made to approximately equal unity if ϕ is chosen to be orthogonal to θ; see Cox and Reid [8]. The factor involving j_θ penalizes values of θ that give relatively high information about ϕ. In examples where the estimate $\hat{\theta}$ is inconsistent, use of $L_M(\theta)$ reduces the inconsistency but may not completely eliminate it. The attraction is that conditioning has implicitly been performed merely by directly using a likelihood function, albeit approximately.

It is unclear what the conceptual and practical properties of $L_M(\theta)$ are when no sufficient statistics for ϕ exist (even when allowed to be functions of θ). There are also cases where conditioning produces a degenerate distribution, primarily when the data are discrete. A simple example is a logistic regression* where interest is in the intercept. The use of modified profile likelihood in such cases has not yet been fully justified.

REFERENCES

1. Barndorff-Nielsen, O. (1978). *Information and Exponential Families*. Wiley, New York.
2. Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika*, **67**, 293–310.
3. Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.
4. Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377–380.
5. Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Ass.*, **72**, 355–366.
6. Buehler, R. J. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Ass.*, **77**, 582–594.
7. Cox, D. R. (1980). Local ancillarity. *Biometrika*, **67**, 279–286.
8. Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. B*, **49**, 1–39.
9. Davison, A. (1988). Approximate conditional inference in generalised linear models. *J. Roy. Statist. Soc. B*, **50**, 445–461.
10. Durbin, J. (1961). Some methods of constructing exact tests. *Biometrika*, **48**, 41–55.
11. Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*, **65**, 457–487.
12. Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.*, **22**, 700–725.
13. Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A*, **144**, 285–307.
14. Godambe, V. P. (1976). Conditional likelihood and unconditional optimal estimating equations. *Biometrika*, **63**, 277–284.
15. Lehmann, E. L. and Scholz, F. W. (1991). Ancillarity. In *Current Issues in Statistical Inference: Essays in honour of D. Basu*, M. Ghosh and P. K. Pathak, eds., IMS Lecture Notes—Monograph Series, pp. 32–51.
16. Lloyd, C. J. (1985). On external ancillarity. *Austral. J. Statist.*, **27**, 202–220.
17. Lloyd, C. J. (1991). Asymptotic expansions of the Fisher information in a sample mean. *Statist. Probab. Lett.*, **11**, 133–137.
18. Lloyd, C. J. (1992). Effective conditioning. *Austral. J. Statist.*, **34**(2), 241–260.

19. McCullagh, P. (1984). Local sufficiency. *Biometrika*, **71**, 233–244.

20. McCullagh, P. (1991). *On the Choice of Ancillary in the Cauchy Location Scale Problem. Tech. Rep. 311*, University of Chicago.

21. Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1–32.

22. Pitman, E. J. G. (1939). The estimation of the location and scale parameter of a continuous population of any given form. *Biometrika*, **30**, 391–421.

23. Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.*, **23**, 470–472.

24. Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *Austral. J. Statist.*, **30**, 299–306.

25. Yip, P. (1991). Conditional inference on a mixture model for the analysis of count data. *Commun. Statist. Theory Methods*, **20**, 2045–2057.

See also Ancillary Statistics, First Derivative; Conditional Inference; Nuisance Parameters; and p^*-Formula.

Christopher J. Lloyd
Evan J. Williams
Paul S. F. Yip

ANCILLARY STATISTICS, FIRST DERIVATIVE

The usual definition of an ancillary statistic* $a(y)$ for a statistical model $\{f(y;\theta):\theta\in\Omega\}$ requires that the distribution of $a(y)$ be fixed and thus free of the parameter θ in Ω. A first derivative ancillary (at θ_0) requires, however, just that the first derivative at θ_0 of the distribution of $a(y)$ be zero; in an intuitive sense this requires ancillarity for θ restricted to a small neighborhood $(\theta_0\pm\delta)$ of θ_0 for δ sufficiently small. The notion of a first derivative ancillary was developed in Fraser [4] and named in Fraser and Reid [6].

Fisher's discussion and examples of ancillaries [3] indicate that he had more in mind than the fixed θ-free distribution property, but this was never revealed to most readers' satisfaction. One direction was to require that y in some sense measure θ, as in the location model (*see* Ancillary Statistics—I); this notion was generalized in terms of structural models* which yield a well-defined ancillary and a well-defined conditional distribution to be used for conditional inference*. In a related manner the usefulness of first derivative ancillaries arises in a context where the variable y in a sense measures θ, locally at a value θ_0, and particularly in some asymptotic results that will be described briefly below.

First we give a simple although somewhat contrived example of a first derivative ancillary. Consider a sample $(y_1,\ldots,y_n)$ from the model $\varphi(y-\theta)(1+y^2\theta^2)/(1+\theta^2+\theta^4)$, where φ denotes the standard normal density. Then $(y_1-\bar{y},\ldots,y_n-\bar{y})$ is first derivative ancillary at $\theta=0$, as this model agrees with the model $\varphi(y-\theta)$ at $\theta=0$, and the first derivatives with respect to θ of the two models are also identical at $\theta=0$. Also $(y_1-\bar{y},\ldots,y_n-\bar{y})$ is the standard configuration ancillary for the model $\varphi(y-\theta)$; *see* Ancillary Statistics—I.

There are a number of likelihood-based methods that lead to highly accurate test procedures for scalar parameters. The typical context has a continuous model $f(y;\theta)$, where y has dimension n and θ has dimension p, and interest centers on a component scalar parameter ψ, where $\theta=(\lambda,\psi)$, say.

In cases where the dimension of the minimal sufficient statistic is the same as the dimension of θ, as in an exponential model or via conditionality with a location or transformation model, there are now well-developed, very accurate methods for obtaining a p-value, $p(\psi_0)$, for testing the value ψ_0 for the parameter ψ; see, for example, Barndorff-Nielsen [1], Fraser and Reid [5,6], and Reid [8].

Extension to the $n>p$ case requires a dimension reduction by sufficiency or conditionality. In this context, sufficiency turns out to be too specialized and conditionality requires an ancillary or approximate ancillary statistic. While suitable approximate ancillary statistics have long been known to exist, there seems not to be any generally applicable construction method. Particular examples are discussed in Barndorff-Nielsen and Cox [2]. The first derivative ancillary [6] provides the ingredients for a simple construction method.

Consider an independent coordinate y of a model with scalar parameter θ, and suppose the distribution function $F(y;\theta)$ is stochastically monotone at θ_0. Then a transformation $x = x(y)$ to a new variable with distribution function $G(x; 0)$ exists such that $G(x;\theta)$ and $G(x - (\theta - \theta_0); \theta_0)$ agree up to their first derivatives at $\theta = \theta_0$. If this property holds for each coordinate, then $(x_1 - \overline{x}, \ldots, x_n - \overline{x})$ is first derivative ancillary at $\theta = \theta_0$, as in the simple example above.

The conditional distribution given this ancillary has tangent direction $(1, \ldots, 1)'$ in terms of the x-coordinates and tangent direction $\upsilon = (\upsilon_1, \ldots, \upsilon_n)'$ in terms of the y-coordinates, where

$$\upsilon_i = \frac{\partial F_i(y_i;\theta)}{\partial\theta} \Bigg/ \frac{\partial F_i(y_i;\theta)}{\partial y_i}\Bigg|_{(y^0,\hat{\theta}^0)};$$

the calculation is at the observed data y^0 with corresponding maximum likelihood estimate $\hat{\theta}^0$, and the subscript i is to designate the ith coordinate distribution function. More concisely, we can write $\upsilon = \partial y/\partial\theta|_{(\tilde{y}^0,\hat{\theta}^0)}$, where the differentiation is for a fixed value of the distribution function. In most problems this is easy to calculate.

An interesting feature of the above type of first derivative ancillary is that it can be adjusted to give an approximate ancillary statistic to third order without altering its tangent directions at the data point. (A third-order approximate ancillary statistic is a statistic whose distribution is free of θ in a specific sense: see Reid [7]). It turns out that for approximation of p-values to third order [meaning with relative error $O(n^{-3/2})$], the only needed information concerning the ancillary is information on the tangent directions at the data point. As a result the first derivative ancillary at the maximum likelihood value provides a means to generalize third-order likelihood asymptotics from the case $n = p$ to the typical general case with $n > p$ [6].

REFERENCES

1. Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, **73**, 307–322.
2. Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.
3. Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed. Oliver & Boyd, Edinburgh. (Includes a discussion of ancillary statistics with many examples.)
4. Fraser, D. A. S. (1964). Local conditional sufficiency. *J. R. Statist. Soc. B*, **26**, 52–62.
5. Fraser, D. A. S. and Reid, N. (1993). Simple asymptotic connections between densities and cumulant generating function leading to accurate approximations for distribution functions. *Statist. Sinica*, **3**, 67–82.
6. Fraser, D. A. S. and Reid, N. (1995). Ancillaries and third-order significance. *Utilitas Math.*, **47**, 33–53.
7. Reid, N. (1995). The roles of conditioning in inference. *Statist. Sci.*, **10**, 138–157.
8. Reid, N. (1996). Asymptotic expansions. In *Encyclopedia of Statistical Sciences Update*, vol. 1, S. Kotz, C. B. Read, and D. L. Banks, eds. Wiley, New York, pp. 32–39.

See also ANCILLARY STATISTICS—I; ASYMPTOTIC EXPANSIONS—II; CONDITIONAL INFERENCE; STRUCTURAL MODELS; and SUFFICIENT STATISTICS.

D. A. S. FRASER
N. REID

ANDERSON–DARLING TEST FOR GOODNESS OF FIT. See GOODNESS OF FIT, ANDERSON–DARLING TEST OF

ANDERSON, OSKAR NIKOLAEVICH

Born: August 2, 1887, in Minsk, Russia.

Died: February 12, 1960, in Munich, Federal Republic of Germany.

Contributed to: correlation analysis, index numbers, quantitative economics, sample surveys, time-series analysis, nonparametric methods, foundations of probability, applications in sociology.

The work and life of Oskar Anderson received a great deal of attention in the periodical statistical literature following his death in 1960. In addition to the customary obituary in the *Journal of the Royal Statistical Society (Series A)*, there was a relatively rare long appreciative article—written by the famous statistician and econometrist H. O. A. Wold,* whose biography also appears in this volume—in the *Annals of Mathematical Statistics*, with an extensive bibliography, and a remarkable obituary and survey of his activities in the *Journal of the American Statistical Association*, written by the well-known statistician G. Tintner, also containing a bibliography. Furthermore there was detailed obituary in *Econometrica*. The first entry in the *International Statistical Encyclopedia* (J. M. Tanur and W. H. Kruskal, eds.) contains a rather detailed biography and an analysis of Anderson's contributions.

Some of this special interest may be associated with unusual and tragic events the first 40 years of Anderson's life, as aptly noted by Wold (1961):

> The course of outer events in Oskar Anderson's life reflects the turbulence and agonies of a Europe torn by wars and revolutions. [In fact Fels (1961) noted that "a daughter perished when the Andersons were refugees; a son a little later. Another son fell in the Second World War."]

Both German and Russian economists and statisticians compete to claim Oskar Anderson as their own. His father became a professor of Finno-Ugric languages at the University of Kazan (the famous Russian mathematician I. N. Lobachevsky, who was born in Kazan, was also a professor at the University). The Andersons were ethnically German. Oskar studied mathematics at the University of Kazan for a year, after graduating (with a gold medal) from gymnasium in that city in 1906. In 1907, he entered the Economics Department of the Polytechnical Institute at St. Petersburg. From 1907 to 1915 he was a assistant to A. A. Chuprov* at the Institute, and a librarian of the Statistical-Geographical "Cabinet" attached to it. He proved himself an outstanding student of Chuprov's, whose influence on Anderson persisted throughout his life. Also, during the years 1912–1917, he was a lecturer at a "commercial gymnasium" in St. Petersburg, and managed to obtain a law degree. Among his other activities at that time, he organized and participated in an expedition in 1915 to Turkestan to carry out an agricultural survey in the area around the Syr Darya river. This survey was on a large scale, and possessed a representativity ahead of contemporary surveys in Europe and the USA. In 1917 he worked as a research economist for a large cooperative society in southern Russia.

In 1917, Anderson moved to Kiev and trained at the Commercial Institute in that city, becoming a docent, while simultaneously holding a job in the Demographic Institute of the Kiev Academy of Sciences, in association with E. Slutskii.*

In 1920, he and his family left Russia, although it was said that Lenin had offered him a very high position in the economic administration of the country. It is possible that his feelings of loyalty to colleagues who were in disfavor with the authorities influenced this decision. For a few years he worked as a high-school principal in Budapest, and then for many years (1924–1943) except for a two-year gap he lived in Bulgaria, being a professor at the Commercial Institute in Varna from 1924 to 1933, and holding a similar position at the University of Sofia from 1935 to 1942. (In the period 1933–1935 he was a Rockefeller Fellow in England and Germany, and his first textbook on mathematical statistics was published.) While in Bulgaria he was very active in various sample surveys and censuses, utilizing, from time to time, the methodology of purposive sampling. In 1940 he was sent to Germany by the Bulgarian government to study rationing, and in 1942, in the midst of World War II, he accepted an appointment at the University of Kiel.

After the war, in 1947, Anderson became a professor of statistics in the Economics Department of the University of Munich, and he remained there till his death in 1960. His son Oskar Jr. served as a professor of economics in the University of Mannheim.

Anderson was a cofounder—with Irving Fisher and Ragnar Frisch—of the Econometric Society. He was also a coeditor of the journal *Metrika*. In the years 1930–1960 he was one of the most widely known statisticians in Central and Western Europe, serving as a link between the Russian and Anglo-American schools in statistics, while working within the German tradition exemplified by such statisticians as Lexis* and von

Bortkiewicz.* He contributed substantially to the statistical training of economists in German universities. His main strengths lay in systematic coordination of statistical theory and practice; he had good intuition and insight, and a superb understanding of statistical problems. His second textbook, *Probleme der Statistischen Methodenlehre*, published in 1954, went through three editions in his lifetime; a fourth edition appeared posthumously in 1962. He was awarded honorary doctorates from the Universities of Vienna and Mannheim and was an honorary Fellow of the Royal Statistical Society.

His dissertation in St. Petersburg, "On application of correlation coefficients in dynamic series," was a development of Chuprov's ideas on correlation. He later published a paper on this topic in *Biometrika*, and a monograph in 1929. While in Varna, he published a monograph in 1928 (reprinted in Bonn in 1929), criticizing the Harvard method of time-series analysis, and developed his well-known method of "variate differences" (concurrently with and independently of W. S. Gosset*). This method compares the estimated variances of different orders of differences in a time series to attempt to estimate the appropriate degree of a polynomial for a local fit. In 1947, Anderson published a long paper in the *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, devoted to the use of prior and posterior probabilities in statistics, aiming at unifying mathematical statistics with the practices of statistical investigators. Anderson was against abstract mathematical studies in economics, and often criticized the so-called "Anglo-American school," claiming that the Econometric Society had abandoned the goals originally envisioned by its founders.

During the last period of his life, he turned to nonparametric methods, advocating, *inter alia*, the use of Chebyshev-type inequalities, as opposed to the "sigma rule" based on assumptions of normality. Some of his endeavors were well ahead of his time, in particular, his emphasis on causal analysis of nonexperimental data, which was developed later by H. Wold,* and his emphasis on the importance of elimination of systematic errors in sample surveys. Although he had severed physical contact with the land of his birth as early as 1920, he followed with close attention the development of statistics in the Soviet Union in the thirties and forties, subjecting it to harsh criticism in several scorching reviews of the Marxist orientation of books on statistics published in the USSR.

REFERENCES

1. Anderson, O. N. (1963). *Ausgewáhlte Schriften*, H. Strecker and H. Kellerer, eds. Mohr Verlag, Tübingen, Germany. 2 vols. (These collected works, in German, of O. N. Anderson include some 150 articles, originally written in Russian, Bulgarian, and English as well as German. They are supplemented by a biography.)
2. Fels, E. (1961). Oskar Anderson, 1887–1960. *Econometrica*, **29**, 74–79.
3. Fels, E. (1968). Anderson, Oskar N. In *International Encyclopedia of Statistics*, J. M. Tanur and W. H. Kruskal, eds. Free Press, New York.
4. Rabikin, V. I. (1972). O. Anderson—a student of A. A. Chuprov, *Uchenye Zapiski po Statistike*, **18**, 161–174.
5. Sagoroff, S. (1960). Obituary: Oskar Anderson, 1887–1960, *J. R. Statist. Soc. A.* **123**, 518–519.
6. Tintner, G. (1961). The statistical work of Oscar Anderson. *J. Amer. Statist. Ass.*, **56**, 273–280.
7. Wold, H. (1961). Oskar Anderson: 1887–1960, *Ann. Math. Statist.*, **32**, 651–660.

ANDREWS FUNCTION PLOTS

Function plots are displays of multivariate data in which all dimensions of the data are displayed. Each observation is displayed as a line or function running across the display. The plots are useful in detecting and assessing clusters* and outliers*. Statistical properties of the plots permit tests of significance* to be made directly from the plot.

The display of data of more than two dimensions requires special techniques. Symbols may be designed to represent simultaneously several dimensions of the data. These may be small symbols used in a scatter plot with two dimensions of the data giving the location of the symbol of the page. Anderson [1] gives examples of such glyphs*. Patterns involving one or more of the plotting dimensions are most easily detected.

Alternatively, these may be larger symbols displayed separately. Chernoff faces* are an example of this type. Although no two dimensions have special status as plotting coordinates, the detection of patterns is more awkward. Function plots are a method of displaying large (page size) symbols simultaneously.

CONSTRUCTION OF PLOTS

Although only few statisticians have experience in displaying items of more than three dimensions, all statisticians are familiar with displays of functions $f(t)$. These may be considered as infinite dimensional. This suggests a mapping of multivariate data, observation by observation, into functions and then displaying the functions. Many such mappings are possible, but the mapping proposed here has many convenient statistical properties.

For each observation involving k dimensions $\mathbf{x}' = (x_1, \ldots, x_k)$, consider the function

$$f_x(t) = x_1\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots$$

plotted for values $-\pi < t < \pi$. Each observation contributes one line running across the display. The completed display consists of several such lines.

STATISTICAL PROPERTIES

The mapping $x \to f_x(t)$ preserves distances. For two points x, y the equation

$$\sum_{i=1}^{k}(x_i - y_i)^2 = \pi^{-1}\int_{-\pi}^{\pi}\left[f_x(t) - f_y(t)\right]^2 dt$$

implies that two functions will appear close *if and only if* the corresponding points are close.

The mapping is linear. This implies that

$$\overline{f}_x(t) = f_{\overline{x}}(t).$$

If the data have been scaled so that the variates are approximately independent with the same variance σ^2, then the variance of $f_x(t)$ is constant, independent of t, or almost constant. Since

$$\begin{aligned}\mathrm{var}(f_x(t)) &= \sigma^2(\tfrac{1}{2}\pi + \sin^2 t + \cos^2 t \\ &\quad + \sin^2 2t + \cos^2 2t + \ldots) \\ &= \sigma^2(\tfrac{1}{2} + k/2 + R),\end{aligned}$$

where $R = 0$ if k is odd and $\sin^2[(k+1)/2]$ if k is even. This relation may be used to produce and display confidence bands* and

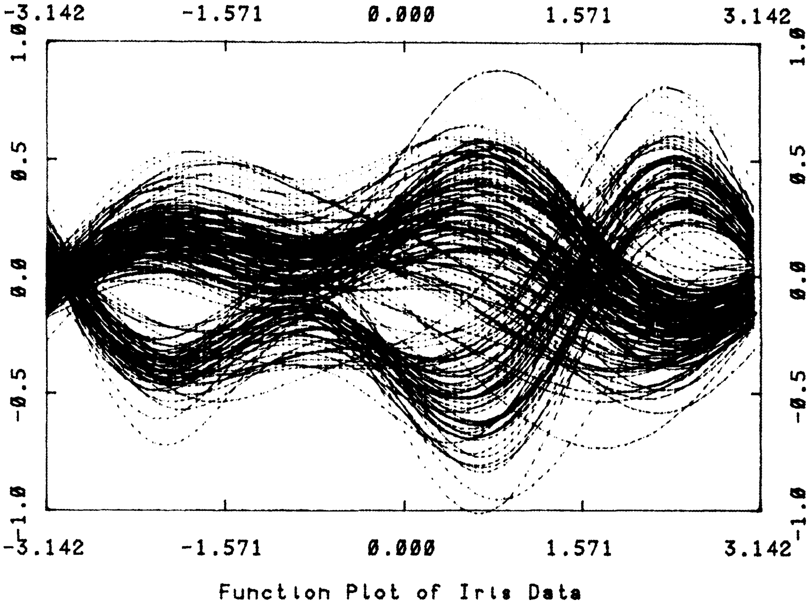


Figure 1. All species—150 observations.

tests for outliers*. These tests may be made for preselected values of t or marginally for all values of t. Scheffé's method of multiple comparison* may be used here.

EXAMPLE

Figure 1 is a plot of the Fisher iris data. These data consist of observations of four variables (log units) on 150 iris flowers. The example is commonly used to demonstrate multivariate techniques. Figure 1 clearly demonstrates the separation of one group. This group consists of one species. This is verified in Fig. 2, which is the plot of this species alone. Note the presence of two "outliers" represented by two straggling lines.

FURTHER NOTES

In some applications, with large data sets, the data may be summarized for each value of t by selected order statistics* of the values $f_x(t)$. Thus a complex plot may be reduced to a plot of the median, the quartiles, the 10% points, and the outlying observations. The order statistics were chosen so that the lines will be almost equally spaced for Gaussian (normal) data.

The order of the variables included in the specification of the function has no effect on the mathematical of statistical properties, although it does affect the visual appearance of the display. Some experience suggests that dominant variables should be associated with the lower frequencies.

GENERAL REFERENCE

1. Anderson, E. (1960). *Technometrics*, **2**, 387–392.

BIBLIOGRAPHY

Andrews, D. F. (1972). *Biometrics*, **28**, 125–136.

Chernoff, H. (1973). *J. Amer. Statist. Ass.*, **68**, 361–368.

Fisher, R. A. (1936). *Ann. Eugen.* (*Lond.*), **7** (Pt. II), 179–188.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

See also CHERNOFF FACES; GRAPHICAL REPRESENTATION OF DATA; and MULTIVARIATE ANALYSIS.

D. F. ANDREWS

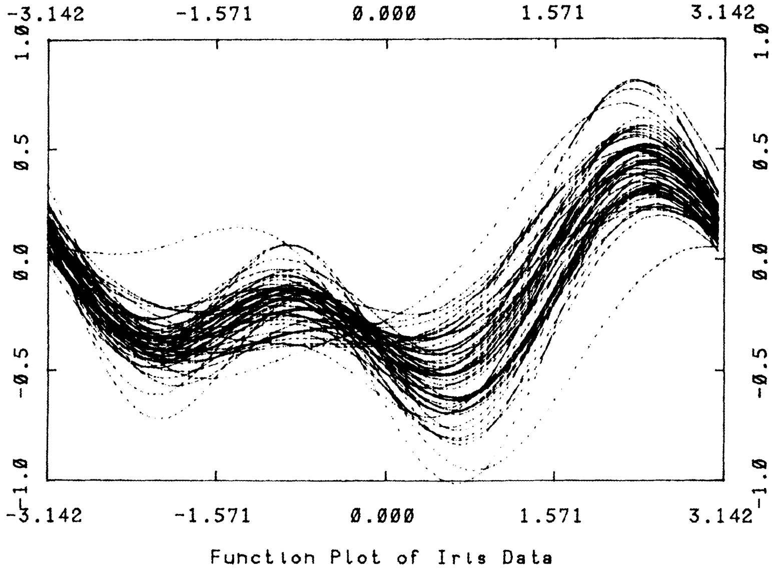


Figure 2. One species—50 observations.

ANGLE BRACKETS

Tukey [4] used angle brackets for symmetric means. These are power product sum* (augmented monomial symmetric functions) divided by the number of terms forming the sum, thus giving the mean power product [5, p. 38].

If observations in a sample are $x_1, \ldots, x_n$, the power product sum $[P] = [p_1 \ldots p_\pi] = \sum_{\neq}^{n} x_i^{p_1} x_j^{p_2} \ldots x_l^{p_\pi}$, where the sum is over all permutations of the subscripts, and no two subscripts are equal. The number of terms forming the sum is $n(n-1)\ldots(n-\pi+1) = n^{(\pi)}$. The symmetric mean or angle bracket is then $\langle P\rangle = [P]/n^{(\pi)}$. Thus

$$\langle 1\rangle = \frac{1}{n}[1] = \frac{1}{n}\sum_{1}^{n} x_i = \bar{x},$$

$$\langle r\rangle = \frac{1}{n}[r] = \frac{1}{n}\sum_{1}^{n} x_i^r = m_r',$$

$$\langle rs\rangle = \frac{1}{n^{(2)}}[rs] = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} x_i^r x_j^s,$$

$$\langle 11\rangle = \frac{1}{n^{(2)}}[11] = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} x_i x_j.$$

But

$$[1]^2 = [2] + [11]$$

as

$$\left(\sum_{1}^{n} x_i\right)^2 = \sum_{1}^{n} x_i^2 + \sum_{i\neq j}^{n} x_i x_j.$$

Hence

$$\langle 11\rangle = \frac{1}{n-1}\{n\langle 1\rangle^2 - \langle 2\rangle\}.$$

Tukey [4] and Schaeffer and Dwyer [3] give such recursion formulas for the computation of angle brackets. Elementary examples of computing angle brackets are given in Keeping [2]. Two angle brackets may be multiplied by the rule

$$\langle r\rangle\langle s\rangle = \frac{n-1}{n}\langle rs\rangle + \frac{1}{n}\langle r+s\rangle.$$

A similar symmetric mean may be defined for the population $x_1, \ldots, x_N$, and denoted by $\langle P\rangle_N = [P]_N/N^{(\pi)}$, where $[P]_N$ denotes the power product sum $\sum_{\neq}^{N}$ for the population. Then, if E_N denotes the expected value for the finite population, it follows from an argument of symmetry [3,4] that $E_N\langle P\rangle = \langle P\rangle_N$. Tukey [4] calls this property "inheritance on the average." It makes angle brackets attractive in the theory of sampling from finite populations*, as sample brackets are unbiased estimates* of corresponding population brackets.

Every expression that is (1) a polynomial, (2) symmetric, and (3) inherited on the average can be written as a linear combination of angle brackets with coefficients that do not depend on the size of the set of numbers concerned [4].

Since Fisher's k-statistic* k_p is defined as $k_p = \sum_p (-1)^{\pi-l}(\pi-1)!C(P)\langle P\rangle$ and the finite population K-parameter is $K_p = \sum_p (-1)^{\pi-l}(\pi-1)!C(P)\langle P\rangle_N$, it directly follows that $E_N(k_p) = K_p$. For infinite populations, $E\langle P\rangle = \mu_{p_1}' \ldots \mu_{p_\pi}'$, and hence $E(k_p) = \kappa_p$. *See* FISHER'S k-STATISTICS for more details.

Tukey [5] defines polykays* by a symbolic multiplication of the k-statistics written as linear combinations of angle brackets. The symbolic product of the brackets is a bracket containing the elements in the brackets multiplied, i.e., $\langle p_1 \ldots p_\pi\rangle \circ \langle q_1 \ldots q_x\rangle = \langle p_1 \ldots p_\pi q_1 \ldots q_x\rangle$. Thus $k_{21} = k_2 \circ k_1 = \{\langle 2\rangle - \langle 11\rangle\} \circ \langle 1\rangle = \langle 21\rangle - \langle 111\rangle$. Tukey [4] uses angle brackets in the consideration of randomized (or random) sums. Hooke [1] extends them to generalized symmetric means for a matrix.

REFERENCES

1. Hooke, R. (1956). *Ann. Math. Statist.*, **27**, 55–79.
2. Keeping, E. S. (1962). *Introduction to Statistical Inference.* D. Van Nostrand, Princeton, N. J.
3. Schaeffer, E., and Dwyer, P. S. (1963). *J. Amer. Statist. Ass.*, **58**, 120–151.
4. Tukey, J. W. (1950). *J. Amer. Statist. Ass.*, **45**, 501–519.
5. Tukey, J. W. (1956). *Ann. Math. Statist.*, **27**, 37–54.

See also FISHER'S k-STATISTICS; POLYKAYS; and POWER PRODUCT SUMS.

D. S. TRACY

ANGULAR TRANSFORMATION.

See VARIANCE STABILIZATION

ANIMAL POPULATIONS, MANLY–PARR ESTIMATORS

The Manly—Parr estimator of population size can be calculated from data obtained by the capture-recapture method* of sampling an animal population. The assumption made by Manly and Parr [5] is that an animal population is sampled on a series of occasions in such a way that on the ith occasion all the N_i animals then in the population have the same probability p_i of being captured. In that case the expected value of n_i, the number of captures, is $E(n_i) = N_i p_i$, so that

$$N_i = E(n_i)/p_i. \qquad (1)$$

Manly and Parr proposed that p_i be estimated by the proportion of the animals known to be in the population at the time of the ith sample that are actually captured at that time. For example, in an open population (where animals are entering through births and leaving permanently through deaths and emigration) any animal seen before the time of the ith sample and also after that time was certainly in the population at that time. If there are C_i individuals of this type, of which c_i are captured in the ith sample, then $\hat{p}_i = c_i/C_i$ is an unbiased estimator of p_i. Using equation (1) the Manly—Parr estimator of N_i is then

$$\hat{N}_i = n_i/\hat{p}_i = n_i C_i/c_i. \qquad (2)$$

Based upon a particular multinomial* model for the capture process, Manly [4] gives the approximate variance

$$\text{var}(\hat{N}_i) \simeq N_i(1-p_i)(1-\theta_i)/(p_i\theta_i), \qquad (3)$$

where θ_i is the probability of one of the N_i animals being included in the class of C_i animals known to certainly be in the population at the time of the ith sample. This variance can be estimated by

$$\hat{\text{Var}}(\hat{N}_i) = \hat{N}_i(C_i - c_i)(n_i - c_i)/c_i^2. \qquad (4)$$

Manly and Parr also proposed estimators for survival rates and birth numbers in the open population situation. Let r_i denote the number of animals that are captured in both the ith and the $(i+1)$th samples. The expected value of this will be $E(r_i) = N_i s_i p_i p_{i+1}$, where s_i is the survival rate over the period between the two samples for the population as a whole. This relationship together with equation (1) suggests the Manly—Parr survival rate estimator

$$\hat{s}_i = r_i/(n_i\hat{p}_{i+1}), \qquad (5)$$

where $\hat{p}_{i+1} = c_{i+1}/C_{i+1}$.

At the time of the $(i+1)$th sample the population size N_{i+1} will be made up of the survivors from the previous sample time, $N_i s_i$, plus the number of new entries B_i to the population (the births). On this basis the Manly—Parr estimator of the number of births is

$$\hat{B}_i = \hat{N}_{i+1} - \hat{s}_i\hat{N}_i. \qquad (6)$$

As an example of the use of the Manly—Parr equations consider the data shown in Table 1. The values of n_i, c_i, C_i, and r_i that are needed for equations (2), (4), (5), and (6) are shown at the foot of the table. Using these, eq. (2) produces the population size estimates $\hat{N}_2 = 94.5$ and $\hat{N}_3 = 82.9$. The square roots of the estimated variances from equation (4) then give the estimated standard errors $S\hat{e}(\hat{N}_2) = 16.9$ and $S\hat{e}(\hat{N}_3) = 16.7$. Eq. (5) gives the estimated survival rates $\hat{s}_1 = 0.812$ and $\hat{s}_2 = 0.625$. Finally, Eq. (6) produces the estimated birth number $\hat{B}_2 = 23.8$. The data in this example are part of the illustrative data used by Manly and Parr [5].

Seber [7], Southwood [8], and Begon [1] discuss the Manly—Parr method in the context of capture—recapture methods in general. There are two principal competitors for the analysis of data from open populations, the Jolly—Seber method [3,6] and the Fisher—Ford method [2]. The main theoretical advantage of the Manly—Parr approach is that it does not require the assumption that the probability of survival is the same for animals of all ages.

REFERENCES

1. Begon, M. (1979). *Investigating Animal Abundance*. Edward Arnold, London. (An introduction to capture—recapture methods aimed mainly at biologists. There is a lengthy discussion on the relative merits of different methods of analysis for capture—recapture data.)
2. Fisher, R. A. and Ford, E. B. (1947). *Heredity*, **1**, 143–174.
3. Jolly, G. M. (1965). *Biometrika*, **52**, 225–247.
4. Manly, B. F. J. (1969). *Biometrika*, **56**, 407–410.
5. Manly, B. F. J. and Parr, M. J. (1968). *Trans. Soc. Brit. Ent.*, **18**, 81–89.
6. Seber, G. A. F. (1965). *Biometrika*, **52**, 249–259.
7. Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, 2nd. ed. Charles Griffin, London. (This is the standard reference for statistical and mathematical aspects of capture—recapture methods.)
8. Southwood, T. R. E. (1978). *Ecological Methods*. Chapman and Hall, London. (This is a standard text for ecologists. One chapter gives a good survey of both statistical and practical aspects of capture—recapture methods.)

See also Animal Science, Statistics in; Capture–Recapture Methods—I; Ecological Statistics; and Wildlife Sampling.

Bryan F. J. Manly

ANIMAL SCIENCE, STATISTICS IN

INTRODUCTION AND HISTORY

The introduction of modern statistical methods has been a slower process in animal science than in agriculture*. In designing experiments there have been limitations in the maintenance costs and duration of some of the experiments. Sampling methods for estimating animal numbers have been difficult to adopt, because animals, being mobile, may be counted repeatedly.

The first application of statistical methods to animal populations was in 1890 when Weldon [44] showed that the distributions of different measurements (ratios to total length) made on four local races of shrimp (*Crangon vulgaris*) closely followed the normal law. Weldon [45] found that the frontal breadths of Naples crabs formed into asymmetric curves with a double hump. It was this problem of dissection of a frequency distribution into two normal components that led to Karl Pearson's [27] first statistical memoir. The capture—mark—recapture method was first used by Petersen [28] in his studies of the European plaice and the earliest beginnings in the quantitative study of bird populations were made by Jones [21].

Some developments in the planning and analysis of animal experiments and in sampling methods for estimating animal numbers will be reviewed in the following sections.

DESIGN OF EXPERIMENTS

Choice of Experimental Unit

In feeding trials the experimental unit may be a pen. Often one pen of animals is fed on each treatment. The animals are assigned at random to each pen and the treatments are assigned at random to the pens. With a single pen of animals receiving a treatment, the effect of factors such as breed and age cannot be separated from treatment differences to provide a valid estimate of error [19,23].

Change-over Designs

One way of controlling the variation among animals in experiments is to adopt change-over designs* in which different treatments are tested on the same animal, such that each animal receives in sequence two or more treatments in successive experimental periods. The analysis of switchback trials for more than two treatments is dealt with in ref. 24. Incomplete block designs*, where the number of treatments exceeds the number of experimental units per block, are given in refs. 25 and 26.

Complete Block Designs

Randomized block designs (RBD) and Latin squares are common in animal experiments. A series of Latin squares is recommended for use in large experiments on rodents [39].

Alternatively, analysis of covariance* has been used in which a concomitant observation is taken for eliminating the effect of variations arising out of the peculiarities of individual animals. Thus, to test the effect of treatments on milk production of cows, the yield in the previous lactation is used as a concomitant variate to reduce the error in the experimental yield. The method assumes that (i) the regression of y on the concomitant variate x is significant and (ii) x does not vary significantly among treatments.

Scientists often object to use of RBDs or Latin squares*, which make dosing and sampling procedures difficult to operate in practice. In such cases, a standard order of groups may be used within each block, e.g., use of a standard Latin square within a battery of cages in rodent studies.

Split-Plot Designs*

Incomplete blocks or fractional replication designs may be used to eliminate the influence of litter differences from treatment effects when the number of treatments is large. An experiment on mice in which larger units are split into smaller ones is described in ref. 8 to provide increased precision on the more interesting comparisons.

Repeated Measurements* Experiments

A RBD in which the experimental units are animals that receive each treatment in turn is a repeated measurement experiment. This is generally analyzed as a split-plot design, which assumes equal correlation for every pair of subplot treatments. Multivariate methods that do not require this assumption are given in ref. 5.

Regression*

A common problem in animal studies is the estimation of x from the regression of y on x when y is easy to measure but x is difficult and expensive. Thus y and x may represent the length and age of a fish and we may want to estimate the age composition to predict the status of the stock in future years. Other examples are given in ref. 16. A polynomial regression* of average daily gains of each of several Holstein calves on time was fitted in ref. 1 and analysis of variance* (ANOVA) done on the regression coefficients. Nonlinear regressions for describing weight—age relationship in cattle are compared in ref. 3. Fitzhugh [14] reviewed analysis of growth curves* and strategies for altering their shape.

Anova Models

Sometimes observations may be lost due to death or premature drying of lactation. The estimation of variance components* in mixed models with unbalanced data are dealt with in ref. 17 and illustrations from animal experiments are given in ref. 2. Analysis of balanced experiments for linear and normal models is discussed in ref. 18.

Transformations

Often the data violate basic assumptions of the methods of analysis. For violations due to heterogeneity, transformation of the individual observations into another scale may reduce heterogeneity, increase precision in treatment comparisons, and reduce nonadditivity [41]. Illustrations from animal experiments are given in ref. 16.

In management trials with pigs a score of x (out of 100) is assigned to each pig for treatment comparisons. Treatment effects will not be additive and the transformation $\log[(x+\frac{1}{2})/(100+\frac{1}{2}-x)]$ given in ref. 8 would rectify this.

Categorical Data

Log-linear* and generalized least-squares approaches have been used for categorical data* on the incidence of pregnancy in ewes and of buller steers [33]. Log-linear models for capture—recapture* data from closed populations are mainly discussed in refs. 6 and 34; extensions to open populations are in ref. 7.

ESTIMATION OF ANIMAL NUMBERS OR DENSITY

Two important reasons for using sampling methods for estimation of animal numbers or density are (1) limitation of funds, time, and resources and (2) lesser disturbance of the population and its environment than by total count operations. Some important methods for estimating animal numbers will be summarized in the following sections.

Strips, Transects, Or Quadrats

A common method is to count the number within a long rectangle of known area called a *transect*, a short rectangle called a *strip*, or a square called a *quadrat*. The average density per unit area is estimated by taking total counts over randomly chosen areas. The total population is estimated by multiplying average density by the total area of the population. Quadrat sampling* is preferred for some big game species and its size and shape will depend upon the habitat, abundance, and mobility of the species. When a population is randomly distributed, the number(s) of quadrats to be sampled in an area is given by $s = S/(1 + NC^2)$, where S is the total number of quadrats, N is the size of the population being sampled, and C is the coefficient of variation* of $\hat{N}$ [34].

When the population density varies over large areas, stratified sampling with optimum allocation* is recommended. Siniff and Skoog [40] allocated a sampling effort for six caribou strata on the basis of preliminary population estimates of big game in aerial surveys of Alaska caribou conducted in 22,000 square miles of the Nelchina area. The method was based on rectangular sample plots 4 square miles in area on which observers attempted to count all animals. Two important advantages of the method were (i) reduction of the observer bias in sighting by using large transects and (ii) the use of an efficient design through optimum allocation of effort based on available information on caribou distribution. Sampling effort allocated to strata reduced the variance by more than half over simple random sampling*. Other examples are given in refs. 13 and 36. Where it is difficult or time consuming to count all animals in all the sampled quadrats, two-phase sampling using ratio or regression methods may be adopted. Somewhat similar to quadrat sampling in birds and mammals is the use of sampling in estimating fish population in a section of a stream. Sample areas are selected and fish captured by seining or electrical shocking. The areas are screened to prevent the dispersal of the fish while they are being caught. Stratification in weirs or seines [20] is common to the Atlantic herring or sardine fishing off the New England and New Brunswick coasts.

Strip Sampling. Parallel lines one strip width ($2W$, say) apart determine the population of strips. All the n animals observed within the sampled strips are counted. The estimate of total number of animals (N) is given by $A \cdot n/(2LW)$, where L and A represent the length and population area of the strip. Strip sampling involves less risk of repeated counting and is generally recommended in aerial surveys [42,43].

Line Transect Sampling. An observer walks a fixed distance L along a transect or set of transects and records for each of the n sighted animals, its right angle distance $y_i (i = 1, 2, \ldots, n)$ from the transect or its distance r_i from the observer or both. The technique is most useful when the animals move so fast that they can only be seen when they are flushed into the open. Various parametric estimators have been suggested in refs. 15 and CAPTURE–RECAPTURE METHODS—I. Nonparametric estimators are given in refs. 4 and 10. References 15 and 34 deal with consequences for departures from the assumptions. All the estimates are of the form $A \cdot n/(2LW)$, where W is some measure of one-half the "effective width" of the strip covered by the observer as he moves down the transect. The transect method can also be used in aerial surveys. The procedure is the same as for walking or driving and is likewise subject to error. *See also* LINE INTERCEPT SAMPLING and LINE INTERSECT SAMPLING.

Capture—Mark—Recapture (CMR) Methods

A number of M animals from a closed population are caught, marked, and released. On a second occasion, a sample of n individuals are captured. If m is the number of marked animals in the sample, a biased estimate of N and its variance are

$$\hat{N} = \frac{n}{m}M,$$

$$\upsilon(\hat{N}) = \frac{\hat{N}^2(\hat{N} - m)(\hat{N} - n)}{Mn(\hat{N} - 1)}.$$

This was first given in ref. 28 using tagged plaice. Because the coefficient of variation of

$\hat{N}$ is approximately given by $1/m^{1/2}$, it follows that for $\hat{N}$ to be efficient, we should have sufficient recaptures in the second sample. For closed populations $\hat{N}$ appears to be the most useful estimate of N if the basic assumptions [34] underlying the method are met. CAPTURE–RECAPTURE METHODS—I contains recent developments in estimating N for "closed" and "open" populations when marking is carried over a period of time. A capture—recapture design robust to unequal probability of capture is given in ref. 29. The linear logistic binary regression model has been used in ref. 30 to relate the probability of capture to auxiliary variables for closed populations. With tag losses due to nonreporting or misreporting of tag returns by hunters, $\hat{N}$ will be an overestimate of N [35]. A treatment of response and nonresponse errors in Canadian waterfowl surveys is given in refs. 37 and 38; response errors were found larger than the sum of sampling and nonresponse errors.

Change-In-Ratio and Catch-Effort Methods. The Change-In-Ratio method estimates the population size by removing individual animals if the change in ratio of some attribute of the animal, e.g., age or sex composition, is known. For a closed population the maximum-likelihood estimator (MLE) of the population total $N_t (t = 1, 2)$, based on samples n_t at the beginning and end of the "harvested period," is given by

$$\hat{N}_t = \left(R_m - R\hat{P}_t\right) / \left(\hat{P}_1 - \hat{P}_2\right),$$

where R_m and $R_f (R = R_m + R_f)$ are, respectively, the males and females caught during the harvested period $\hat{P}_t = m_t/n_t (t = 1, 2)$, where m_t is the number of males at time t. The method assumes (i) a closed population, (ii) $\hat{P}_t = P = \text{const}$ for all t, and (iii) R_m and R_f are known exactly. A detailed discussion when the assumptions are violated is given in ref. 34. In fisheries attempts have been made to estimate populations using harvest and age structure data [32].

In the catch-effort method, one unit of sampling effort is assumed to catch a fixed proportion of the population. It is shown in refs. 9 and 22 that

$$E(C_t|k_t) = K(N - k_t),$$

where C_t is the catch per unit effort in the tth period, k_t is the cumulative catch through time period $(t - 1)$, K is the catchability coefficient, and N is the initial population size. C_t plotted against k_t provides estimates of the intercept KN and the slope K, whence N can be estimated.

Indices

Indices are estimates of animal population derived from counts of animal signs, e.g., pellet group, roadside count of breeding birds, nest counts, etc. The investigator records the presence or absence of a species in a quadrat and the percentage of quadrats in which the species is observed, giving an indication of its relative importance. Stratified sample surveys are conducted annually in U.S.A. and Canada for measuring changes in abundance of nongame breeding birds during the breeding season [12,31].

Big game populations are estimated by counting pellet groups in sample plots or transects. A method for calibrating an index by using removal data is given in ref. 11.

SUMMARY

A basic problem in experimental design with animals is substantial variation among animals owing to too few animals per unit. Hence, the need for concomitant variables* and their relationship with the observations for reducing the effect of variations due to individual animals in a unit. The alternative approach, to reduce this variation by picking experimental animals from a uniform population is not satisfactory, since the results may not necessarily apply to populations with much inherent variability.

Animal scientists are often not convinced of the importance of good experimentation, so that laboratory experiments are not necessarily the best to detect small differences. Closer cooperation is needed between the animal scientist and the statistician to yield the best results. Where the underlying assumptions are false, it is important to investigate whether this would invalidate the conclusions. Animal scientists should be alert to the severity of bias in the use of regression for calibration of one variable as an indicator for another when the problem is an inverse one and the relation between the variables is not very close.

In experiments with repeated observations, with successive observations correlated on the same animal, the animal scientist should consult the statistician in the use of the most appropriate method of analysis.

Sampling techniques for estimating the size of animal populations are indeed difficult to implement since animals often hide from us or move so fast that the same animal may be counted more than once. The choice of method would depend on the nature of the population, its distribution, and the method of sampling. Where possible, the design should be flexible enough to enable the use of more than one method of estimation.

In the past, the Petersen method has been used with too few captures and recaptures, leading to estimates with low precision. As far as possible, more animals should be marked and recaptured to ensure higher precision. Planning for a known precision has been difficult owing to lack of simple expressions for errors. Where the total sample can be split into a number of interpenetrating subsamples*, separate estimates of population size can be formed, resulting in simpler expressions for overall error. This point needs examination for population estimates.

The CMR method assumes that both mortality and recruitment are negligible during the period of data collection*. Another assumption underlying the method is that marked and unmarked animals have the same probability of being caught in the second sample. When these assumptions are violated, it is useful to compare the results with other estimating procedures and, if possible, test on a known population. The latter is recommended for use in animal populations when the methods are likely to vary widely in accuracy.

Nonresponse and response errors often form a high proportion of the total error in estimation of animal numbers or their density. As an example of such errors, visibility bias of the observer in aerial surveys, which results in a proportion of the animal being overlooked, may be cited. In such cases, the population total or its density should be estimated by different groups of experienced observers, either through the use of interpenetrating subsamples or bias corrected by use of alternative methods, e.g., air—ground comparisons.

REFERENCES

1. Allen, O. B., Burton, J. H., and Holt, J. D. (1983). *J. Animal Sci.*, **55**, 765–770.
2. Amble, V. M. (1975). *Statistical Methods in Animal Sciences*. Indian Society of Agricultural Statistics, New Delhi, India.
3. Brown, J. E., Fitzhugh, H. A., and Cartwright, T. C. (1976). *J. Animal Sci.*, **42**, 810–818.
4. Burnham, K. P. and Anderson, D. R. (1976). *Biometrics*, **32**, 325–336.
5. Cole, J. W. L. and Grizzle, J. E. (1966). *Biometrics*, **22**, 810–828.
6. Cormack, R. M. (1979). In *Sampling Biological Populations*, R. M. Cormack, G. P. Patil, and D. S. Robson, eds. Satellite Program in Statistical Ecology, International Cooperative Publishing House, Fairland, MD, pp. 217–255.
7. Cormack, R. M. (1984). *Proc. XIIth International Biom. Conf.*, pp. 177–186.
8. Cox, D. R. (1958). *Planning of Experiments*. Wiley, New York.
9. De Lury, D. B. (1947). *Biometrics*, **3**, 145–167.
10. Eberhardt, L. L. (1978). *J. Wildl. Manag.*, **42**, 1–31.
11. Eberhardt, L. L. (1982). *J. Wildl. Manag.*, **46**, 734–740.
12. Erskine, A. J. (1973). *Canad. Wildl. Serv. Prog. Note*, **32**, 1–15.
13. Evans, C. D., Troyer, W. A., and Lensink, C. J. (1966). *J. Wildl. Manag.*, **30**, 767–776.
14. Fitzhugh, H. A. (1976). *J. Animal Sci.*, **42**, 1036–1051.
15. Gates, C. E. (1969). *Biometrics*, **25**, 317–328.
16. Gill, J. L. (1981). *J. Dairy Sci.*, **64**, 1494–1519.
17. Henderson, C. R. (1953). *Biometrics*, **9**, 226–252.
18. Henderson, C. R. (1969). In *Techniques and Procedures in Animal Science Research*. American Society for Animal Sciences, pp. 1–35.
19. Homeyer, P. G. (1954). *Statistics and Mathematics in Biology*. O. Kempthorne et al., eds. Iowa State College Press, Ames, IA, pp. 399–406.
20. Johnson, W. H. (1940). *J. Fish. Res. Board Canad.*, **4**, 349–354.

21. Jones, L. (1898). *Wils. Orn. Bull.*, **18**, 5–9.
22. Leslie, P. H. and Davis, D. H. S. (1939). *J. Animal Ecol.*, **8**, 94–113.
23. Lucas, H. L. (1948). *Proc. Auburn Conference on Applied Statistics*, p. 77.
24. Lucas, H. L. (1956). *J. Dairy Sci.*, **39**, 146–154.
25. Patterson, H. D. (1952). *Biometrika*, **39**, 32–48.
26. Patterson, H. D., and Lucas, H. L. (1962). Tech. Bull. No. 147, N. C. Agric. Exper. Stn.
27. Pearson K. (1894). *Philos. Trans. R. Soc. Lond., A*, **185**, 71–110.
28. Petersen, C. E. J. (1896). *Rep. Dan. Biol. Stat.*, **6**, 1–48.
29. Pollock, K. H. and Otto, M. C. (1983). *Biometrics*, **39**, 1035–1049.
30. Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). *Biometrics*, **40**, 329–340.
31. Robbins, C. S. and Vanvelzen, W. T. (1967). *Spec. Sci. Rep. Wildl. No. 102*, U.S. Fish Wildl. Serv.
32. Ricker, W. E. (1958). *Bull. Fish. Board Canad.*, **119**, 1–300.
33. Rutledge, J. J. and Gunsett, F. C. (1982). *J. Animal Sci.*, **54**, 1072–1078.
34. Seber, G. A. F. (1980). *The Estimation of Animal Abundance and Related Parameters*, 2nd ed. Griffin, London, England.
35. Seber, G. A. F. and Felton, R. (1981). *Biometrika*, **68**, 211–219.
36. Sen, A. R. (1970). *Biometrics*, **26**, 315–326.
37. Sen, A. R. (1972). *J. Wildl. Manag.*, **36**, 951–954.
38. Sen, A. R. (1973). *J. Wildl. Manag.*, **37**, 485–491.
39. Shirley, E. (1981). *Bias*, **8**, 82–90.
40. Siniff, D. B. and Skoog, R. O. (1964). *J. Wildl. Manag.*, **28**, 391–401.
41. Tukey, J. W. (1949). *Biometrics*, **5**, 232–242.
42. Watson, R. M., Parker, I. S. C., and Allan, T. (1969). *E. Afr. Wildl. J.*, **7**, 11–26.
43. Watson, R. M., Graham, A. D., and Parker, I. S. C. (1969). *E. Afr. Wildl. J.*, **7**, 43–59.
44. Weldon, W. F. R. (1890). *Proc. R. Soc. Lond.*, **47**, 445–453.
45. Weldon, W. F. R. (1893). *Proc. R. Soc. Lond.*, **54**, 318–329.

BIBLIOGRAPHY

Seber, G. A. F. (1986). *Biometrics*, **42**, 267–292. (Reviews animal abundance estimation techniques developed between 1979 and 1985, some of which supplant previous methods. This is a very thorough survey, with over 300 references.)

See also CAPTURE–RECAPTURE METHODS—I; ECOLOGICAL STATISTICS; LINE TRANSECT SAMPLING; QUADRAT SAMPLING; and WILDLIFE SAMPLING.

A. R. SEN

ANNALES DE L'INSTITUT HENRI POINCARÉ (B)

The *Annales de l'Institut Henri Poincaré, Section B, Probabilités et Statistiques* publishes articles in French and English, and is an international journal covering all aspects of modern probability theory and mathematical statistics, and their applications.

The journal is published by Elsevier; website links are www.elsevier.com/locate/anihpb and www.sciencedirect.com/science/journal/02460203 . The Editor-in-Chief works with an international Editorial Board of twelve or so members.

ANNALS OF APPLIED PROBABILITY

[This entry has been updated by the Editors.]

The *Annals of Applied Probability (AAP)* is the youngest of the "*Annals*" series to be created by the Institute of Mathematical Statistics* (IMS).

The website for *AAP* is www.imstat.org/aap/.

In 1973 the original *Annals of Mathematical Statistics* was split into the *Annals of Statistics** and the *Annals of Probability**. Over the next two decades both of these expanded noticeably in size: the 1973–1979 volumes of *Annals of Probability* contained around 1100 pages, but the 1993 volume more than doubled this.

By the late 1980s, the Council of the IMS had decided to further split its publications in probability theory and its applications into two journals—the continuing *Annals of Probability* and the *Annals of Applied Probability*. There were factors other than increasing size involved in this decision. Paramount amongst them was the feeling that the *Annals of*

Probability was failing to attract outstanding papers in applied probability with the ease that it was attracting such papers in the theory of probability. The IMS felt that this was partly because the *Annals of Probability* was seen as being aimed at an audience and an authorship interested in the purer aspects of probability theory, despite the efforts of successive editors to solicit papers with applications.

The first volume of the *Annals of Applied Probability* appeared in 1991, and *AAP* quickly established itself as one of the leading journals in the area, seamlessly carrying on the tradition of the other *Annals* published by the IMS.

EDITORIAL POLICY

The *Annals of Applied Probability* has two overriding criteria for accepting papers, other than formal correctness and coherence. These are

1. that the results in the paper should be genuinely applied or applicable; and
2. that the paper should make a serious contribution to the mathematical theory of probability, or in some other sense carry a substantial level of probabilistic innovation, or be likely to stimulate such innovation.

The first criterion in particular can be hard to define, and in some cases it has a broad interpretation. But in differentiating this journal from its sibling journal, the *Annals of Probability*, it is a criterion that is applied with some care. The Editorial Board has rejected a number of excellent papers because the authors did not make a convincing case for applicability, and indeed several of these have been forwarded (with the author's permission) to the other journal, and some have been accepted there.

The second, more mathematical criterion is also taken seriously, and in this the *Annals of Applied Probability* follows the tradition of the original *Annals of Mathematical Statistics*, which was set up to provide an outlet for the more mathematical papers written by members of the American Statistical Association. Thus the *Annals of Applied Probability* has rejected a number of excellent papers in, say, queueing theory, or in the theory of biological models, where the applications were indeed of interest but where the contribution to the mathematical or probabilistic theory was felt to be limited.

The editorial policy of *AAP* is stated on the website exhibited above. In addition to the two criteria just discussed, it states:

> "*The Annals of Applied Probability* aims to publish research of the highest quality reflecting the varied facets of contemporary Applied Probability. Primary emphasis is placed on importance and originality ... .
> "Mathematical depth and difficulty will not be the sole criteria with respect to which submissions are evaluated. Fundamentally, we seek a broad spectrum of papers which enrich our profession intellectually, and which illustrate the role of probabilistic thinking in the solution of applied problems (where the term "applied" is often interpreted in a side sense)."
> "Most papers should contain an Introduction which presents a discussion of the context and importance of the issues they address and a clear, non-technical description of the main results. The Introduction should be accessible to a wide range of readers. Thus, for example, it may be appropriate in some papers to present special cases or examples prior to general, abstract formulations. In other papers a discussion of the general scientific context of a problem might be a helpful prelude to the main body of the paper. In all cases, motivation and exposition should be clear".

All papers are refereed.

TYPES OF PAPERS CARRIED

Stating such general criteria for publication is one thing; seeing them in operation in practice is another. The style of a journal is always best described in terms of its actual content, and although this is by no means stationary in time, the following areas are ones in which the *Annals of Applied Probability* carried very significant contributions in its first years:

1. Queueing networks, with the Lanchester Prize-winning paper on loss networks by Frank Kelly, papers on unexpectedly unstable networks, and much

of the seminal literature on fluid model approximations;

2. Stochastic models in finance*, with applications to investment strategies, options analysis, arbitrage models, and more;
3. Rates of convergence papers for Markov chains*, with contributions ranging from finite to general space chains;
4. Foundational work on Markov-chain Monte Carlo* models (*see* also Gibbs Sampling);
5. Development of stochastic algorithms in computing, and analysis of their properties.

This list is by no means exhaustive; the behavior of diffusion processes*, branching processes*, and the like is of course widely studied in applied probability, and excellent papers in these areas have appeared in the *Annals of Applied Probability*. However, to be encouraged by the Editors, there needs to be more than formal technical novelty in such papers, and authors are encouraged to consider this carefully when submitting to the journal.

STRUCTURE AND ACCEPTANCE PROCESS

Currently, the *Annals of Applied Probability* has an Editorial Board, consisting of the Editor (appointed by the IMS Council for a three-year term), a Managing Editor (with shared responsibilities for the *Annals of Probability*) and 25 or so Associate Editors from around the world. Past Editors of the *Annals of Probability* are:

J. Michael Steele, 1991–1993,
Richard L. Tweedie, 1994–1996,
Richard T. Durrett, 1997–1999,
Søren Asmussen, 2000–2002,
Robert Adler, 2003–.

FUTURE PLANNING

As with many journals, the *Annals of Applied Probability* is moving towards an era of improved publication speed using electronic means.

See also *Annals of Probability*; *Annals of Statistics*; and Institute of Mathematical Statistics.

Paul S. Dwyer

ANNALS OF EUGENICS. See *Annals of Human Genetics*

ANNALS OF HUMAN GENETICS

The *Annals of Human Genetics* was one of the earliest and has remained one of the foremost journals concerned with research into genetics*; it is specifically concerned with human genetics*. The home page for the journal is www.gene.ucl.ac.uk/anhumgen.

The original title was *Annals of Eugenics*, subtitled “A Journal for the scientific study of racial problems.” The word “Eugenics” had been coined by Sir Francis Galton*, who defined it in 1904 as “the science which deals with all the influences that improve the inborn qualities of a race; also those that develop them to the utmost advantage.” In his foreword to the first volume of the *Eugenics Review* in 1909 he said that “the foundation of Eugenics is laid by applying mathematical statistical treatment to a large collection of facts.” He was a man of wide scientific interests, which included stockbreeding, psychology, and the use of fingerprints for identification, and he was a cousin of Charles Darwin. He was also one of the early white explorers of Africa and a prolific writer on these and many other subjects.

The journal was founded in 1925 by Karl Pearson*, who was head of the Department of Applied Statistics and of the Galton Laboratory at University College; it was printed by Cambridge University Press. The Galton Laboratory had been founded under Galton’s will, and Karl Pearson was the first occupant of the Galton Chair of Eugenics. He had been an intimate friend and disciple of Galton and remained faithful to many of his ideas. The journal’s aims were set out in a foreword to the first volume, where eugenics is defined as “the study of agencies under social control that may improve or impair the racial qualities of future generations either physically or mentally.” The quotations from Galton and Darwin still retained on the cover of the

Annals indicate its commitment to mathematical and statistical techniques. Until the arrival of the computer and the resulting enormous increase in the amount of information collected, the journal emphasized the necessity of publishing data with papers; this is now most often deposited in record offices when it is extensive.

Karl Pearson retired in 1933 and was succeeded as editor in 1934 by Ronald A. Fisher*, who was also keenly interested in the development of statistical techniques, but was critical of some of Pearson's statistical methods. He became Galton Professor in 1934, and the editorship of the *Annals* went with the Chair then as now.

Fisher innovated many well-known statistical techniques and showed how they could be applied to genetical problems. He changed the subtitle of the *Annals* to "A Journal devoted to the genetic study of human populations." He co-opted several members of the Eugenics Society on to the editorial board. This society had been founded independently in 1908 as the Eugenics Education Society; its purpose was to propagate Galton's ideas and the work of the Laboratory, and Galton had accepted the presidency. Ronald Fisher had been an active member from its early days. This partnership seems only to have lasted until the outbreak of war, when Fisher returned to Rothamsted Experimental Station. He stayed there until 1943, when he accepted the Chair of Genetics at Cambridge and his editorship ended. In the foreword to the first volume of the *Annals* which he edited, Vol. 6, he announces its policy to be consistent with the aims of its founder:

> The contents of the journal will continue to be representative of the researches of the Laboratory and of kindred work, contributing to the further study and elucidation of the genetic situation in man, which is attracting increasing attention from students elsewhere. The two primary disciplines which contribute to this study are genetics and mathematical studies.

In 1945, Lionel S. Penrose succeeded him as editor of the *Annals*. He was a distinguished medical man and alienist, and under him the journal became more medical in content. Some of his papers on Down's anomaly, a permanent interest of his, and other aspects of inherited mental illness appeared in the journal. A feature of it in his time was the printing of pedigrees of inherited diseases covering several generations. He was responsible for changing the title from *Annals of Eugenics* to *Annals of Human Genetics*, a change for which it was necessary for an act of Parliament to be passed. The subtitle was also changed again to "A Journal of Human Genetics." He retired in 1965, and so did M. N. Karn, who had been the assistant editor since prewar days. Penrose was editor for a longer period than either of his predecessors and under his guidance the journal broadened its coverage and drew its contributions from a wider field.

Harry Harris, also a medical man and biochemist, succeeded to the Galton Chair and the editorship in 1965, and coopted C. A. B. Smith, mathematician and biometrician, who had been on the editorial board since 1955, to be coeditor. Professor Harris was also head of a Medical Research Council Unit of Biochemical Genetics which became associated with the Galton Laboratory. Reflecting the editors' interests, the contents of the *Annals* inevitably became more concerned with biochemical genetics and statistics; *Annals of Eugenics* was dropped from the title page.

In 1975, Harris accepted the Chair of Human Genetics at the University of Pennsylvania, and for the next two years Cedric Smith was virtually the sole editor, as the Galton Chair remained vacant. In 1978, Elizabeth B. Robson was appointed to it, and she and C. A. B. Smith became the editors. The journal has always been edited at the Galton Laboratory.

The journal is a specialized one dealing with human genetics, but has changed with the progress of research and changing methods, currently publishing material related directly to human genetics or to scientific aspects of human inheritance.

As the website for the journal points out, during the latter part of the twentieth century it became clear that our understanding of variation in the human genome could be enhanced by studying the interaction between the fields of population genetics and molecular pathology. Accordingly, contemporary topics included in the *Annals of Human*

Genetics include human genome variation, human population genetics, statistical genetics, the genetics of multifactorial diseases, Mendelian disorders and their molecular pathology, and pharmacogenetics. Animal as well as human models may be considered in some of these areas.

Most articles appearing in the journal report full-length research studies, but many issues include at least one review paper. Shorter communications may also be published.

For many years the journal was published by the Galton Laboratory under the ownership of University College London. Since 2003 it is published by Blackwell. It has an Editor-in-Chief, a Managing Editor, five or so Senior Editors, a Reviews Editor and an international Editorial Board of 30 or so members. All papers are rigorously referred.

See also English Biometric School; Fisher, Ronald Aylmer; Galton, Francis; Human Genetics, Statistics in—II; and Pearson, Karl—I.

Jean Edmiston
The Editors

ANNALS OF MATHEMATICAL STATISTICS. See *Annals of Statistics*

ANNALS OF PROBABILITY

[This entry has been updated by the Editors.]

The *Annals of Probability (AP)* was one of two journals that evolved from the *Annals of Mathematical Statistics (AMS)* in 1973, the other being the *Annals of Statistics* (AS)*. All three journals are (or were) official publications of the Institute of Mathematical Statistics* (IMS).

Readers are referred to the entries *Annals of Statistics* and Institute of Mathematical Statistics for the evolution both of *AP* and *AS* out of *AMS*.

The *Annals of Probability* is published bimonthly, each volume consisting in the six issues of each calendar year. Due to the expansion of *AP* in the 1970s and 1980s, the IMS Council decided to split *AP* into two journals; *Annals of Probability* continued publication, but with a theoretical focus, and *Annals of Applied Probability* (AAP)* with a focus on applications began publication in 1991. See the entry on *AAP* for further discussion on the rationale for the split.

Table 1. Editors, The Annals of Probability

Ronald Pyke,	1972–1975
Patrick Billingsley,	1976–1978
R. M. Dudley,	1979–1981
Harry Kesten,	1982–1984
Thomas M. Liggett,	1985–1987
Peter Ney,	1988–1990
Burgess Davis,	1991–1993
Jim Pitman,	1994–1996
S. R. S. Varadhan,	1996–1999
Thomas K. Kurtz,	2000–2002
Steve Lalley,	2003–2005

The editorship of AP has been held roughly for three-year periods; see Table 1. In the first issue (Feb. 1973) appeared the policy statement:

> "The main aim of the *Annals of Probability* and the *Annals of Statistics* is to publish original contributions related to the theory of statistics and probability. The emphasis is on quality, importance and interest; formal novelty and mathematical correctness alone are not sufficient. Particularly appropriate for the *Annals* are important theoretical papers and applied contributions which either stimulate theoretical interest in important new problems or make substantial progress in existing applied areas. Of special interest are authoritative expository or survey articles, whether on theoretical areas of vigorous recent development, or on specific applications. All papers are referred.

The current editorial policy of AP appears on the journal website www.imstat.org/aop/:

"The *Annals of Probability* publishes research papers in modern probability theory, its relation to other areas of mathematics, and its applications in the physical and biological sciences. Emphasis is on importance, interest, and originality—formal novelty and correctness are not sufficient for publication. The *Annals* will also publish authoritative review papers and surveys of areas in vigorous development."

Currently the Editorial Board is comprised of the Editor, a Managing Editor, and 25 or so Associate Editors from around the world.

See also ANNALS OF APPLIED PROBABILITY; ANNALS OF STATISTICS; and INSTITUTE OF MATHEMATICAL STATISTICS.

ANNALS OF STATISTICS

[This entry has been updated by the Editors.]

The *Annals of Statistics (AS)*, first published in 1973, is one of the two journals resulting from a split of the old *Annals of Mathematical Statistics*, the other journal being the *Annals of Probability**. Three current *Annals* are official publications of the Institute of Mathematical Statistics* (IMS); in the late 1980s the IMS Council decided to split its publications on probability theory and its applications further; *see* ANNALS OF APPLIED PROBABILITY.

The website for the *Annals of Statistics* is www.imstat.org/aos/.

The original *Annals of Mathematical Statistics (AMS)* started before the Institute existed, and in fact was originally published by the American Statistical Association* in 1930. At that time it had become apparent that the *Journal of the American Statistical Association* (JASA)* could not adequately represent the interests of mathematically inclined statisticians, who were beginning to do important research. Willford King, writing in a prefatory statement to the first issue of *AMS*, said:

> The mathematicians are, of course, interested in articles of a type which are not intelligible to the non-mathematical readers of our Journal. The Editor of our Journal [JASA] has, then, found it a puzzling problem to satisfy both classes of readers.
>
> Now a happy solution has appeared. The Association at this time has the pleasure of presenting to its mathematically inclined members the first issue of the *ANNALS OF STATISTICS*, edited by Prof. Harry C. Carver of the University of Michigan. This Journal will deal not only with the mathematical technique of statistics, but also with the applications of such technique to the fields of astronomy, physics, psychology, biology, medicine, education, business, and economics. At present, mathematical articles along these lines are scattered through a great variety of publications. It is hoped that in the future they will be gathered together in the *Annals*.

The seven articles that followed in that first issue covered a very wide range indeed, as their titles suggest:

Remarks on Regression

Synopsis of Elementary Mathematical Statistics

Bayes Theorem

A Mathematical Theory of Seasonal Indices

Stieltjes Integrals in Mathematical Statistics

Simultaneous Treatment of Discrete and Continuous Probability by Use of Stieltjes Integrals

Fundamentals of Sampling Theory

Harry Carver, the founding editor, took on sole responsibility for the young journal when in 1934 the *ASA* stopped its financial support. He continued to publish privately until 1938 when the IMS took over the financial responsibility. Actually, the IMS had come into existence in 1935, and the *Annals* had been its official publication from the start. But Carver, a prime mover in starting the IMS, insisted that it not be tied down by support of the journal.

After the crucial period of Carver's editorship, S. S. Wilks was appointed editor, a post he held from 1938 until 1949. Since that time the editorship has been held for 3-year periods, initially by Wilks's appointed successors (Table 1).

Each editor has headed a distinguished editorial board, but none compares to the illustrious board established in 1938: Wilks, Craig, Neyman (co-editors), Carver, Cramér, Deming, Darmois, R. A. Fisher*, Fry, Hotelling*, von Mises, Pearson, Rietz, and Shewhart.

With this auspicious start, the *AMS* became the focal point for developments in theoretical statistics, particularly those developments associated with the general mathematical theories of estimation, testing, distribution theory, and design of experiments. The full impact of the *Annals*, past and present, is clearly seen in the bibliographies of most books on theoretical statistics. Many of the newer statistical methods can be traced back to pioneering research papers in *AMS*.

Table 1.

Editors of *Annals of Mathematical Statistics*	
H. C. Carver (1930–1938)	William H. Kruskal (1958–1961)
S. S. Wilks (1938–1949)	J. L. Hodges, Jr. (1961–1964)
T. W. Anderson (1950–1952)	D. L. Burkholder (1964–1967)
E. L. Lehmann (1953–1955)	Z. W. Birnbaum (1967–1970)
T. E. Harris (1955–1958)	Ingram Olkin (1970–1972)

Editors of *Annals of Statistics*	
Ingram Olkin (1972–1973)	Lawrence D. Brown and
I. R. Savage (1974–1976)	John A. Rice, 1995–1997
R. G. Miller, Jr. (1977–1979)	James O. Berger and
David Hinkley (1980–1982)	Hans R. Kunsch, 1998–2000
Michael D. Perlman, 1983–1985	John I. Marden and
Willem R. Van Zwet, 1986–1988	Jon A. Wellner, 2001–2003
Arthur Cohen, 1989–1991	Morris L. Eaton and
Michael Woodroofe, 1992–1994	Jianqing Fan, 2004–2006

After some 40 years of growing strength and size, the *AMS* was split in 1972 during the editorship of I. Olkin, who continued as first editor of *AS*. Each volume of *AS* is devoted entirely to research articles. (The news items and notices that used to appear in *AMS* until 1972 are published in the *IMS Bulletin*, issued bimonthly.) The journal does not publish book reviews. A volume currently consists of six bimonthly issues. All papers are referred under the general guidelines of editorial policy.

EDITORIAL POLICY

The following statement was made in the 1938 volume of *JASA*:

> The *Annals* will continue to be devoted largely to original research papers dealing with topics in the mathematical theory of statistics, together with such examples as may be useful in illustrating or experimentally verifying the theory. However, in view of the purpose of the Institute of Mathematical Statistics which, interpreted broadly, is to stimulate research in the mathematical theory of statistics and to promote cooperation between the field of pure research and fields of application, plans are being made to extend the scope of the *Annals* to include expository articles from time to time on various fundamental notions, principles, and techniques in statistics. Recognizing that many theoretical statistical problems have their origin in various fields of pure and applied science and technology, papers and shorter notes dealing with theoretical aspects of statistical problems arising in such fields will be welcomed by the editors.

The current editorial policy of the journal is stated on its website www.imstat.org/aos/, as follows:

"*The Annals of Statistics* aims to publish research papers of highest quality, reflecting the many facets of contemporary statistics. Primary emphasis is placed on importance and originality, not on formalism.

The discipline of statistics has deep roots in both mathematics and in substantive scientific fields. Mathematics provides the language in which models and the properties of statistical methods are formulated. It is essential for rigor, coherence, clarity and understanding. Consequently, our policy is to continue to play a special role in presenting research at the forefront of mathematical statistics, especially theoretical advances that are likely to have a significant impact on statistical methodology or understanding. Substantive fields are essential for continue vitality of statistics, since they provide the motivation and direction for most of the future developments in statistics. We thus intend to also publish papers relating to the role of statistics in inter-disciplinary investigations in all fields of natural, technical and social sciences. A third force that is reshaping statistics is the computational revolution, and the *Annals* will also welcome developments in this area. Submissions in these two latter categories will be evaluated primarily by the relevance of the issues addressed and the creativity of the proposed solutions.

"Lucidity and conciseness of presentation are important elements in the evaluation of submissions. The introduction of each paper should be accessible to a wide range of readers. It should thus discuss the context and importance of the issues addressed and give a clear, nontechnical description of the main results. In some papers it may, for example, be appropriate to present special cases or specific examples prior to general, abstract formulations, while in other papers discussion of the general scientific context of a problem might be a helpful prelude to the body of the paper."

Currently two Editors, a Managing Editor, and 40 or so Associate Editors in many countries serve on the Editorial Board.

AS continues to recognize its singular and historic role as publisher of general theory, while reflecting the impact that theory does and should have on practical problems of current interest. For that reason, relevance and novelty are at least as important as mathematical correctness.

See also *Annals of Applied Probability*; *Annals of Probability*; and Institute of Mathematical Statistics.

D. V. Hinkley

ANNALS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

The first issue of this journal appeared in 1949. It is published in English by Kluwer.

The aims and scope of *AISM* are presented at the journal's website www.kluweronline.com/issn/0020-3157, as follows:

> "*Annals of the Institute of Statistical Mathematics* (AISM) provides an international forum for open communication among statisticians and researchers working with the common purpose of advancing human knowledge through the development of the science and technology of statistics.
>
> AISM will publish broadest possible coverage of statistical papers of the highest quality. The emphasis will be placed on the publication of papers related to: (a) the establishment of new areas of application; (b) the development of new procedures and algorithms; (c) the development of unifying theories; (d) the analysis and improvement of existing procedures and theories; and the communication of empirical findings supported by real data.
>
> "In addition to papers by professional statisticians, contributions are also published by authors working in various fields of application. Authors discussing applications are encouraged to contribute a complete set of data used in their papers to the AISM Data Library. The Institute of Statistical Mathematics will distribute it upon request from readers (see p. 405 and 606, Vol. 43, No. 3, 1991). The final objective of AISM is to contribute to the advancement of statistics as the science of human handling of information to cope with uncertainties. Special emphasis will thus be placed on the publication of papers that will eventually lead to significant improvements in the practice of statistics."

AISM currently has an Editor-in-Chief, six Editors and 40 Associate Editors. All papers published in the journal are refereed.

ANOCOVA TABLE. See Analysis of Covariance; Anova Table

ANOVA TABLE

An ANOVA (analysis of variance) table is a conventional way of presenting the results of an analysis of variance*. There are usually four columns, headed

1. Source (of variation)
2. Degrees of freedom*
3. Sum of squares
4. Mean square*

Columns 1 and 2 reflect the size and pattern of the data being analyzed and the model being used. Column 4 is obtained by dividing the entry (in the same row) in column 3 by that in column 2.

Sometimes there is a fifth column, giving the ratios of mean squares to a residual mean square* (or mean squares). These statistics are used in applying the standard F-test* used in the analysis of variance.

The value of the ANOVA table is not only in its convenient and tidy presentation of the quantities used in applying analysis-of-variance tests. The juxtaposition of all the quantities used for a number of different tests (or the mean squares for many different sources of variation) can provide valuable insight into the overall structure of variation. For example, in the analysis of a factorial experiment*, the groups of interactions* of specified order can provide evidence of relatively great variation arising when a particular factor (or group of factors) is involved.

The term “ANOCOVA table” is also used (although rather infrequently) to describe similar tables relevant to the analysis of covariance*.

See also ANALYSIS OF VARIANCE.

ANSARI—BRADLEY *W*-STATISTICS.

See SCALE TESTS, ANSARI—BRADLEY

ANSCOMBE DATA SETS

A celebrated classical example of role of residual analysis and statistical graphics in statistical modeling was created by Anscombe [1]. He constructed four different data sets $(X_i, Y_i), i = 1, \ldots, 11$ that share the same descriptive statistics $(\overline{X}\,\overline{Y}, \hat{\beta}_0, \hat{\beta}_1, MSE, R^2, F)$ necessary to establish linear regression fit $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

The following statistics are common for the four data sets:

Sample size N	11
Mean of $X(\overline{X})$	9
Mean of $Y(\overline{Y})$	7.5
Intercept $(\hat{\beta}_0)$	3
Slope $(\hat{\beta}_1)$	0.5
Estimator of σ, (s)	1.2366
Correlation $r_{X,Y}$	0.816

A linear model is appropriate for Data Set 1; the scatterplots and residual analysis suggest that the Data Sets 2–4 are not amenable to linear modeling.

						Set 1					
X	10	8	13	9	11	14	6	4	12	7	5
Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
						Set 2					
X	10	8	13	9	11	14	6	4	12	7	5
Y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
						Set 3					
X	10	8	13	9	11	14	6	4	12	7	5
Y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
						Set 4					
X	8	8	8	8	8	8	8	19	8	8	8
Y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

REFERENCE

1. Anscombe, F. (1973). Graphs in Statistical Analysis, *American Statistician*, **27** [February 1973], 17–21.

FURTHER READING

Tufte, E. R., *The Visual Display of Quantitative Information*, Graphic Press, 1983.

ANSCOMBE, FRANCIS JOHN

Frank Anscombe was born in Hertfordshire and grew up in Hove, England. His parents were Francis Champion Anscombe (1870–1942) and Honoria Constance Fallowfield Anscombe (1888–1974). His father worked for a pharmaceutical company in London and his mother was an early female graduate of the University of Manchester. Frank attended Trinity College, Cambridge, on a merit scholarship. He obtained a B.A. in

Mathematics in 1939 with first class honors, and an M.A. in 1943. During the war years, he was with the British Ministry of Supply, concerned with assessment of weapons and quality control of munitions production. From 1945 to 1948, he was in the statistics department at Rothamsted Experimental Station, Hertfordshire. In 1954, he married John Tukey's sister-in-law, Phyllis Elaine Rapp. They had four children (Francis, Anthony, Frederick, and Elizabeth). Anscombe died in 2001 after a long illness.

ACADEMIC CAREER

From 1948 to 1956, Frank Anscombe was lecturer in mathematics at the Statistical Laboratory, the University of Cambridge in England. In 1956, he moved to the mathematics department of Princeton University, as associate and then full professor. He left Princeton in 1963 to found the statistics department at Yale University. He chaired that department for six years, developing a graduate program. The department became known for its careful balance of theory and applications. He was a member of important advisory and evaluation committees. He retired in 1988.

STATISTICAL CONTRIBUTIONS

Anscombe made important contributions to the British World War II effort. In particular, he was concerned with the deployment of weapons, the aiming of anti-aircraft rockets, and the development of strategies for massing guns. After the war, he worked at Rothamsted on the applications of statistics to agriculture. During these years, he published papers in *Nature, Biometrika*, and *Biometrics,* and had a discussion paper in the *Journal of the Royal Statistical Society* [1–3,12]. He often wrote on the problems of sampling inspection and sequential estimation. One in particular [5] is a discussion paper on sequential analysis invited by the Royal Statistical Society. Later, however [7], he wrote, “Sequential analysis is a hoax.”

The next decade saw Anscombe evolving into a subjective Bayesian on the one hand, but on the other delving directly into data analysis concerned with uses of residuals* and into tricky problems created by outliers*. The Fourth Berkeley Symposium paper [6] and the *Technometrics* paper with John Tukey [13] were landmarks in the history of residual analysis.

The final stages of Anscombe's research career saw a move into computing [8,9]. In the Preface of the book [9], he writes that the work is “a festivity in ... honor” of J. W. Tukey and K. E. Iverson. Anscombe's last published paper [10] concerned testing in clinical trials*.

Two of Anscombe's theoretical papers are often referred to. Reference 4 presents Anscombe's Theorem, which provides conditions for replacing a fixed sample size by a random stopping time. The result was later extended by Rényi [14]. Reference 11 contains the Anscombe–Auman work showing that Savage's derivation of expected utility* can be considerably simplified.

CONCLUDING REMARKS

Anscombe was an elected member of the International Statistical Institute* and a charter member of the Connecticut Academy of Science and Engineering. He was the R. A. Fisher Lecturer in 1982. Throughout his career, he prepared pithy book reviews and contributed to discussions in a lively manner.

Anscombe was concerned with the popularization and simplification of statistics. For example, he had papers in *The American Statistician* and once wore a sombrero at an Institute of Mathematical Statistics meeting in an attempt to enliven it. He had important interests outside of statistics, including classical music, poetry, art, and hiking.

REFERENCES

1. Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246–254.
2. Anscombe, F. J. (1948). The validity of comparative experiments (with discussion). *J. R. Stat. Soc. A*, **111**, 181–211.
3. Anscombe, F. J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, **5**, 165–173.
4. Anscombe, F. J. (1952). Large-sample theory of sequential estimation. *Proc. Cambridge Philos. Soc.*, **48**, 600–607.

5. Anscombe, F. J. (1953). Sequential estimation (with discussion). *J. R. Stat. Soc. B*, **15**, 1–29.
6. Anscombe, F. J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, **1**, 1–36.
7. Anscombe, F. J. (1963). Sequential medical trials. *J. Am. Stat. Assoc.*, **58**, 365–383.
8. Anscombe, F. J. (1968). "Regression Analysis in the Computer Age". *Proc. Thirteenth Conf. Design Expts. Army Res. Dev. Testing*. U.S. Army Research Office, Durham, NC, pp. 1–13.
9. Anscombe, J. J. (1981). *Computing in Statistical Science Through APL*. Springer-Verlag, New York.
10. Anscombe, F. J. (1990). The summarizing of clinical experiments by significance levels. *Stat. Med.*, **9**, 703–708.
11. Anscombe, F. J. and Auman, J. (1963). A definition of subjective probability. *Ann. Math. Stat.*, **34**, 199–205.
12. Anscombe, F. J. and Singh, B. N. (1948). Limitation of bacteria by micro-predators in soil. *Nature*, **161**, 140–141.
13. Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141–160.
14. Rényi, A. (1960). On the central limit theorem for the sum of a random number of independent random variables. *Acta Math. Acad. Sci. Hung.*, **11**, 97–102.

See also Bayesian Inference; Outliers; Residuals; and Sequential Estimation.

David R. Brillinger

ANTHROPOLOGY, STATISTICS IN

Statistical methods were introduced into physical anthropology by Adolphe Quetelet* and Francis Galton* during the nineteenth century. From their work grew the "Biometrics School," headed by Pearson* and Karl Weldon W.F.R.*, whose members studied the variation between local races in an attempt to clarify the processes of inheritance and evolution. Human skulls, in particular, were intensively studied because of the availability of historical material. The statistical treatment of data from skulls raised many new problems, and Pearson [23] took one of the first steps in multivariate analysis* by introducing the coefficient of racial likeness*, a statistic based on all measurements and used to assess the significance of differences between groups. There is less emphasis on craniometry today than there was in Pearson's time, but the basic statistical problem is still with us. How should variation in the shape of complex objects such as bones be described?

Pearson believed passionately in measurement* as the basis of all science, although he recognized that most advances in the study of shape had actually relied on visual comparisons. Unfortunately, the measurement of shape is very difficult. In the majority of published examples the procedure has been to identify common landmarks on the objects and then to measure angles and linear distances. The hope is that a statistical summary of these data will embody a summary of shape. The early biometricians were restricted to simple statistical summaries by primitive computing equipment, but techniques of multivariate analysis are now commonly used to attempt a more explicit description of shape variation within a group and to make comparisons between groups.

Landmarks are hard to find on some objects, and an alternative approach is to record the coordinates of a large number of points on the object. Measures of shape must then be based on geometric properties of the surfaces containing the points, and should be independent of which points are chosen. The technique has so far been limited to outlines from sections of the original objects for which curves rather than surfaces are relevant. This has been for practical rather than theoretical reasons. A plot of radial distance versus angle has been used to summarize outlines, but this has the disadvantage of depending on the choice of origin. An alternative is a plot of tangent direction versus distance round the outline [27]. In both cases the plots may be described quantitatively using Fourier series*. The geometric approach is well reviewed in Bookstein [5].

SHAPE AND SIZE

The special problem of determining the extent to which shape is related to size is referred to as allometry*. Apart from this concern, size variation is usually of little interest in studies of shape. However, linear distances inevitably reflect the size of an object so that size variation can be a nuisance. Since only the relative magnitude of distances is important for shape, the distances are often replaced by ratios of one distance to another. If shape is related to size, such ratios still might well be related to size, but the degree of relationship will be much less than for the original distances. It is usual to choose one distance that is strongly influenced by size and to use this as the denominator when expressing other distances as ratios. Mosimann [18] and Corruccini [7] have suggested the use of a symmetric function of all the size-dependent variables as the denominator.

VARIATION WITHIN A GROUP

Measurements made on an object are regarded as a vector of observations x on a vector of variables X. The individual variables in X are referred to as $X_1, \ldots, X_\upsilon$. Data from a group of n objects consists of n vectors, $x_1, \ldots, x_n$, which together form a $n \times \upsilon$ data matrix. The rows of this matrix, which are the vectors x_i, may be represented as n points in υ-space and the columns as υ points in n-space. The two representations are sometimes referred to as Q and R, respectively. The Euclidean metric is used in both spaces so that in row space (Q) the distance between two objects is, in matrix notation, $(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})$. Thus two objects that are close in row space are similar in respect of the υ measurements.

The usual statistical summary is based on the mean* and standard deviation* of each variable, together with correlations* between pairs of variables. This depends on the distribution being roughly multivariate normal*, an assumption that may be partially tested by inspecting the distribution of each variable separately (which should be normal) and each possible bivariate plot (which should be linear). If there is no correlation, then the variation in shape is uninteresting: objects vary, but not in any consistent way. Suppose now that all objects are roughly the same size. Then a high correlation (positive or negative) is regarded as evidence that the two variables are constrained by the necessity for the object to stay in the same class of shapes and that jointly they are measuring a single aspect of shape. A negative correlation can be converted to a positive one by using the reciprocal of a measurement or ratio, or the complement of an angle, and this is usually done to ensure positive correlations as far as possible. If a group of variables has high positive intercorrelations, then the group is taken to be measuring a single aspect of shape. Different aspects of shape will have relatively low correlation, by definition. Statistically, this amounts to grouping the variables on the basis of their correlations. It may be done by eye for a small number of variables or by extracting principal components for a larger number ($\upsilon > 10$). If the objects do not have the same size, then the interpretation of correlations depends on the nature of the variables. Correlation among linear distances will almost certainly be partly, perhaps largely, due to size variation. Correlation among angles and ratios will generally indicate constraints of shape.

Principal component analysis* extracts the components $(Z_1, \ldots, Z_\upsilon)$ from the covariances* between the original variables. Each component is a linear combination of the variables $(X_1, \ldots, X_\upsilon)$. If $Z_1 = a_1X_1 + \cdots + a_\upsilon X_\upsilon$, then a_i is called the loading* of X_i on the first component and is proportional to the covariance between Z_1 and X_i. When the data have been standardized by reducing each variable by its mean and scaling each to have unit standard deviation, then covariance equals correlation. In this case the loadings are used to group variables according to their correlations with the first few components and hence with each other. Sometimes the procedure is reversed and the first few components are "named" according to the variables they are associated with. This is logically equivalent to grouping the variables*. Howells [14] gives a good example.

If the observed values of X for an object are substituted in the expression for Z_1, the result is a *score* for that object on Z_1. The scores on Z_1 and Z_2 may be plotted, using rectangular axes, to give an approximation to the representation of objects in row space. The quality of the approximation depends on how much of the overall variability between objects has been reproduced by Z_1 and Z_2. The plot is useful for spotting any lack of homogeneity* in the group of objects. (Rao [25] gives a very detailed account of the different uses of principal components.)

VARIATION BETWEEN GROUPS

If the comparison of several groups is to be meaningful, each must have a representative shape. In other words, the groups must be homogeneous, displaying some variation, but not too much. We shall assume that there are k groups in all, distinguishing between them by using different letters $(x, y, z, \ldots)$ to refer to a typical vector of measurements in each group. The vector of means for each group is regarded as representing a mean shape. For example, McLearn et al. [16] reconstruct a typical profile from the mean of measurements taken from a large number of individual profiles of the human face. In this example the result still looked like a human face, i.e., $\overline{x}$ satisfied the same geometric constraints as each x_i in $\overline{x} = \sum x_i/n$, but since these constraints are in general nonlinear, this will not always be the case.

When comparing groups there are two kinds of questions: comparison between pairs of groups and an overall comparison. The latter requires some metric* enabling one to decide whether group x is closer to y than to z, and a wide variety have been proposed. A good review is given in Weiner [26]. The most commonly used metric is now $(\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{y})$, where $\mathbf{\Sigma}$ is a positive definite $p \times p$ matrix. This may be interpreted as follows. If a multivariate normal density with covariance $\mathbf{\Sigma}$ is centered at x, then all points y that are equiprobable in this density are equidistant from x in this metric. If $\mathbf{\Sigma} = \mathbf{I}$, the contours of equal probability are spheres; otherwise, they are ellipsoids. The metric is satisfactory only for groups with similar patterns of covariance, in which case $\mathbf{\Sigma}$ is taken equal to $\mathbf{S}$, the pooled covariance matrix within groups. This leads to D^2, equal to $(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T \mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}})$, as the measure of distance between groups. It is clear from the derivation that the comparison of two large values of D^2 is unlikely to be satisfactory. In fact, although D^2 takes account of correlations within groups, it is, like all metrics, rather a blunt instrument when used for assessing affinity.

With a lot of groups the pairs of D^2 values can be confusing, so a visual overall picture of the interrelationships between groups is produced using principal components*. A $k \times \upsilon$ data matrix in which the rows are now the group means is used and principal component analysis is carried out in the D^2 metric [12]. The resulting components are called canonical variates or sometimes discriminant functions*. Scores for each group on the first two or three canonical variates are plotted using rectangular axes. Good examples of the use of canonical variates are those of Ashton, et al. [2], Day and Wood [9], and Oxnard [19].

If something is known about the function of the objects, then it may be possible to order the groups according to this function, at least roughly, e.g., low, medium, and high. We assume that function is not quantifiable, so that only a rough ordering is possible. The plot will show whether or not this ordering is associated with the major dimensions of shape variation. Of course, there may well be other measures of shape more highly associated with function than the first few canonical variates. Aspects of shape that are highly variable are not necessarily those most highly associated with function.

The D^2-metric itself is not without disadvantages. First the assumption that the pattern of variation is constant for all groups is inherently unlikely with shape studies. Second, when a large number of variables is used to ensure that shape is adequately described, they often contain redundancies which lead to nearly singular matrices $\mathbf{S}$ and hence to very large and meaningless D^2-values. Some progress has been made in overcoming these and other difficulties. To avoid distortion from studying scores on just the first two or three canonical variates, Andrews [1] has suggested representing each group by a weighted combination of trigonometric functions of θ with weights equal to the scores on all υ canonical variates. As θ varies, each group is represented by a smooth curve. Burnaby [6] has shown how the D^2-metric may be adjusted to measure change only in certain directions in row space. This can be useful when it is required to avoid a direction corresponding to growth or size change. Penrose [24] showed how the adjusted D^2-metric becomes equal to the Euclidean metric when all correlations are equal. To avoid D^2 altogether, some authors have pooled the groups and studied the extent to which the individual objects can be clustered, either visually using principal components and the Euclidean metric, or automatically using various clustering* algorithms [20]. The techniques of multidimensional scaling* [15] and principal coordinates* [11] are also relevant here.

FOSSILS

The stimulus to study variation in modern groups of man often comes from a particularly important fossil find. If canonical variates for the modern groups have been evaluated, then scores on these variates can be obtained for the fossil, and it can be placed on the plot of modern group means relative to the first two or three canonical variates. This plot indicates its position relative to the modern groups, but the result must be treated with caution. Both the modern groups and the fossil should fit well into the two- or three-dimensional plot, for otherwise their relative positions will be distorted. If the fit is poor, then the actual D^2 distances of the fossil from the modern groups should be compared, but if these are all large, then the assessment of affinity is bound to be unsatisfactory. Day and Wood [9], after associating a canonical variate with function, used it to predict function for a fossil that was very different from the modern groups used to derive the canonical variate. This situation is in some ways analogous to that encountered when using a regression line* to predict values outside the range on which the line was based.

PREVALENCE OF ATTRIBUTES*

Consider υ attributes of an object measured by $X_1, \ldots, X_\upsilon$, where these now take only two possible values (presence/absence). The comparison between groups rests on a comparison of the prevalence for each attribute. If $p(X_i)$ is the prevalence for X_i, then the difference between two groups is measured on the transformed scale $\theta = \sin^{-1}\sqrt{p}$. On this scale the standard deviation of θ is approximately $1/(4n)$. *See* VARIANCE STABILIZATION. If the υ attributes are independent, then υ differences, $\theta_i - \theta_i'$, may be combined to provide an overall distance $d^2 = \sum(\theta_i - \theta_i')^2$. Berry and Berry [4] give an example based on attributes of the skull. Edwards [10] has generalized this to cover attributes with more than two states, such as blood groups.

STATISTICAL TESTS AND PREDICTION

Since there is rarely any element of randomization* in data collection for physical anthropology, the role of significance tests* is less important than that of description. Three tests are commonly performed: equality of covariance matrices between groups, difference between two groups based on D^2, and zero intercorrelation within a set of variables. All three are based on the multivariate normal distribution*. Full details are given in Morrison [17].

A vector $\mathbf{x}$ which is incomplete cannot be used in multivariate analysis* without the missing values being replaced by some estimates. The group means for the relevant variables are sometimes used, but a better method is to use the set of complete vectors to predict the missing values using multiple regression* [3].

Predicting the sex of bones can be dealt with statistically if reference groups of known sex are available. Each unknown bone is allocated to the closer of the male and female reference groups using the D^2-metric. Day and Pitcher–Wilmott [8] give an example. The maturity of bones is assessed from characteristics that can be ordered with respect to maturity. Scaling techniques* are used to make the assessment quantitative [13].

REFERENCES

1. Andrews, D. F. (1972). *Biometrics*, **28**, 125–136.
2. Ashton, E. H., Healy, M. J. R., and Lipton, S. (1957). *Proc. R. Soc. Lond. B*, **146**, 552–572.
3. Beale, E. M. L. and Little, R. J. A. (1975). *J. R. Statist. Soc. B*, **37**, 129–145.
4. Berry, A. C. and Berry, R. J. (1967). *J. Anat.*, **101**, 361–379.

5. Bookstein, F. L. (1978). The Measurement of Biological Shape and Shape Change. *Lect. Notes Biomath.*, 24. Springer-Verlag, Berlin.
6. Burnaby, T. P. (1966). *Biometrics*, **22**, 96–110.
7. Corruccini, R. S. (1973). *Amer. J. Phys. Anthropol.*, **38**, 743–754.
8. Day, M. H. and Pitcher-Wilmott, R. W. (1975). *Ann. Hum. Biol.*, **2**, 143–151.
9. Day, M. H. and Wood, B. A. (1968). *Man*, **3**, 440–455.
10. Edwards, A. W. F. (1971). *Biometrics*, **27**, 873–881.
11. Gower, J. C. (1966). *Biometrika*, **53**, 325–338.
12. Gower, J. C. (1966). *Biometrika*, **53**, 588–590.
13. Healy, M. J. R. and Goldstein, H. (1976). *Biometrika*, **63**, 219–229.
14. Howells, W. W. (1972). In *The Functional and Evolutionary Biology of Primates: Methods of Study and Recent Advances*, R. H. Tuttle, ed. Aldine-Atherton, Chicago, pp. 123–151.
15. Kruskal, J. B. (1964). *Psychometrika*, **29**, 1–27.
16. McLearn, I., Morant, G. M., and Pearson, K. (1928). *Biometrika*, **20B**, 389–400.
17. Morrison, D. F. (1967). *Multivariate Statistical Methods.* McGraw-Hill, New York.
18. Mosimann, J. E. (1970). *J. Amer. Statist. Ass.*, **65**, 930–945.
19. Oxnard, C. E. (1973). *Form and Pattern in Human Evolution*. University of Chicago Press, Chicago.
20. Oxnard, C. E. and Neely, P. M. (1969). *J. Morphol.*, **129**, 1–22.
21. Pearson, E. S. (1936). *Biometrika*, **28**, 193–257.
22. Pearson, E. S. (1938). *Biometrika*, **29**, 161–248.
23. Pearson, K. (1926). *Biometrika*, **18**, 105–117.
24. Penrose, L. S. (1954). *Ann. Eugen. (Lond.)*, **18**, 337–343.
25. Rao, C. R. (1964). *Sankhyā A*, **26**, 329–358.
26. Weiner, J. S. (1972). *The Assessment of Population Affinities*, J. S. Weiner, ed. Clarendon Press, Oxford.
27. Zahn, C. T. and Roskies, R. Z. (1972). *IEEE Trans. Computers*, **C-21**, 269–281.

FURTHER READING

The book by Oxnard [19] provides the best available general introduction to the subject, and contains a good bibliography. The series of papers by Howells (1951 onward; see the bibliography in ref. [19]) contain good examples of most of the important methodological contributions made during this period. The use of canonical variates is well explained in Ashton et al. [2]. Bookstein's monograph [5] provides a useful antidote to the uncritical use of interlandmark distances; although more mathematical than the other references cited it is well worth the effort. Technical details about principal components and canonical variates are best obtained from Rao [25], Gower [11,12], and Morrison [17]. For historical details the early volumes of *Biometrika* should be consulted, particularly Volume 1 and the account of Karl Pearson's life and work given by E. S. Pearson [21,22].

See also ALLOMETRY; CLUSTER ANALYSIS; CORRELATION; DISCRIMINANT ANALYSIS; MULTIDIMENSIONAL SCALING; MULTIVARIATE ANALYSIS; *PATTERN RECOGNITION*; PRINCIPAL COMPONENT ANALYSIS, GENERALIZED; and REGRESSION.

M. HILLS

ANTIEIGENVALUES AND ANTIEIGENVECTORS

INTRODUCTION

It follows from the celebrated Cauchy–Schwarz* inequality that for a real symmetric positive definite matrix $\mathbf{A}$ of order $p \times p, (\mathbf{x}'\mathbf{A}\mathbf{x})^2 \leqslant \mathbf{x}'\mathbf{A}^2\mathbf{x} \cdot \mathbf{x}'\mathbf{x}$, with equality if and only if $\mathbf{A}\mathbf{x}$ is proportional to $\mathbf{x}$. Thus, the optimization problem

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\sqrt{\mathbf{x}'\mathbf{A}^2\mathbf{x} \cdot \mathbf{x}'\mathbf{x}}}, \tag{1}$$

has its optimum value at 1 and it is attained when $\mathbf{x}$ is an eigenvector* of $\mathbf{A}$. Thus, if $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p\}$ is the set of orthogonal eigenvectors of $\mathbf{A}$ corresponding to the eigenvalues* $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$, then any of these eigenvectors will solve Equation 1. The corresponding minimization problem

$$\min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\sqrt{\mathbf{x}'\mathbf{A}^2\mathbf{x} \cdot \mathbf{x}'\mathbf{x}}}, \tag{2}$$

however, is solved by $\left(\frac{\lambda_p}{\lambda_1+\lambda_p}\right)^{1/2} \boldsymbol{\gamma}_1 \pm \left(\frac{\lambda_1}{\lambda_1+\lambda_p}\right)^{1/2} \boldsymbol{\gamma}_p$ and the corresponding minimum value is called the *Kantorovich bound* [17] (*see* KANTOROVICH INEQUALITY), namely, $\frac{2\sqrt{\lambda_1\lambda_p}}{\lambda_1+\lambda_p}$. Since this corresponds to the minimization problem rather than maximization, it may be appropriate to call the quantity

$$\mu_1 = \frac{2\sqrt{\lambda_1\lambda_p}}{\lambda_1 + \lambda_p}$$

an *antieigenvalue* and the solution vectors $\left(\frac{\lambda_p}{\lambda_1+\lambda_p}\right)^{1/2}\boldsymbol{\gamma}_1 \pm \left(\frac{\lambda_1}{\lambda_1+\lambda_p}\right)^{1/2}\boldsymbol{\gamma}_p$ the corresponding *antieigenvectors*. In an attempt to imitate the theory of eigenvalues and eigenvectors, we may call μ_1 the smallest antieigenvalue and the corresponding antieigenvector may be denoted by $\boldsymbol{\eta}_1$ in a generic way. The next antieigenvalue μ_2 and corresponding antieigenvector $\boldsymbol{\eta}_2$ are defined as the optimized value and the solution to the optimization problem

$$\min_{\mathbf{x}\neq\mathbf{0},\mathbf{x}\perp\boldsymbol{\eta}_1} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\sqrt{\mathbf{x}'\mathbf{A}^2\mathbf{x}\cdot\mathbf{x}'\mathbf{x}}}. \qquad (3)$$

The other antieigenvalues and the corresponding antieigenvectors are similarly defined by requiring that the corresponding antieigenvector be orthogonal to all those previously obtained. A set of these orthogonal antieigenvectors is given by $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k\}$, where $k = [\frac{p}{2}]$ and

$$\boldsymbol{\eta}_i = \left(\frac{\lambda_{p-i+1}}{\lambda_i + \lambda_{p-i+1}}\right)^{1/2}\boldsymbol{\gamma}_i + \left(\frac{\lambda_i}{\lambda_i + \lambda_{p-i+1}}\right)^{1/2}\boldsymbol{\gamma}_{p-i+1},$$

$i = 1, 2, \ldots, k$. The corresponding antieigenvalues are given by

$$\mu_i = \frac{2\sqrt{\lambda_i\lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}}.$$

Clearly, because of the arbitrariness of the signs of the eigenvectors, there are 2^k such sets for a given set $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p\}$ of eigenvectors. Further, if there were repeated eigenvalues, say for example, if $\lambda_i = \lambda_{p-i+1}$ for some i, then any vector in the plane generated by $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_{p-i+1}$ is also an eigenvector. Consequently, in this case, $\boldsymbol{\eta}_i$ are eigenvectors as well as antieigenvectors of $\mathbf{A}$. The corresponding antieigenvalue μ_i is equal to 1, which is the value of the maximum as well as of the minimum. From a statistical point of view, these antieigenvalues are uninteresting.

The development of the mathematical theory of antieigenvalues and antieigenvectors can be traced to Gustafson [6] and Davis [3]. Notable contributions were made by Gustafson and his associates in a series of papers [7–16]. Khattree [20–22] provides some statistical interpretations and computational details involving antieigenvalues. Khattree [21] also defines what he terms as the *generalized antieigenvalue of order r* of a symmetric positive definite matrix $\mathbf{A}$ and the corresponding *antieigenmatrix* through the optimization problem

$$\min_{\mathbf{X},\mathbf{X}'\mathbf{X}=\mathbf{I}_r} \frac{|\mathbf{X}'\mathbf{A}\mathbf{X}|}{\sqrt{|\mathbf{X}'\mathbf{A}^2\mathbf{X}|}}, \qquad (4)$$

where the minimized value $\mu_{[r]}$ is the generalized antieigenvalue of order $r(\leqslant [\frac{p}{2}])$, and the corresponding $p \times r$ suborthogonal matrix $\mathbf{X}$ solving the above problem is the antieigenmatrix. Clearly, $\mathbf{X}$ is not unique; for a given $\mathbf{X}$, $\mathbf{PX}$, where $\mathbf{P}$ is orthogonal, also solves Equation 4. In fact [21] one choice of $\mathbf{X}$ is the matrix whose columns are $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_r$. Correspondingly,

$$\mu_{[r]} = \prod_{i=1}^{r} \frac{2\sqrt{\lambda_i\lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \mu_1\mu_2\ldots\mu_r. \qquad (5)$$

A paper by Drury et al. [4] nicely connects antieigenvalues and the generalized antieigenvalue stated in Equation 5 with many matrix inequalities. In that context, although these authors also talk about the quantity given in Equation 5 above, their usage of the term "generalized antieigenvalue" should be contrasted with our definition. In fact, their generalization, as stated in their Theorem 1, is toward replacing the positive definiteness of $\mathbf{A}$ by nonnegative definiteness, and in the context of Löwner ordering.

Gustafson [11] provides an interesting interpretation of $\cos^{-1}(\mu_1)$ as the largest angle by which $\mathbf{A}$ is capable of turning any vector $\mathbf{x}$. The quantities $\cos^{-1}(\mu_i), i = 2, 3, \ldots$ can be similarly defined, subject to the orthogonality conditions as indicated in Equation 3.

STATISTICAL APPLICATIONS

The concepts of antieigenvalues and antieigenvectors were independently used in statistics about the same time as these were discovered in mathematics, although without any such nomenclatures. It is appropriate to interpret these quantities statistically and provide a reference to the contexts in which they naturally arise.

Let $\mathbf{z}$ be a $p \times 1$ random vector with variance-covariance matrix* proportional to the identity matrix $\mathbf{I}_p$. The problem of finding a nonnull vector $\mathbf{x}$, such that the correlation* between the linear combinations $\mathbf{z}'\mathbf{x}$ and $\mathbf{z}'\mathbf{A}\mathbf{x}$ (with $\mathbf{A}$ symmetric positive definite) is minimum, is essentially the problem stated in Equation 2 [22]. Under certain conditions on the correlation structure, it is also the minimum possible value of the parent-offspring correlation [22].

Venables [27], Eaton [5], and Schuenemeyer and Bargman [25] have shown that a lower bound on certain canonical correlation* is given by $1 - \mu_1^2$. Furthermore, this bound is sharp. Bartmann and Bloomfield [1] generalize this result by stating that

$$\prod_{i=1}^{r}(1 - \rho_i^2) \geqslant \mu_{[r]}, \tag{6}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

with $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_{11}$, and $\boldsymbol{\Sigma}_{22}$ as symmetric positive definite variance-covariance matrices of order $p \times p, r \times r$ and $(p - r) \times (p - r)$, respectively, and the canonical correlations $\rho_i, i = 1, 2, \ldots, r$ are defined by $\rho_i^2 = i^{th}$ eigenvalue of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Without any specific reference to the term antieigenvalue, their problem of minimizing the left-hand side of Equation 6 essentially boils down to that stated in Equation 4.

Bloomfield and Watson [2] and Knott [23] in two back to back articles in *Biometrika** consider the problem of determining the efficiency of least square* estimator of $\boldsymbol{\beta}$ in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{A})$ set up. Let $\mathbf{X}$ be of order $n \times k$ with $k \leqslant [\frac{n}{2}]$. If the least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is used instead of the general least squares estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{y},$$

then the relative efficiency e of $\hat{\boldsymbol{\beta}}$ relative to $\tilde{\boldsymbol{\beta}}$ may be defined as the ratio of the generalized variances

$$e = \frac{|D(\tilde{\boldsymbol{\beta}})|}{|D(\hat{\boldsymbol{\beta}})|} = \frac{|\mathbf{X}'\mathbf{X}|^2}{|\mathbf{X}'\mathbf{A}^{-1}\mathbf{X}||\mathbf{X}'\mathbf{A}\mathbf{X}|},$$

where $D(.)$ stands for the variance-covariance matrix. The above-mentioned authors were then interested in finding a lower bound on the above, which if attained, would represent the worst-case scenario. This problem can be reduced to Equation 4 with the transformation $\mathbf{Z} = \mathbf{A}^{-\frac{1}{2}}\mathbf{X}$. Thus, the worst case occurs when the columns of $\mathbf{Z}$ (or $\mathbf{X}$) span the subspace generated by the antieigenvectors $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k\}$. The minimum value of the relative efficiency is given by $\mu_{[k]}$ as in Equation 5.

Venables [27] has considered the problem of testing sphericity* $(H_0 : \boldsymbol{\Sigma} = \sigma^2\mathbf{I})$ of a variance-covariance matrix $\boldsymbol{\Sigma}$. The sample variance-covariance matrix $\mathbf{S}$ is given as data and $f\mathbf{S} \sim W_p(f, \boldsymbol{\Sigma})$, the Wishart distribution* with f degrees of freedom along with $E(\mathbf{S}) = \boldsymbol{\Sigma}$. He argues that the null hypothesis is true if and only if any arbitrary r-dimensional subspace $(r \leqslant p)$ of $\boldsymbol{\Sigma}$ is an invariant subspace. Thus, the null hypothesis is the intersection of all such subhypotheses stating the invariance, and the intersection is taken over all such subspaces. The likelihood ratio test for all such subhypotheses can be derived, and with an argument that one accepts H_0 if and only if all such subhypotheses are accepted, a test statistic for H_0 can be obtained by taking the maximum of such test statistics. This maximization problem turns out to be equivalent to Equation 4 and hence the test statistic for testing H_0 is given by $\mu_{[r]}$ computed from the matrix $\mathbf{S}$. The result can be intuitively justified from the fact that each product term in Equation 5 provides a measure of the eccentricity of the ellipsoids defined by $\mathbf{S}$. Another article [18] and its correction in Reference 19 deal with the same problem but using a different mathematical argument, and still arrives at $\mu_{[r]}$ as the test statistic. Also see Reference 26.

Since the Kantorovich inequality plays an important role in determining the rates of convergence [24] in many mathematical optimization problems, the smallest antieigenvalue

$$\mu_1 = \frac{2\sqrt{\lambda_1\lambda_p}}{\lambda_1+\lambda_p} = 2\left[\sqrt{\frac{\lambda_1}{\lambda_p}}+\sqrt{\frac{\lambda_p}{\lambda_1}}\right]^{-1},$$

has a one-to-one correspondence with the condition number λ_1/λ_p. Thus, μ_1 and, more generally, $\mu_1, \mu_2 \ldots$ as well as $\mu_{[r]}$ can be viewed as the condition indices. In linear models, their applications hold promise in assessing multicollinearity* problems.

The topic of antieigenvalues and antieigenvectors may not be very well known among mathematicians. Perhaps statisticians have been slightly ahead of mathematicians in using the concepts implicit in the definitions of the antieigenvalues and antieigenvectors. It is anticipated that some of the mathematical research done on this topic will find further applications in statistics and will provide insight into many other statistical problems.

Acknowledgment
The author's work was partially sponsored by an Oakland University Research fellowship.

REFERENCES

1. Bartmann, F. C. and Bloomfield, P. (1981). Inefficiency and correlation. *Biometrika*, **68**, 67–71.
2. Bloomfield, P. and Watson, G. S. (1975). The inefficiency of least squares. *Biometrika*, **62**, 121–128.
3. Davis, C. (1980). Extending the Kantorovich inequalities to normal matrices. *Lin. Alg. Appl.*, **31**, 173–177.
4. Drury, S. W., Liu, S., Lu, C.-Y, Puntanen, S., and Styan, G. P. H. (2002). *Some Comments on Several Matrix Inequalities with Applications to Canonical Correlations: Historical Background And Recent Developments*. Sankhyā, A64, 453–507.
5. Eaton, M. L. (1976). A maximization problem and its application to canonical correlation. *J. Multivariate Anal.*, **6**, 422–425.
6. Gustafson, K. (1972). "Antieigenvalue Inequalitities in Operator Theory". In *Inequalities III, Proceedings of the Los Angeles Symposium 1969*, O. Shisha, ed. Academic Press, New York, pp. 115–119.
7. Gustafson, K. (1994a). Operator trigonometry. *Lin. Mult. Alg.*, **37**, 139–159.
8. Gustafson, K. (1994b). Antieigenvalues. *Lin. Alg. Appl.*, **208/209**, 437–454.
9. Gustafson, K. (1995). Matrix trigonometry. *Lin. Alg. Appl.*, **217**, 117–140.
10. Gustafson, K. (1996). *Operator Angles (Gustafson), Matrix Singular Angles (Wielandt), Operator Deviations (Krein), Collected Works of Helmug Wielandt II*, B. Huppert and H. Schneider, eds. De Gruyters, Berlin, Germany.
11. Gustafson, K. (1999). The geometrical meaning of the Kantorovich-Wielandt inequalities. *Lin. Alg. Appl.*, **296**, 143–151.
12. Gustafson, K. (2000). An extended operator trigonometry. *Lin. Alg. Appl.*, **319**, 117–135.
13. Gustafson, K. and Rao, D. K. M. (1977a). Numerical range and accretivity of operator products. *J. Math. Anal. Appl.*, **60**, 693–702.
14. Gustafson, K. and Rao, D. K. M. (1997b). *Numerical Range: The Field of Values of Linear Operators and Matrices*. Springer-Verlag, New York.
15. Gustafson, K. and Seddighin, M. (1989). Antieigenvalue bounds. *J. Math. Anal. Appl.*, **143**, 327–340.
16. Gustafson, K. and Seddighin, M. (1993). A note on total antieigenvectors. *J. Math. Anal. Appl.*, **178**, 603–611.
17. Kantorovich, L. B. (1948). Functional analysis and applied mathematics. *Uspehi. Mat. Nauk.*, **3**, 89–185 (Translated from the Russian by Benster, C. D. (1952), Report 1509, National Bureau of Standards, Washington, D.C).
18. Khatri, C. G. (1978). Some optimization problems with applications to canonical correlations and sphericity tests, *J. Multivariate Anal.*, **8**, 453–467 (Corrected by Khatri [19]).
19. Khatri, C. G. (1982). *J. Multivariate Anal.*, **12**, 612; Erratum (to Khatri [18]).
20. Khattree, R. (2001). On calculation of antieigenvalues and antieigenvectors. *J. Interdiscip. Math.*, **4**, 195–199.
21. Khattree, R. (2002). Generalized antieigenvalues of order *r*. *Am. J. Math. Manage. Sci.*, **22**, 89–98.
22. Khattree, R. (2003). Antieigenvalues and antieigenvectors in statistics. *J. Stat. Plann. Inference* (Special issue in celebration of C.R. Rao's 80th birthday), **114**, 131–144.
23. Knott, M. (1975). On the minimum efficiency of least squares. *Biometrika*, **62**, 129–132.

24. Luenberger, D. (1984). *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, Mass.
25. Schuenemeyer, J. H. and Bargman, R. E. (1978). Maximum eccentricity as a union-intersection test statistic in multivariate analysis. *J. Multivariate Anal.*, **8**, 268–273.
26. Srivastava, M. S. and Khatri, C. G. (1978). *An Introduction to Multivariate Statistics*, North Holland, New York.
27. Venables, W. (1976). Some implications of the union-intersection principle for tests of sphericity, *J. Multivariate Anal.*, **6**, 185–190.

See also CANONICAL ANALYSIS; KANTOROVICH INEQUALITY; LINEAR ALGEBRA, COMPUTATIONAL; and SPHERICITY, TESTS OF.

RAVINDRA KHATTREE

ANTIMODE

An antimode is the opposite of a mode* in the sense that it corresponds to a (local) minimum frequency. As with the mode, it is sometimes desired that the name should be applied only to global, and not to local minima. The more common use, however, includes local minima.

Note that whereas $x = 1$ is a mode of the PDF*,

$$f_X(x) = \begin{cases} 2x & 0 \leqslant x \leqslant 1, \\ 0 & \text{elsewhere,} \end{cases}$$

$x = 0$ is not an antimode of this PDF. On the other hand,

$$f_X(x) = \begin{cases} |x| & -1 \leqslant x \leqslant 1, \\ 0 & \text{elsewhere,} \end{cases}$$

has an antimode at $x = 0$ (and modes at $x = -1$ and $x = 1$).

The antimode itself refers to the frequency (or PDF) at the antimodal value of the argument.

See also MEAN, MEDIAN, AND MODE.

ANTIRANKS

Nonparametric tests and estimates are generally based on certain statistics which depend on the sample observations $X_1, \ldots, X_n$ (real-valued) only through their ranks* $R_1, \ldots, R_n$, where

$$R_i = \text{number of indices } r \quad (1 \leqslant r \leqslant n) : X_r \leqslant X_i, \qquad (1)$$

for $i = 1, \ldots, n$. If $X_{n:1} \leqslant \cdots \leqslant X_{n:n}$ stand for the sample order statistics*, then we have

$$X_i = X_{n:R_i}, \quad i = 1, \ldots, n. \qquad (2)$$

Adjustment for ties can be made by dividing equally the total rank of the tied observations among themselves. Thus if $X_{n:k} < X_{n:k+1} = \cdots = X_{n:k+q} < X_{n:k+q+1}$, for some $k, 0 \leqslant k \leqslant n-1$, and $q \geqslant 1$ (where $X_{n:0} = -\infty$ and $X_{n:n+1} = +\infty$), then for the q tied observations (with the common value $X_{n:k+1}$), we have the *midrank* $k + (q+1)/2$.

Let us now look at (2) from an opposite angle: For which index (S_k), is X_{S_k} equal to $X_{n:k}$? This leads us to define the *antiranks* $S_1, \ldots, S_n$ by

$$X_{n:i} = X_{S_i}, \quad \text{for } i = 1, \ldots, n. \qquad (3)$$

Note the inverse operations in (2) and (3) as depicted below:

$$\begin{tikzcd}
X_1, & \ldots, & X_i, \arrow[drr] & \ldots, & X_{S_i}, \arrow[dll] & \ldots, & X_n \\
X_{n:1}, & \ldots, & X_{n:i}, & \ldots, & X_{n:R_i}, & \ldots, & X_{n:n},
\end{tikzcd} \qquad (4)$$

so that, we have

$$R_{S_i} = S_{R_i} = i, \quad \text{for } i = 1, \ldots, n, \qquad (5)$$

and this justifies the terminology: Antiranks.

Under the null hypothesis that $X_1, \ldots, X_n$ are exchangeable random variables, $\mathbf{R} = (R_1, \ldots, R_n)$, the vector of ranks, takes on each permutation of $(1, \ldots, n)$ with the common probability $(n!)^{-1}$. By virtue of (5), we obtain that under the same null hypothesis, $\mathbf{S} = (S_1, \ldots, S_n)$, the vector of antiranks, has the same (discrete) uniform permutation distribution. In general, for the case of ties neglected, (5) can be used to obtain the distribution of $\mathbf{S}$ from that of $\mathbf{R}$ (or vice versa), although when the X_i are not exchangeable, this distribution may become quite cumbrous. Under the null hypothesis of exchangeability*, for suitable functions of $\mathbf{R}$ (i.e. rank- statistics), *permutational central limit theorems** provide asymptotic solutions, and by virtue of (5), these remain applicable to antirank statistics as well.

For mathematical manipulations, often $\mathbf{S}$ may have some advantage over $\mathbf{R}$. To illustrate this point, consider a typical *linear rank statistic** (T_n) of the form $\sum_{i=1}^{n} c_i a_n(R_i)$, where the c_i are given constants and $a_n(1), \ldots, a_n(n)$ are suitable scores. By (5), we have

$$T_n = \sum_{i=1}^{n} c_{S_i} a_n(i), \qquad (6)$$

and this particular form is more amenable to censoring schemes (*see* PROGRESSIVE CENSORING SCHEMES). If we have a type II censoring (at the kth failure), then the censored version of T_n in (6) is given by

$$T_{nk} = \sum_{i=1}^{k} (c_{S_i} - \bar{c}_n)[a_n(i) - a_n^*(k)], \quad k \geqslant 0, \qquad (7)$$

where

$$\bar{c}_n = n^{-1} \sum_{i=1}^{n} c_i,$$

$$a_n^*(k) = (n-k)^{-1} \sum_{j=k+1}^{n} a_{n(j)},\ k < n$$

and

$$a_n^*(n) = 0.$$

For the classical *Kolmogorov–Smirnov tests**, these antiranks may be used to express the statistics in neat forms and to study their distributions in simpler manners; we may refer to Hájek and Šidák [1] for a nice account of these.

In life-testing* problems and clinical trials*, for rank procedures in a time-sequential setup, the antiranks play a vital role; for some details, see Sen [2,3].

REFERENCES

1. Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic, New York.
2. Sen, P. K. (1981). *Sequential Nonparametrics*. Wiley, New York.
3. Sen, P. K. (1985). *Theory and Application of Sequential Nonparametrics*. SIAM, Philadelphia.

See also EMPIRICAL DISTRIBUTION FUNCTION (EDF) STATISTICS; KOLMOGOROV–SMIRNOV STATISTICS; LIMIT THEOREM, CENTRAL; LINEAR RANK TESTS; ORDER STATISTICS; PERMUTATIONAL CENTRAL LIMIT THEOREMS; PROGRESSIVE CENSORING SCHEMES; and TIME-SEQUENTIAL INFERENCE.

P. K. SEN

ANTISMOOTHING. See FOURIER SERIES CURVE ESTIMATION AND ANTISMOOTHING

ANTITHETIC VARIATES

If T_1 and T_2 are unbiased estimators of a parameter θ, then $T = \frac{1}{2}(T_1 + T_2)$ is also unbiased and has variance $\frac{1}{4}\{\text{var}(T_1) + \text{var}(T_2) + 2\,\text{cov}(T_1, T_2)\}$. This variance is reduced (for fixed $\text{var}(T_1)$ and $\text{var}(T_2)$) by reducing $\text{cov}(T_1, T_2)$ and making the correlation between T_1 and T_2 negative and as large, numerically, as possible. Pairs of variates constructed with this aim in view are called *antithetic variates*.

The concept arose in connection with estimation of integrals by simulation experiments [2,3] (*see* MONTE CARLO METHODS). The following example [1, p. 61] may help to clarify ideas. If X is uniformly distributed* between 0 and 1, then for any function $g(x)$,

$$E\left[g(X)\right] = \int_0^1 g(x)\,dx,$$

so $g(x)$ is an unbiased estimator* of $\int_0^1 g(x)dx$. So is $g(1-X)$, since $(1-X)$ is also uniformly distributed between 0 and 1. If $g(x)$ is a monotonic function of x, $g(x)$ and $g(1-X)$ are negatively correlated and are "antithetic variates." In particular,

$$\text{var}(g(X) + g(1-X)) \leqslant \tfrac{1}{2}\,\text{var}(g(X)).$$

The construction of a variate antithetic to $g(X)$ can be extended in a simple way to cases when the function $g(x)$ is not monotonic, but the interval 0 to 1 can be split into a finite number of intervals in each of which $g(x)$ is monotonic.

The method can also be applied to estimation of multivariate integrals by a straightforward extension. Use of antithetic variables can be a powerful method of increasing accuracy of estimation from simulation in appropriate situations.

REFERENCES

1. Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Methuen, London.
2. Hammersley, J. M. and Mauldon, J. G. (1956). *Proc. Camb. Philos. Soc.*, **52**, 476–481.
3. Hammersley, J. M. and Morton, K. W. (1956). *Proc. Camb. Philos. Soc.*, **52**, 449–475.
4. Tukey, J. W. (1957). *Proc. Camb. Philos. Soc.*, **53**, 923–924.

See also MONTE CARLO METHODS and NUMERICAL INTEGRATION.

APL. See STATISTICAL SOFTWARE

APPLIED PROBABILITY

Applied probability is that field of mathematical research and scholarship in which the theory and calculus of probability are applied to real-life phenomena with a random component. Such applications encompass a broad range of problems originating in the biological, physical, and social sciences, as well as engineering and technology.

The term "applied probability" first appeared as the title of the proceedings of a symposium on the subject, held by the American Mathematical Society in 1955 [3]. It became popular through its use by the *Methuen Monographs in Applied Probability and Statistics*, edited from 1959 by M. S. Bartlett. The two fields are closely related: applied probability is concerned primarily with modeling random phenomena* (*see* STOCHASTIC PROCESSES), while statistics serves to estimate parameters* and test the goodness of fit* of models to observed data. Bartlett has expressed the opinion that neither field could exist without the other; this is a viewpoint shared by many applied probabilists.

There are currently several periodicals publishing material in applied probability; the principal ones in order of their dates of first publication are *Teoriya Veroyatnostei i ee Primeneniya** (1956), (English translation: *Theory of Probability and Its Applications**) *Zeitschrift für Wahrscheinlichkeitstheorie* (1962), the *Journal of Applied Probability** (1964), *Advances in Applied Probability** (1969), *Stochastic Processes and Their Applications** (1973), and *Annals of Probability** (1973). Other mathematical or statistical journals publish the occasional paper in applied probability, and there is considerable discussion of applied probability models in journals of biology, physics, psychology, operations research, and engineering. Among these are *Theoretical Population Biology*, the *Journal of Statistical Physics**, *Psychometrika, Operations Research**, the *Journal of Hydrology*, and the *Journal of the Institute of Electrical Engineers*.

It is impossible to give a comprehensive description of current work in applied probability; perhaps a few illustrative examples selected at random from the recent literature in each of the biological, physical, social, and technological areas will serve to indicate the breadth of the field.

BIOLOGICAL SCIENCES: BIRD NAVIGATION; OPTIMAL HUNTING

David Kendall [2] investigated some interesting models of bird navigation. Ornithologists have surmised that birds navigate instinctively by reference to the sun and stars; Kendall constructed two models to simulate such navigation realistically.

The first is referred to as the Manx model, after the Manx shearwater, which flies across the Atlantic to its European breeding grounds. In this model the bird flies laps of approximately 20 miles each, at a speed of roughly 40 mph. At the end of each lap it redirects itself toward its goal but commits an angular error, having the von Mises or wrapped normal distribution*. When it arrives within a radius of approximately 10 miles of its home, it recognizes its destination and heads directly for it. The second Bessel model allows both the lap length and the deflection from the correct direction to be random. Under appropriate conditions, the two models converge to Brownian motion*.

In his paper, Kendall analyzes bird data on which to base his models and tests these by repeated simulations*, leading to graphical representations of Manx and Bessel flights. He compares these models theoretically and numerically, carries out some diffusion* approximations, and concludes with a lengthy study of the hitting times to the circumference of the homing target. Kendall's work is a model of what is most illuminating in the applied probabilist's approach to a scientific problem. Practical data are carefully considered, and a suitable model is found to fit it. The apprentice applied probabilist could well base his methods and style on Kendall's work. *See also* ORNITHOLOGY, STATISTICS IN.

Another of the many interesting problems which arise in a biological context is that of determining optimal hunting or harvesting policies for animal populations. Abakuks [1] has discussed such a policy for a population growing according to a stochastic logistic* scheme, subject to natural mortality. The object is to maximize the long-term average number of animals hunted or harvested per unit time. It is shown that there is a critical population size x_c such that hunting or harvesting is optimal if and only if the population is greater or equal to this number.

PHYSICAL SCIENCES: ISING LATTICES

In statistical mechanics, one may need to determine the partition function for large lattices of points representing crystals, particles, or atoms. The Ising model, which helps to characterize qualitative changes at critical parameter values, is therefore important in theoretical physics (*see* LATTICE SYSTEMS).

Consider a rectangular lattice of $N = m \times n$ points in a plane; these points are labeled 1 to N. At each site i, the random variable X_i may take the values ± 1, where $+1$ may, for example, correspond to the formation of a crystal. The joint distribution of the site variables is given by

$$P\{\mathbf{X} = \mathbf{x}\} = k^{-1}(a) \exp\left(a \sum x_i x_j\right),$$

where $k(a) = \sum_{\mathbf{x}} \exp\left(a \sum x_i x_j\right)$ is the partition function, and the $\sum x_i x_j$ is taken over all nearest-neighbor* pairs; $\mathbf{X}$ is a simple Markov random field*.

Pickard [4] has recently obtained some limit theorems for the sample correlation* between nearest neighbors in an Ising lattice for the noncritical case. This provides a model for asymptotic testing and estimation of the correlation between nearest neighbors, based on experimental data.

SOCIAL SCIENCES: MANPOWER SYSTEMS; ECONOMIC OPTIMIZATION

A firm or company is a hierarchically graded manpower system, usually modeled by a Markov chain. The probabilities p_{ij} of this chain denote annual promotion rates from grade i to grade j of the firm; the state of the graded manpower system is described from year to year by the numbers of individuals in each grade (*see* MANPOWER PLANNING).

The simpler manpower models are mainly linear, but Vassiliou [7] considered a high-order nonlinear Markovian model for promotion based on three principles. The first is the ecological principle that promotions should be proportional to suitable staff available for them, as well as to vacancies for promotion. The second is that resignations from different grades are different, and the third is an inertia principle which prescribes that when there is a reduced number of vacancies, promotions may still be made faster than the ecological principle suggests.

Difference equations* for the mean number of individuals in each grade are obtained, and the model is used to provide detailed numerical forecasts for probabilities of promotion in organizations with five grades. A comparison is made between predictions based on an earlier linear model*, and the present nonlinear model*; actual data from a large British firm are found to be adequately described by the latter.

Optimization may be important in a variety of sociological contexts, some purely economic, others more technological. Vered and Yechiali [8] studied the optimization of a power system for a private automatic telephone exchange (PABX). Several maintenance policies are considered, which depend on the set of parameters (m, u, μ, υ) of the PABX system, where n is the number of independent rectifiers in parallel, m the minimal number of units that will keep the system operative, μ their mean time to failure* and υ the period of time between regular maintenance visits. Repairs are carried out every υ units of time, or when the system fails. The authors determine the optimal (m, n, μ, υ) to minimize costs for a required level of reliability* of the PABX system, and provide tables of numerical results for the optimal parameters in both the cases of periodic and emergency maintenance.

TECHNOLOGICAL SCIENCES: RELIABILITY

All engineering components have a failure point; it is therefore of importance to study the reliability of mechanical or electronic parts to determine the probability of their failure times. In this context, two recent problems, the first studied by Szász [5] concerning two lifts, and the second by Taylor [6] on the failure of cables subjected to random loads, will be of interest.

Szász [5] examines a building that has two lifts working independently of each other. The functioning of each lift forms an alternating renewal process* with working-time distribution F and repair-time distribution G. Suppose that this latter distribution $G = G(x, \epsilon)$ depends on $\epsilon > 0$ in such a way that its mean $\int x\,dG(x, \epsilon)$ tends to zero as $\epsilon \to 0$. The author sets out to find the asymptotic distribution of the first instant τ^{ϵ} at which both lifts are simultaneously out of order, as $\epsilon \to 0$.

It is shown that under certain conditions, as $\epsilon \to 0$, the normalized point process W^{ϵ} : $w_1^{\epsilon} < w_2^{\epsilon} < \cdots$, where $w_k^{\epsilon} = \epsilon\tau_k^{\epsilon}$, tends to a Poisson process* with parameter $2\lambda^{-2}$, where $\lambda = \int x\,dF(x)$. Thus subject to certain very general conditions, one must expect breakdowns of both lifts to occur according to a Poisson process. For the favorable case in which λ is very large, this process will have a very small mean.

Taylor [6] is concerned with the reliability of deep-sea cables made up of fiber bundles, and the effect on them of random loads generated by waves that rock the ocean vessels deploying them. A simple model with random loads is studied, subject to power-law breakdown, such that the failure time T under constant load L follows the negative exponential distribution*

$$\Pr[T > x] = \exp(-KL^{\rho}x) \qquad (x \geqslant 0),$$

where $K > 0$ and $\rho \geqslant 1$.

The asymptotic distribution of T under random loads is derived and Taylor shows that random loads have a significant effect on the lifetime of a cable. The loss in mean lifetime cannot be predicted from the first few moments of the load process; it depends on the entire marginal probability distribution of the load, as well as the power-law exponent ρ. It is shown that the asymptotic variance of the lifetime has two components, the first due to the variation of individual fibers, and the second to the variation of the load.

EXTENT AND FUTURE OF THE FIELD

It is of interest to know the methods of probability theory that are most used in attacking problems of applied probability. The most commonly applied areas are found to be Markov chains and processes (including diffusion processes*), branching processes* and other stochastic processes* (mostly stationary), limit theorems*, distribution theory and characteristic functions*, methods of geometrical probability*, stopping times, and other miscellaneous methods.

To list the subsections of the main categories given in the first four sections exhaustively would be impossible, but in the biological sciences one finds that population processes, mathematical genetics, epidemic theory, and virology are the major subfields. In the technological sciences, operations research*, queueing theory*, storage (*see* DAM THEORY), and traffic theory are possibly the most active areas.

Applied probability is a very broad subject; as we have seen, it encompasses real-life problems in a variety of scientific and other fields. Although the subject feeds on practical problems, it requires a very high level of theoretical competence in probability. In solving these problems, every approach that proves successful is a useful one. Classical mathematical analysis, numerical analysis, statistical calculations, limit theorems*, simulation, and every other branch of mathematics are legitimate weapons in the search for a solution. Applied probability, although a relatively small branch of mathematics, relies on the resources of the entire subject. It maintains itself successfully without specific affiliation to any particular school or tradition, whether it be British, French, Russian, or North American, while drawing on the best aspects of them all. Its strength lies in the universality of its traditions and the versatility of its mathematical methods.

A further point of importance is the delicate interrelation of theory and practice in applied probability. Without practice (which involves computation and statistics), applied probability is trivial; without theory, it becomes shallow. Close contact is required with experiment and reality for the healthy development of the subject. The collection and analysis of data cannot be avoided, and a certain amount of numerical work will always prove necessary. In attacking problems of applied probability there is a complete cycle from the examination of data to the development of a theoretical model, followed by the statistical verification of the model and its subsequent refinement in the light of its goodness of fit*.

It seems possible that too much effort has been diverted into model building for its own sake, as well as in following through the mathematical refinements of new models. The further development of applied probability requires consideration of real-life problems and the validation of models for these based on observed data. Research workers in the field are aware that only by paying close attention to data and considering genuine problems can their contributions to the subject achieve full scientific stature. For interested readers, a selected bibliography is appended.

REFERENCES

Note: The following references are all highly technical.

1. Abakuks, A. (1979). *J. Appl. Prob.*, **16**, 319–331.
2. Kendall, D. G. (1974). *J. R. Statist. Soc. B*, **36**, 365–402.
3. McColl, L. A., ed. (1957). *Applied Probability* [Proc. Symp. Appl. Math., 7 (1955)]. McGraw-Hill, New York (for the American Mathematical Society.)
4. Pickard, D. (1976). *J. Appl. Prob.*, **13**, 486–497.
5. Szász, D. (1977). *Ann. Prob.*, **5**, 550–559.
6. Taylor, H. M. (1979). *Adv. Appl. Prob.*, **11**, 527–541.
7. Vassiliou, P.-C. G. (1978). *J. R. Statist. Soc. A*, **141**, 86–94.
8. Vered, G. and Yechiali, U. (1979). *Operat. Res.*, **27**, 37–47.

BIBLIOGRAPHY

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*. Charles Griffin, London. (Detailed survey of modeling of infectious diseases.)

Barlow, R. E. and Proschan, F. (1965). *Mathematical Theory of Reliability*. Wiley, New York. (Basic principles of reliability.)

Cohen, J. W. (1969). *The Single Server Queue*. North-Holland, Amsterdam. (Comprehensive treatise on queueing.)

Ewens, W. J. (1969). *Population Genetics*. Methuen, London. (Short monograph on mathematical genetics.)

Iosifescu, M. and Tautu, P. (1973). *Stochastic Processes and Applications in Biology and Medicine*, 2 vols. Springer-Verlag, Berlin. (Compendium of stochastic methods in biology and medicine.)

Keyfitz, N. (1968). *Introduction to Mathematics of Population*. Addison-Wesley, Reading, Mass. (Basic principles of demography.)

Kimura, M. and Ohta, T. (1971). *Theoretical Aspects of Population Genetics*. Monographs in Population Biology 4. Princeton University Press, Princeton, N.J. (Monograph on population genetics with emphasis on diffusion.)

Newell, G. F. (1971). *Applications of Queueing Theory*. Chapman & Hall, London. (Short practical account of queueing applications.)

Pielou, E. C. (1969). *An Introduction to Mathematical Ecology*. Wiley-Interscience, New York. (Basic principles of mathematical ecology.)

Pollard, J. H. (1973). *Mathematical Models for the Growth of Human Populations*. Cambridge University Press, Cambridge. (Good introduction to human population models.)

Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco. (Elementary text with broad range of examples.)

Syski, R. (1960). *Introduction to Congestion Theory in Telephone Systems*. Oliver & Boyd, Edinburgh. (Stochastic processes in telephone traffic.)

Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press, New York. (Introductory treatment of queueing theory.)

Thompson, C. J. (1972). *Mathematical Statistical Mechanics*. Macmillan, New York. (Introduction to statistical mechanics.)

See also CRYSTALLOGRAPHY, STATISTICS IN; DAMAGE MODELS; DAM THEORY; ECOLOGICAL STATISTICS; QUEUEING THEORY; RELIABILITY; RENEWAL THEORY; and STATISTICAL GENETICS; STOCHASTIC PROCESSES.

J. GANI

APPLIED PROBABILITY JOURNALS

[*Journal of Applied Probability (JAP); Advances in Applied Probability (AAP)*]

[This entry has been updated by the Editors.]

JAP is an international journal which first appeared in June 1964; it is published by the Applied Probability Trust, based in the University of Sheffield, England. For some years it was published also in association with the London Mathematical Society. The journal provides a forum for research papers and notes on applications of probability theory to biological, physical, social, and technological problems. A volume of approximately 900 to 1,250 pages is published each year, consisting of four issues, which appear in March, June, September, and December. *JAP* considers research papers not exceeding 20 printed pages in length, and short communications in the nature of notes or brief accounts of work in progress, and pertinent letters to the editor.

AAP is a companion publication of the Applied Probability Trust, launched in 1969. It publishes review and expository papers in applied probability, as well as mathematical and scientific papers of interest to probabilists. A volume of approximately 900 to 1,200 pages is published each year; it also consists of four issues appearing in March, June, September, and December. *AAP* considers review papers; longer papers in applied probability which may include expository material, expository papers on branches of mathematics of interest to probabilists; papers outlining areas in the biological, physical, social, and technological sciences in which probability models can be usefully developed; and papers in applied probability presented at conferences that do not publish their proceedings. In addition, *AAP* has a section featuring contributions relating to stochastic geometry and statistical applications (*SGSA*). Occasionally, a special *AAP* supplement is published to record papers presented at a conference of particular interest.

JAP and *AAP* have a wide audience with leading researchers in the many fields where stochastic models are used, including operations research, telecommunications, computer engineering, epidemiology, financial mathematics, information systems and traffic management.

EARLY HISTORY

In 1962, the editor-in-chief of the two journals, J. Gani, made his first attempts to launch the *Journal of Applied Probability*. At that time, a large number of papers on applications of probability theory were being published in diverse journals dealing with general science, statistics, physics, applied mathematics, economics, and electrical engineering, among other topics. There were then only two probability journals, the Russian *Teoriya Veroyatnostei* (English translation: *Theory of Probability and Its Applications**) and the German *Zeitschrift für Wahrscheinlichkeitstheorie**. Neither of these specialized in applications of probability theory, although papers in applied probability occasionally appeared in them.

Having assembled an editorial board in 1962, Gani attempted to launch the *Journal of Applied Probability* with assistance from the Australian National University and the Australian Academy of Science. He was not successful in this, and it was only after raising private contributions from himself, Norma McArthur, and E. J. Hannan*, both of the Australian National University, that he was able to provide half the finance necessary for the publication of *JAP*. In May 1963, with the support of D. G. Kendall of the University of Cambridge, he was able to persuade the London Mathematical Society to donate the remaining half of the funds required, and thus collaborate in the publication of the journal.

In February 1964, agreement was reached that the ownership of *JAP* should be vested in the Applied Probability Trust, a non-profit-making organization for the advancement of research and publication in probability, and more generally mathematics. The four trustees were to include the three Australian sponsors of the journal and one trustee nominated by the London Mathematical Society. The agreement was ratified legally on June 1, 1964, although de facto collaboration had already begun several months before.

The first issue of *JAP* appeared in June 1964. Every effort was made to prevent the time lag between submission and publication of a paper exceeding 15 months; this became an established policy for both *JAP* and *AAP*.

ORGANIZATION

The office of the Applied Probability Trust is located at the University of Sheffield, England. The website for the Trust, and a link to the journals, is www.shef.ac.uk/uni/companies/apt.

Each of the two journals has an international Editorial Board consisting of an Editor-in-Chief, two (*JAP*) or three (*AAP*) Coordinating Editors and between 20 and 30 Editors, some of whom serve on the Board for both journals.

It is impossible for the *JAP* and *AAP* to collect together every diverse strand of the subject, but probabilists can now find most of the applied material they require in a few periodicals. Among these are the two mentioned earlier, as well as the more recent *Stochastic Processes and Their Applications**, and the Institute of Mathematical Statistics' *Annals of Probability**, both first published in 1973, also the *Annals of Applied Probability**, launched by the IMS in 1991, and the journal *Bernoulli*, since 1995 a publication of the International Statistical Institute*. It is the policy of *JAP* and *AAP* to contribute to future development.

J. Gani
The Editors

APPLIED STATISTICS. See *Journal of the Royal Statistical Society*

APPROGRESSION

Approgression is the use of regression functions (usually linear) to approximate the truth, for simplicity and predictive efficiency. Some *optimality* results are available.

The term seems to be due to H. Bunke in

Bunke, H. (1973). Approximation of regression functions. *Math. Operat. Statist.*, **4**, 314–325.

A forerunner of the concept is the idea of an *inadequate* regression model in

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Ass.*, **54**, 622–654.

The paper:

Bandemer, H. and Näther, W. (1978). On adequateness of regression setups. *Biom. J.*, **20**, 123–132

helps one to follow some of the treatment in Section 2.7 of

Bunke, H. and Bunke, O., eds. (1986). *Statistical Inference in Linear Models*. Wiley, New York,

which is exceedingly general and notationally opaque. The Bunke and Bunke account has about 12 relevant references. More recent work is to be found in

Zwanzig, S. (1980). The choice of appropriate models in non-linear regression. *Math. Operat. Statist. Ser. Statist.*, **11**, 23–47 and in

Bunke, H and Schmidt, W. H. (1980). Asymptotic results on non-linear approximation of regression functions and weighted least-squares. *Math. Operat. Statist. Ser. Statist.*, **11**, 3–22,

which pursue the concept into nonlinearity.

A nice overview is now available in

Linhart, H. and Zucchini, W. (1986) *Model Selection*. Wiley, New York.

See also Linear Model Selection; Regression (Various Entries); and Response Surface Designs.

P.K. Sen

APPROXIMATING INTEGRALS, TEMME'S METHOD

Let $h_n(x)$ be a sequence of functions of a real variable x such that h_n and its derivatives are of order $O(1)$ as $n \to \infty$. (We shall suppress the subscript n in the sequel.) Consider the integral

$$\int_z^\infty h(x)\sqrt{n}\phi(\sqrt{n}x)dx, \qquad (1)$$

where $\phi(\cdot)$ is the standard normal density function. Let $g(x) = [h(x) - h(x)]/x$, $g_1(x) = [g'(x) - g'(0)]/x$, and $g_{j+1}(x) = [g_j'(x) - g_j'(0)]/x$.

Temme [1] approximates the "incomplete" integral (1) by

$$[1 - \Phi(\sqrt{n}z)] \int_{-\infty}^\infty h(x)\sqrt{n}\phi(\sqrt{n}x)dx + \frac{1}{n}\sqrt{n}\phi(\sqrt{n}z)[g(z) + R(z)],$$

where $\Phi(\cdot)$ denotes the standard normal cdf and

$$R(z) = \frac{1}{n}g_1(z) + \frac{1}{n^2}g_2(z) + \ldots.$$

Let $\bar{c}$ be the constant such that

$$\bar{c}\int_\infty^\infty h(x)\sqrt{n}\phi(\sqrt{n}x)dx = 1.$$

Assuming that $\bar{c} = 1 + O(n^{-1})$, one can approximate

$$\int_z^\infty \bar{c}h(x)\sqrt{n}\phi(\sqrt{n}x)dx. \qquad (2)$$

by

$$[1 - \Phi(\sqrt{n}z)] + \frac{1}{n}\sqrt{n}\phi(\sqrt{n}z)\frac{h(z) - h(0)}{z}. \quad (3)$$

Knowledge of the value of $\bar{c}$ is thus not needed for this approximation. The error of (3) as an approximation to (2) is $\phi(\sqrt{n}z)$ $O(n^{-\frac{3}{2}})$.

For fixed z, the *relative* error of the approximation is $O(n^{-1})$.

These formulas provide approximations to tail probabilities, which, unlike those derived via the saddle-point* method, can be integrated analytically.

REFERENCE

1. Temme, N. M. (1982). The uniform asymptotic expansion of a class of integrals related to cumulative distribution functions. *SIAM J. Math. Anal.*, **13**, 239–253.

APPROXIMATIONS TO DISTRIBUTIONS

HISTORICAL DEVELOPMENTS

The main currents of thought in the development of the subject can be traced to the works of Laplace* (1749–1827), Gauss* (1777–1855), and other scientists of a century or so ago. The normal* density played an important role because of its alleged relation to the distribution of errors of measurement, and it is said that Gauss was a strong "normalist" to the extent that departures were due to lack of data. Natural and social phenomena called for study, including the search for stability, causality, and the semblance of order. Quetelet* (1796–1874) exploited the binomial* and normal* distributions, paying particular attention to the former in his investigations of data reflecting numbers of events. Being well versed in mathematics and the natural sciences (he was a contemporary of Poisson*, Laplace, and Fourier), he searched for order in social problems, such as comparative crime rates and human characteristics, posing a dilemma for the current notions of free will. His contributions to social physics, including the concepts for stability and stereotypes, paved the way for modern sociology. Lexis (1837–1914) opened up fresh avenues of research with his sequences of dependent trials, his aim being to explain departures from the binomial exhibited by many demographic* studies.

Urn problems*, involving drawing of balls of different colors, with and without replacement, have played an important part in distributional models. Karl Pearson* [58] interested himself in a typical discrete distribution (the hypergeometric*) originating from urn sampling, and was led to the formulation of his system of distributions by a consideration of the ratio of the slope to ordinate of the frequency polygon*. Pearson's system (see the section "The Pearson System"), including a dozen distinct types, was to play a dominant role in subsequent developments.

The effect of transformation, with natural examples such as the links between distribution of lengths, areas, and volumes, also received attention; for example, Kapteyn [40] discussed the behavior of a nonlinear mapping of the normal.

Which models have proved useful and enlightening? It is impossible to answer this question purely objectively, for utility often depends on available facilities. Statistical development over this century has reacted specifically, first to all applied mathematics, and second, to the influence of the great digital computer invasion. So models exploiting mathematical asymptotics have slowly given way to insights provided by massive numerical soundings.

For example, the Pearson system probability integrals, involving awkward quadratures* for implementation, have recently [56] been tabulated and computerized. Its usefulness has become obvious in recent times. Similarly, the Johnson [35] translation system*, at one time prohibitive in numerical demands, now has extensive tabulated solutions, which, however, present no problem on a small computer.

Questions of validation of solutions are still mostly undecided. As in numerical quadrature, different appropriate formulas on differential grids form the basis for error analysis, and so for distributional approximation, we must use several approaches for comparison; a last resort, depending on the circumstances, would be simulation* studies.

In fitting models by moments it should be kept in mind that even an infinite set of moments may not determine a density uniquely; it turns out that there are strange nonnull functions, mathematical skeletons, for which all moments are zero. This aspect of the problem, usually referred to as the problem of moments, has been discussed by Shohat and Tamarkin [66].

On the other hand, from Chebyshev-type inequalities*, distributions having the same first r moments, cannot be too discrepant in their probability levels. For the four-moment case, the subject has been studied by Simpson and Welch [67]. They consider $\Pr[x < y] = \alpha$ and the problem of bounds for y given α. Also, it should be kept in mind that moments themselves are subject to constraints; thus for central moments, $\mu_4\mu_2 - \mu_3^2 \geqslant \mu_2^3$.

TEST STATISTICS IN DISTRIBUTION

The distribution of test statistics (Student's t*, the standard deviation*, the coefficient of variation*, sample skewness*, and kurtosis*) has proved a problem area and excited interest over many decades. Here we are brought face to face with the fact that normality rarely occurs, and in the interpretation of empirical data conservatism loosened its grip slowly and reluctantly.

Outstanding problems are many. For example, consider statistics which give information over and above that supplied by measures of scale and location. Skewness and kurtosis are simple illustrations, the former being assessed by the third central moment, and the latter by the fourth. The exact distributions of these measures under the null hypothesis, normal universe sampled, are still unknown, although approximations are available (see "Illustrations"). Similarly, Student's t and Fisher's F*, although known distributionally under normality, are in general beyond reach under alternatives. These problems direct attention to approximating distributions by mathematical models.

In passing, we note that whereas large sample assumptions usually slant the desired distribution toward a near neighborhood of normality, small-sample assumptions bring out the amazing intricacies and richness of distributional forms. McKay [44] showed the density of the skewness in samples of four from a normal universe to be a complete elliptic integral; Geary [30] shows the density for samples of five and six to be a spline function*, and remarkably, smoothness of density becomes evident for larger samples. Hoq et al. [33] have derived the exact density of Student's ratio, a noncentral version, for samples of three and four in sampling from an exponential distribution*. Again, spline functions* appear, and it becomes evident that the problem presents insuperable difficulties for larger samples.

In this article we discuss the following systems:

1. K. Pearson's unimodal curves based on a differential equation.
2. Translation systems* based on a mapping of a normal, chi-squared*, or other basic density.
3. Perturbation models based on the normal density, arising out of studies of the central limit theorem*.
4. Multivariate models.
5. Discrete distributions*.

Literature

Johnson and Kotz [37] give an up-to-date comprehensive treatment, the first of its kind, with many references. Discrete, continuous univariate, and multivariate distributions are discussed. A comprehensive account of distributions, with many illuminating examples and exercises, is provided by Kendall and Stuart [42]. Patil and Joshi [53] give a comprehensive summary with most important properties. Bhattacharya and Ranga Rao [4] give a theoretical and mathematical account of approximations related to the central limit theorem*. Some specialized approaches are given in Crain [16] and Dupuy [22].

THE PEARSON SYSTEM

The Model

The model (*see* PEARSON SYSTEM OF DISTRIBUTIONS) is defined by solutions of the differential equation

$$y' = -(x+a)y/(Ax^2+Bx+C), \qquad (1)$$

$y(\cdot)$ being the density. We may take here (but not necessarily in the following discussion) $E(x)=0$, $\mathrm{var}(x)=1$, so that $E(x^3)=\sqrt{\beta_1}$, $E(x^4)=\beta_2$, where the skewness $\beta_1=\mu_3^2/\mu_2^3$ and the kurtosis $\beta_2=\mu_4/\mu_2^2$. Arranging (1) as

$$x^s(Ax^2+Bx+C)y' + x^s(x+a)y = 0 \qquad (2)$$

and integrating, using $s=0,1,2,3$, we find

$$\begin{aligned} A &= (2\beta_2 - 3\beta_1 - 6)/\Delta, \\ B &= \sqrt{\beta_2}(\beta_2+3)/\Delta, \\ C &= (4\beta_2 - 3\beta_1)/\Delta, \qquad a = B, \\ \Delta &= 10\beta_2 - 12\beta_1 - 18. \end{aligned}$$

[If $\Delta=0$, then define $A\Delta=\alpha, B\Delta=\beta$, $C\Delta=\gamma$, so that (1) becomes

$$y' = 2y\sqrt{\beta_1}/(x^2 - 2x\sqrt{\beta_1} - 3),$$

leading to a special case of type 1.]

Note that $\sqrt{\beta_1}$, β_2 uniquely determine a Pearson density. Moreover, it is evident that solutions of (1) depend on the zeros of the quadratic denominator. Although Karl Pearson* identified some dozen types, it is sufficient here, in view of the subsequent usage of the system, to describe the main densities. We quote noncentral (μ_s') or central (μ_s) moments for convenience, and expressions for $\sqrt{\beta_1}$ and β_2.

Normal:

$$y(x) = (2\pi)^{-1/2}\exp(-\tfrac{1}{2}x^2) \qquad (x^2<\infty). \ (3a)$$

$$\mu_1' = 0, \qquad \mu_2 = 1,$$

$$\sqrt{\beta_1} = 0, \qquad \beta_2 = 3.$$

Type 1 (or Beta*):

$$y(x) = \Gamma(a+b)x^{a-1}(1-x)^{b-1}/(\Gamma(a)\Gamma(b))$$
$$(0 < x < 1; a, b > 0). \qquad (3b)$$

$$\mu_1' = a/\alpha_0, \qquad \mu_2 = ab/(\alpha_0^2\alpha_1),$$
$$\mu_3 = 2ab(b-a)/(\alpha_0^3\alpha_1\alpha_2),$$
$$\mu_4 = 3ab(ab(\alpha_0-6)+2\alpha_0^2)/(\alpha_0^4\alpha_1\alpha_2\alpha_3);$$
$$\alpha_s = a+b+s.$$

Gradient at $x = 0$ is finite if $a > 1$.

Gradient at $x = 1$ is finite if $b > 1$.

$$\sqrt{\beta} \geqslant 0 \qquad \text{if } b \geqslant a.$$

Type I may be U-, Ω-, or J-shaped according to the values of a and b.

Type III (Gamma*, Chi-Squared*):

$$y(x) = (x/a)^{\rho-1}\exp(-x/a)/(a\Gamma(\rho))$$
$$(0 < x < \infty; a, \rho > 0). \qquad (3c)$$

$$\mu_1' = a\rho, \qquad \mu_2 = a^2\rho,$$
$$\sqrt{\beta_1} = 2/\sqrt{\rho}, \quad \beta_2 = 3 + 6/\rho.$$

If the density is zero for $x < -s$, then Type III becomes

$$y(x) = ((x+s)/a)^{\rho-1}$$
$$\cdot \exp(-(x+s)/a)/(a\Gamma(\rho))$$
$$(-s < x < \infty),$$

and the only modification in the moments is that $\mu_1' = s + a\rho$. For other types, see Elderton and Johnson [26].

Recurrence for Moments

For the standardized central moments $\nu_s = \mu_s/\sigma^s(\mu_2 = \sigma^2)$, we have from (1),

$$\nu_{s+1} = \frac{s}{D_s}\Big\{(\beta_2+3)\nu_s\sqrt{\beta_1}$$
$$+ (4\beta_2 - 3\beta_1)\nu_{s-1}\Big\}$$
$$(s = 1, 2, \ldots; \nu_0 = 1, \nu_1 = 0), \qquad (4)$$

where $D_s = 6(\beta_2 - \beta_1 - 1) - s(2\beta_2 - 3\beta_1 - 6)$. Note that if $2\beta_2 - 3\beta_1 - 6 < 0$, then $D_s > 0$ since $\beta_2 - \beta_1 - 1 > 0$. Thus in the Type I region of the (β_1, β_2) plane, all moments exist, whereas below the Type III line, $2\beta_2 - 3\beta_1 - 6 > 0$, only a finite number of moments exists; in this case the highest moment ν_s occurs when $s = [x] + 1$, where $x = 6(\beta_2 - \beta_1 - 1)/(2\beta_2 - 3\beta_1 - 6)$ and $[x]$ refers to the integer part of x.

Evaluation of Percentage Points

First, for a given probability level α, we may seek t_α, where

$$\int_{t_\alpha}^{\infty} y(x)\,dx = \alpha \qquad (5)$$

and $y(\cdot)$ is a solution of the Pearson differential equation, with $\alpha = 1$ when $t_\alpha = -\infty$. It is possible to solve (5) by fragmentation, employing tailor-made procedures for each subregion of the $(\sqrt{\beta}_1, \beta_2)$ space. An alternative, encompassing the whole system, is to use (5) and (1) as simultaneous equations for the determination of t_α. Computer programs have been constructed by Amos and Daniel [1] and Bouver and Bargmann [5].

The converse problem of finding α given t_α has also been solved by the two approaches.

This problem of the relation between t_α and α, as one might expect, transparently reflects the fashions and facilities available over its history of a century or so. Computerized numerical analysis*, in modern times, has dwarfed the quadrature* and inverse interpolation* problems involved, and directed attention away from applied mathematical expertise. Nonetheless, there is a rich literature on probability integral problems, much of it relating to various aspects of closest approximation.

The Johnson et al. [39] tables have been available for two decades and give lower and upper points at the percent levels 0.1, 0.25, 0.5, 1.0, 2.5, 5.0, 10.0, 25.0, and 50.0 in terms of (β_1, β_2); the tables are given with additions in the Pearson and Hartley tables [56]. Interpolation is frequently necessary, however. Approximation formulas for standard percent levels 1.0, 5.0, etc., have been given by Bowman and Shenton [8,9].

Illustrations

Pearson curves have been fitted to distributions of the following statistics.

Skewness Statistic. $\sqrt{b_1}$, defined as $m_3/m_2^{3/2}$, where for the random sample $X_1, X_2, \ldots, X_n$, $m_i = \sum(X_j - \overline{X})^i/n$, and $\overline{X}$ is the sample mean. For sampling from the normal, Fisher [28] demonstrated the independence of $\sqrt{b_1}$ and m_2, from which exact moments can be derived. E. S. Pearson [55] gives comparisons of approximations, including Pearson Type VII, for $n = 25(5)40(10)60$ and probability levels $\alpha = 0.05$ and 0.01.

D'Agostino and Tietjen [20] also give comparisons for $n = 7$, 8, 15, 25, 35 at $\alpha = 0.1$, 0.05, 0.025, 0.01, 0.005, and 0.001.

Mulholland [40] in a remarkable study of $\sqrt{b_1}$ has developed approximations to its density for samples $4 \leqslant n \leqslant 25$. The work uses an iterative integral process based on examination of the density discontinuities and to a certain extent follows earlier work of this kind by Geary [30].

When nonnormal sampling is involved, independence property of $\overline{X}$ and S^2 breaks down, and a Taylor series* for $\sqrt{b_1}$ and its powers can be set up leading to an asymptotic series in n^{-1}. Examples of Pearson approximations have been given by Bowman and Shenton [9] and Shenton and Bowman [65].

Kurtosis Statistic b_2. This is defined as m_4/m_2^2, and the independence of $\overline{X}$ and S^2 in normal sampling enables exact moments to be evaluated. Pearson approximations are given in Pearson [54].

Noncentral χ^2. For independent $X_i \in N(0, 1)$ define

$$\chi'^2 = \sum(a_i + X_i)^2.$$

Pearson approximations are given in Pearson and Hartley [56, pp. 10–11, 53–56]; see also Solomon and Stephens [68].

Watson's U_N^{2*}. For approximations to this goodness-of-fit statistic, see Pearson and Hartley [56,77].

Miscellaneous. Bowman et al. [11] have studied Pearson approximations in the case of Student's t under nonnormality.

Four moments of Geary's ratio* of the mean deviation to the standard deviation in normal samples show the density to be near the normal for $n \geqslant 5$; Pearson curves give an excellent fit [7].

Discussion

In approximating a theoretical density by a four-moment Pearson density, we must remember that the general characteristics of the one must be reflected in the other. Bimodel and multimodal densities, for example, should in general be excluded. Also, for $\sqrt{b_1}$ and $n = 4$ under normality, the true distribution consists of back-to-back J-shapes. However, the Pearson curve is ∩-shaped (see McKay [44], and Karl Pearson's remarks). A comparison is given in Table 1.

Since the four-moment fit ignores end points of a statistic (at least they are not fitted in the model), approximations to extreme percentage points will deteriorate sooner or later. Thus, the maximum value of $\sqrt{b_1}$ is known [61] to be $(n - 2)/\sqrt{n - 1}$, so a Type 1 model will either over or under estimate this end point (see Table 1). Similarly, an associated matter concerns tail abruptness; Pearson models usually deteriorate at and near the blunt tail.

Again, exact moments of a statistic may be intractable. Approximations therefore induce further errors, but usually percentage points are not finely tuned to the measures of skewness and kurtosis.

Finally, in general a unimodal density near the normal ($\sqrt{\beta_1}$ small, $\beta_2 = 3$ approximately) should be well approximated by a Pearson density.

Literature

The classical treatise is Elderton's *Frequency Curves and Correlation* [25]. This has now been revised and appears as *Systems of Frequency Curves* [26] by W. P. Elderton and N. L. Johnson (Cambridge University Press). It contains a complete guide to fitting Pearson curves to empirical data, with comments on other approximation systems.

The basic papers by Karl Pearson are those of 1894 [57], 1895 [58], and 1901 [59]. They are of considerable interest in showing the part played by the normal law of errors and modifications of it; for example, Pearson argues that data might be affected by a second normal component to produce the appearance of nonnormality, and solves the two-component normal mixture problem completely. It has since been shown [10]

Table 1. Pearson and Johnson Approximations

Population	Statistic	Sample Size	A[a]	$\alpha = 0.01$	0.05	0.10	0.90	0.95	0.99
Exponential	$\sqrt{m_2}$	$n = 3$[b]	E	0.045	0.109	0.163	1.293	1.610	2.340
			P	0.057	0.113	0.164	1.296	1.612	2.337
			P*	0.368	0.368	0.369	1.290	1.632	2.374
			S_B	0.019	0.102	0.164	1.289	1.609	2.358
		$n = 4$	E	0.095	0.180	0.245	1.342	1.632	2.303
			P	0.108	0.184	0.245	1.343	1.633	2.293
			P*	0.397	0.397	0.399	1.347	1.654	2.322
			S_B	0.083	0.178	0.246	1.340	1.632	2.308
	$\upsilon = (s_x/m'_1)$	$n = 10$	M	0.481	0.587	0.648	1.231	1.351	1.617
			P	0.499	0.591	0.646	1.237	1.353	1.598
			P*	0.531	0.600	0.650	1.238	1.356	1.601
			S_B^*	0.509	0.596	0.650	1.237	1.355	1.604
				$\alpha = 0.90$	0.950	0.975	0.990	0.995	0.999
Normal	$\sqrt{b_1}$	$n = 4$[c]	E	0.831	0.987	1.070	1.120	1.137	1.151
			P	0.793	0.958	1.074	1.178	1.231	1.306
			S_B	0.791	0.955	1.071	1.179	1.237	1.327
		$n = 8$	E	0.765	0.998	1.208	1.452	1.606	1.866
			P	0.767	0.990	1.187	1.421	1.583	1.929
			S_U	0.767	0.990	1.187	1.421	1.583	1.929
				$\alpha = 0.01$	0.05	0.10	0.90	0.95	0.99
	b_2	$n = 10$	M	1.424	1.564	1.681	3.455	3.938	4.996
			P	1.495	1.589	1.675	3.471	3.931	4.921
			S_B	1.442	1.577	1.677	3.456	3.921	4.933
			D	1.39	1.56	1.68	3.53	3.95	5.00
			A_G	1.287	1.508	1.649	3.424	3.855	4.898

[a] E, exact; P, four-moment Pearson on statistic; P*, four-moment Pearson on square of statistic; M, Monte Carlo of 100,000 runs; A_G, Anscombe and Glynn [2]; D, D'Agostino and Tietjen [19]; E for $\sqrt{m_2}$ derived by Lam [43]; S_U, S_B, Johnson; S_B^*, Johnson on square of statistic.

[b] For $n = 3$, P* has density $c_0(m_2 - 0.136)^{\alpha}/(m_2 + 13.168)^{\beta}$, $\alpha = -0.740$, $\beta = 7.770$; P has density $c_1(\sqrt{m_2} - 0.024)^{\alpha}/(\sqrt{m_2} + 13.076)^{\beta}$, $\alpha = 0.738$, $\beta = 39.078$.

[c] For $n = 4$, P is the density $c_2(1.896^2 - b_1)^{3.26}$; from theory $\sqrt{b_1}$ max. is 1.155.

that if exact moments are available (five are needed) then there may be one or two solutions to the problem, or quite possibly none at all. Pearson's examples note these cases.

Again, Pearson from time to time expresses his displeasure at the sloppy tabulation and abbreviation of data. One example he considers concerns the distribution of 8,689 cases of enteric fever received into the Metropolitan Asylums Board Fever Hospitals, 1871–1893. There are 266 cases reported for "under age 5" and 13 for "over age 60." Fitting types I and III to the data by moments, he notes that both models suggest an unlikely age for first onset of the disease (−1.353 and −2.838 years, respectively) and similarly an unlikely upper limit to the duration of life (385 years). (It is quite possible that he was well aware of the small chances involved, and we should keep in mind the point that to arrive at bounds of any kind for the data was something not supplied by conventional normal theory modeling.)

It should be emphasized that a model for a distribution of experimental data encounters problems that are different from those of the corresponding situation for a theoretical statistic. In the latter, errors are due solely to choice of model, although different solution procedures are possible; for example, parameters may be determined by moments, or by percentiles, and also in part from a knowledge of end points. However, in the case of experimental data, parameters are estimated by different procedures (least squares*, maximum likelihood*, moments*, etc.) and there is the question of biases*, variances, etc., and more detailed knowledge of the sampling distribution of the estimates.

Since higher moments of data are subject to considerable sampling errors, a recent [48] method of fitting is recommended when one (or both) end point(s) is known. The procedure calculates a modified kurtosis using the end point and four moments.

TRANSLATION SYSTEMS

Early Ideas

Suppose that the berries of a fruit have radii (measured in some decidable fashion) normally distributed $N(R, \sigma^2)$. Then surface area $S = \pi r^2$ will no longer be normal and indeed will be skewed distributionally. This idea was developed a little later than the introduction of the Pearson system by Kapteyn [40] in his treatment of skew frequency curves in biology. From a different point of view and using a hypothesis relating to elementary errors, Wicksell [71] traced the relation between certain transformation of a variate and a genetic theory of frequency.

Kapteyn considered the transformations

$$y = (X + h)^q - m \qquad [y \in N(0, 1)]$$

with $-\infty < q < \infty$, the special case $q \to 0$ leading to a logarithmic transformation.

The development of the subject was perhaps retarded because of overanxiety to trace a transformation's relation to natural phenomena and mathematical difficulties. Moreover, the Kapteyn mapping proved intractable mathematically in most cases.

Johnson's Systems

Johnson [35] introduced two transformations of the normal density (*see* JOHNSON'S SYSTEM OF DISTRIBUTIONS). His S_U system relates to a hyperbolic sine function, has doubly infinite range, and

$$y = \sinh\left(\frac{X-\gamma}{\delta}\right) \qquad (-\infty < y < \infty, \delta > 0), \tag{6}$$

where $X \in N(0, 1)$.

Similarly, the S_B system relates to densities with bounded range, and

$$y = 1/(1 + e^{(\gamma - X)/\delta}) \qquad (0 < y < 1, \delta > 0), \tag{7}$$

where again $X \in N(0, 1)$.

These transformations are both one-to-one, and the S_U case readily yields to moment evaluation, whereas S_B does not.

For S_U, the first four moments are

$$\begin{aligned}
\mu_1'(y) &= -\sqrt{\omega}\sinh\Omega,\\
\mu_2(y) &= (\omega - 1)(\omega\cosh(2\Omega) + 1)/2,\\
\mu_3(y) &= -(\omega - 1)^2\sqrt{\omega}\{(\omega^2 + 2\omega)\sinh(3\Omega)\\
&\qquad + 3\sinh\Omega\}/4, \qquad (8)\\
\mu_4(y) &= (\omega - 1)^2\{d_4\cosh(4\Omega)\\
&\qquad + d_2\cosh(2\Omega) + d_0\},
\end{aligned}$$

where

$$\begin{aligned}
d_4 &= \omega^2(\omega^4 + 2\omega^3 + 3\omega^2 - 3)/8,\\
d_2 &= \tfrac{1}{2}\omega^2(\omega + 2),\\
d_0 &= 3(2\omega + 1)/8, \qquad \text{and}\\
\ln\omega &= 1/\delta^2, \quad \Omega = \gamma/\delta.
\end{aligned}$$

Note that since $\omega > 1$, $\mu_3(y)$ has the sign of $(-\Omega)$, and unlike the structure of the Pearson system, here all moments are functions of ω and Ω. (For the Pearson system, mean and variance are not affected by $\sqrt{\beta_1}$ and β_2, whereas they are for S_U.)

If Y is a variate, then set

$$Y = (y + p)/q \tag{9a}$$

so that

$$\begin{aligned}
\nu_1' &= E(Y) = (\mu_1'(y) + p)/q\\
\nu_2 &= \mathrm{var}(Y) = (\mu_2(y))/q^2
\end{aligned} \tag{9b}$$

Determine p and q. The values of $\mu_1'(y)$ and $\mu_2(y)$ are set by equating the skewness $(\sqrt{\beta_1})$, and kurtosis (β_2) of Y to those of y; assistance here is given in Tables 34 and 35 in Pearson and Hartley [56].

Johnson's S_U system [35] immediately provides percentiles from those of the normal, and also an equivalent normal variate,

$$X = \gamma + \delta\sinh^{-1}y.$$

However, one must keep in mind that the S_U density is still only a four-moment approximation.

To fix the domain of validity of S_U, let $\Omega = -k$ and $k \to \infty$, so that the mean of y tends to ∞ and $\sigma^2 \sim \omega(\omega-1)e^{2k}/4$. Then if $t = y/\sigma$ from (6), $X = c + \ln t$, which corresponds to a log-normal transformation, and from (8)

$$\begin{aligned}\sqrt{\beta_1} &= (\omega+2)\sqrt{\omega-1},\\ \beta_2 &= \omega^4 + 2\omega^3 + 3\omega^2 - 3, \qquad (10)\end{aligned}$$

the parametric form of the boundary in the $(\sqrt{\beta_1}, \beta_2)$ plane.

Examples of S_U and S_B are given in Table 1.

Literature

An adequate description of S_U and S_B is given by Pearson and Hartley [56, pp. 80–87], including tables to facilitate fitting. Moreover, the iterative scheme for the evaluation of the parameters of S_U given by Johnson [36] is readily programmed for a small computer. A rational fraction solution for ω, leading also to a value of Ω, has been developed by Bowman and Shenton [10]; a similar scheme is also available for the S_B-system [13].

SERIES DEVELOPMENTS

These originated in the nineteenth century, and are related to procedures to sharpen the central limit theorem*. For large n, for $z_1, z_2, \ldots, z_n$ mutually independent variates with common standard deviation σ, the distribution of $s = (\sum z)/(\sigma\sqrt{n})$ is approximately normal. A better approximation appears from Charlier's A-series [14],

$$\phi(x) + (a_3/3!)\phi^{(3)}(x) + (a_4/4!)\phi^{(4)}(x) + \cdots \qquad (11)$$

where $\phi(x) = (2\pi)^{-1/2}\exp\left(-\frac{1}{2}x^2\right)$ is the standard normal density. [This approximation involves derivatives of $\phi(x)$.] Cramér [17,18] has proved that certain asymptotic properties of (11) hold. Another version of the series development (11) takes the form $\Phi(d/dx)\phi(x)$, where

$$\Phi(t) \equiv \exp \sum \epsilon_j(-t)^j/j!, \qquad (12)$$

the operator changing the cumulants* (κ_r) of $\phi(\cdot)$ to $(\kappa_r + \epsilon_r)$. This, from Cramér's work (see his note and references in ref. 18), has similar asymptotic properties with respect to the central limit theorem*.

Since the derivatives of the normal density are related to Hermite polynomials* (*see* CHEBYSHEV–HERMITE POLYNOMIALS), (11) may be written

$$[1 - (a_3/3!)H_3(x) + (a_4/4!)H_4(x) - \cdots]\phi(x), \qquad (13)$$

and if this approximates a density $f(x)$, then using orthogonality,

$$a_s = (-1)^s E[H_s(x)],$$

where the expectation* operator refers to $f(x)$. Thus in terms of cumulants,

$$\begin{aligned}a_3 &= -\kappa_3, \quad a_4 = \kappa_4,\\ a_5 &= -\kappa_5, \quad a_6 = \kappa_6 + 10\kappa_3^2,\end{aligned}$$

etc., the coefficients a_1, a_2 being zero because of the use of the standard variate x.

If empirical data are being considered, only the first five or six terms of (1) can be contemplated because of large sampling errors of the moments involved. However, this problem does not arise in the approximation of theoretical structures, and cases are on record (for example, ref. 47) in which the twentieth polynomial has been included; quite frequently the density terms turn out to have irregular sign and magnitude patterns.

The Edgeworth* form [23,24],

$$\begin{aligned}\phi(x)\{1 &+ (\kappa_3/3!)H_3(x) + (\kappa_4/4!)H_4(x)\\ &+ (\kappa_5/5!)H_5(x)\\ &+ (\kappa_6 + 10\kappa_3^2)H_6(x)/6! + \cdots\}\end{aligned}$$

has one advantage over (1), in that when applied to certain statistics, the standardized cumulants κ_r may be shown to be of order $1/n^{(1/2)r-1}$, where n is the sample size, so that if carried far enough the coefficients will tend to exhibit a regular magnitude pattern.

Nonnormal Kernels

The basic function $\phi(\cdot)$ need not be normal, and Romanowsky [62] introduced generalizations such as gamma- and beta*-type densities, with associated Laguerre* and Jacobi* orthogonal polynomials. The normal density is, however, generally favored because of the simplicity of the Chebyshev-Hermite system of polynomials.

Cornish–Fisher Expansions

Using a differential series, such as (12) with normal kernel, Cornish and Fisher [15] derived a series for the probability integral of a variate whose cumulants are known (*see* CORNISH–FISHER AND EDGEWORTH EXPANSIONS). By inversion of series, they derived a series for the deviate at a given probability level in terms of polynomials in the corresponding normal deviate. Fisher and Cornish [29] extended the earlier study to include terms of order n^{-3}, these involving the eighth cumulant and polynomials of degree seven. Further generalizations are due to Finney [27], who treated the case of several variates; Hill and Davis [32], who gave a rigorous treatment indicating the procedure for the derivation of a general term; and Draper and Tierney [21], who tabulated the basic polynomials involved in the terms to order n^{-4} and cumulants up to κ_{10}; unpublished results of Hill and Davis for the higher-order polynomials agreed with those found by Draper and Tierney.

Some Applications

In his paper on testing for normality* (*see* DEPARTURES FROM NORMALITY, TESTS FOR), Geary [31] considered the distribution of t under nonnormality, basing it on what he called a differential series [expression (12)] with kernel the density of t^* under normality, i.e., $c_n(1 + t^2/(n-1))^{-n/2}$. He derived a series to order n^{-2} for the probability integral, using it cautiously for a sample of 10 under only moderate nonnormality. In an earlier study [30] searching for precise forms for the density of the sample skewness under normality, he used the Cornish-Fisher series to assess probability levels using cumulants up to the eighth. Mulholland [47] developed the subject further, providing a recursive scheme for evaluating at least theoretically any moment of $\sqrt{b_1}$, and a Charlier series for the probability integral up to polynomials of degree 20.

Series involving Laguerre polynomials and a gamma density have been exploited by Tiku. For example, his study [69] of the variance ratio* and Student's t [70], both under nonnormality, involve terms up to order n^{-2}.

Discussion

In approximating distributions, convergence* questions, and even to a less extent, asymptotic properties are irrelevant, for we are concerned with a finite number of terms, since very rarely can general terms be found. Thus questions of nonnegativity of the density arise [3,64], and internal checks for closeness of approximation via convergence or otherwise are not available. At best the Charlier and Edgeworth series can only serve as backup models.

MULTIVARIATE DENSITIES

The Pearson system defined in (1) becomes, in bivariate form,

$$\frac{1}{y}\frac{\partial y}{\partial x_i} = \frac{P_i(x_1, x_2)}{Q_i(x_1, x_2)} \qquad (i = 1, 2),$$

where P_i and Q_i are functions of x_1 and x_2 of degrees 1 and 2, respectively. There are again several types, including the bivariate normal* with density

$$\phi(x_1, x_2) = (1-\rho^2)^{-1/2}(2\pi)^{-1} \times \exp\left\{\frac{-\frac{1}{2}x^2 - \rho xy - \frac{1}{2}y^2}{1-\rho^2}\right\}, \rho^2 < 1,$$

in standard form, which reduces to a product form when $\rho = 0$ (i.e., when the variates are uncorrelated).

Another well-known type is the Dirichlet* density,

$$y(x_1, x_2) = cx_1^{a-1}x_2^{b-1}(1 - x_1 - x_2)^{c-1} \quad (a, b, c > 0)$$

with domain $x_1, x_2 > 0, x_1 + x_2 \leqslant 1$.

Series development based on the Charlier model take the form

$$\{1 + a_{10}\partial_1 + a_{01}\partial_2 + (a_{20}/2!)\partial_1^2 + a_{11}\partial_1\partial_2 + (a_{02}/2!)\partial_2^2 + \cdots\}\phi(x_1, x_2),$$

where $\partial_i \equiv \partial/\partial x_i$.

Similarly, there are multivariate translation systems, following the Johnson $\mathrm{S_U}$ and $\mathrm{S_B}$ models.

The Pearson system, for which marginal distributions are of Pearson form, can be fitted by moments, a new feature being the necessity to use product moments*, the simplest being related to the correlation between the variates. A similar situation holds for Charlier multivariate developments, and again, as with the single-variate case, negative frequencies can be a problem.

Literature

A sustained study of empirical bivariate data, at least of historical interest, is given in K. Pearson's 1925 paper [60].

A brief discussion of frequency surfaces* will be found in Elderton and Johnson [26]. A more comprehensive treatment, including modern developments, is that of N. L. Johnson and S. Kotz [37]. Mardia's work [45] is a handy reference.

DISCRETE DISTRIBUTIONS

There is now a richness of variety of discrete distributions, and the time is long passed when there were possibly only three or so choices, the binomial*, the Poisson*, and the hypergeometric*; the geometric appearing infrequently in applications.

Just as a differential equation is used to define the Pearson system of curves, so analogs in finite differences* may be used to generate discrete density functions [41,49, 58]. Again, if $G_i(t)$, $i = 1, 2, \ldots$, are the probability generating functions* of discrete variates, then new distributions arise from G_1 $(G_2(t))$ and similar structures; for example, Neyman's contagious distributions* are related to choosing G_1 and G_2 as Poisson generating functions.

There are many applications where distribution approximation models are required for random count data* [51,52]. We also mention occupancy problems* [38], meteorological phenomena relating to drought frequency, storm frequency, degree-days, etc. However, the need for approximating discrete distributions is not common, especially with the advent of computer facilities. A classical exception is the normal approximation to the binomial distribution when the index n is large; in this case, if the random variate is x, then with probability parameter p, we consider the approximation $(x - np)/\sqrt{npq}$ to be nearly normal (for refinements, see, e.g., Molenaar [46]).

Again the Poisson or binomial density functions may be used as the basis for Charlier-type approximation expansions. Thus for the Poisson function $\psi(x) = e^{-m}m^x/x!$, we consider

$$f(x) = \{1 + \alpha_2\nabla_x^2/2! + \alpha_3\nabla_x^3/3!t + \cdots\}\psi(x),$$

where $\nabla g(x) \equiv g(x) - g(x-1)$ is a backward difference*. This type of approximation may be considered for random variates taking the values $x = 0, 1, \ldots$, when moments exist and the density to be approximated is complicated.

Literature

Historical information has been given by Särndal [63], and comprehensive accounts are those of Johnson and Kotz [37] and Patil and Joshi [53]. A short account, with new material, is given by J. K. Ord [50].

Acknowledgment

This research was operated by Union Carbide Corporation under Contract W-7405-eng-26 with the U.S. Department of Energy.

REFERENCES

1. Amos, D. E. and Daniel, S. L. (1971). *Tables of Percentage Points of Standardized Pearson Distributions. Rep. No. SC-RR-71 0348*, Sandia Laboratories, Albuquerque, N.M.
2. Anscombe, F. and Glynn, W. J. (1975). *Distribution of the Kurtosis Statistics b_2 for Normal Samples. Tech. Rep. 37*, Dept. of Statistics, Yale University, New Haven, Conn.

3. Barton, D. E. and Dennis, K. E. (1952). *Biometrika*, **39**, 425–427.
4. Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal Approximations and Asymptotic Expansions*. Wiley, New York.
5. Bouver, H. and Bargmann, R. (1976). *Amer. Stat. Ass.: Proc. Statist. Computing Sect.*, pp. 116–120.
6. Bowman, K. O., Beauchamp, J. J., and Shenton, L. R. (1977). *Int. Statist. Rev.*, **45**, 233–242.
7. Bowman, K. O., Lam, H. K., and Shenton, L. R. (1980). *Reports of Statistical Application Res., Un. Japanese Scientists and Engineers*, **27**, 1–15.
8. Bowman, K. O., Serbin, C. A., and Shenton, L. R. (1981). *Commun. Statist. B*, **10**(1), 1–15.
9. Bowman, K. O. and Shenton, L. R. (1973). *Biometrika*, **60**, 155–167.
10. Bowman, K. O. and Shenton, L. R. (1973). *Biometrika*, **60**, 629–636.
11. Bowman, K. O. and Shenton, L. R. (1979). *Biometrika*, **66**, 147–151.
12. Bowman, K. O. and Shenton, L. R. (1979). *Commun. Statist. B*, **8**(3), 231–244.
13. Bowman, K. O. and Shenton, L. R. (1980). *Commun. Statist. B*, **9**(2), 127–132.
14. Charlier, C. V. L. (1905). *Ark. Mat. Astron. Fys.*, **2**(15).
15. Cornish, E. A. and Fisher, R. A. (1937). *Rev. Inst. Int. Statist.*, **4**, 1–14.
16. Crain, B. R. (1977). *Siam J. Appl. Math.*, **32**, 339–346.
17. Cramér, H. (1928). *Skand. Aktuarietidskr*, **11**, 13–74, 141180.
18. Cramér, H. (1972). *Biometrika*, **59**, 205–207.
19. D'Agostino, R. B. and Tietjen, G. L. (1971). *Biometrika*, **58**, 669–672.
20. D'Agostino, R. B. and Tietjen, G. L. (1973). *Biometrika*, **60**, 169–173.
21. Draper, N. R. and Tierney, D. E. (1973). *Commun. Statist.*, **1**(6), 495–524.
22. Dupuy, M. (1974). *Int. J. Computer Math. B*, **4**, 121–142.
23. Edgeworth, F. Y. (1905). *Camb. Philos. Trans.*, **20**, 36–66, 113141.
24. Edgeworth, F. Y. (1907). *J. R. Statist. Soc.*, **70**, 102–106.
25. Elderton, W. P. (1960). *Frequency Curves and Correlation*. Cambridge University Press, Cambridge.
26. Elderton, W. P. and Johnson, N. L. (1969). *Systems of Frequency Curves*. Cambridge University Press, Cambridge.
27. Finney, D. J. (1963). *Technometrics*, **5**, 63–69.
28. Fisher, R. A. (1928). *Proc. Lond. Math. Soc.*, **2**, 30, 199–238.
29. Fisher, R. A. and Cornish, E. A. (1960). *Technometrics*, **2**, 209–225.
30. Geary, R. C. (1947). *Biometrika*, **34**, 68–97.
31. Geary, R. C. (1947). *Biometrika*, **34**, 209–242.
32. Hill, G. W. and Davis, A. W. (1968). *Ann. Math. Statist.*, **39**, 1264–1273.
33. Hoq, A. K. M. S., Ali, M. M., and Templeton, J. G. (1977). Distribution of Student's Ratio Based on the Exponential Distribution. *Working Paper No. 77-001*, Dept. of Industrial Engineering, University of Toronto, Toronto.
34. Hotelling, H. (1961). *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, Vol. 1. University of California Press, Berkeley, Calif., pp. 319–359.
35. Johnson, N. L. (1949). *Biometrika*, **36**, 149–176.
36. Johnson, N. L. (1965). *Biometrika*, **52**, 547–558.
37. Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.
38. Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Application*. Wiley, New York.
39. Johnson, N. L., Nixon, E., Amos, D. E., and Pearson, E. S. (1963). *Biometrika*, **50**, 459–498.
40. Kapteyn, J. C. (1903). *Skew Frequency Curves in Biology and Statistics*. Noordhoff, Groningen.
41. Katz, L. (1963). *Proc. Int. Symp. Discrete Distrib.*, Montreal, pp. 175–182.
42. Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics*, 3rd ed., Vol. 1: Distribution Theory. Charles Griffin, London/Hafner Press, New York.
43. Lam, H. K. (1978). The Distribution of the Standard Deviation and Student's *t* from Nonnormal Universes. Ph.D. dissertation, University of Georgia.
44. McKay, A. T. (1933). *Biometrika*, **25**, 204–210.
45. Mardia, K. V. (1970). Families of Bivariate Distributions. *Griffin's Statist. Monogr. No. 20*.
46. Molenaar, W. (1970). Approximation to the Poisson, Binomial, and Hypergeometric Distribution Functions. *Math. Centre Tracts No. 31*. Mathematisch Centrum, Amsterdam.

47. Mulholland, H. P. (1977). *Biometrika*, **64**, 401–409.
48. Müller, P.-H. and Vahl, H. (1976). *Biometrika*, **63**, 191–194.
49. Ord, J. K. (1967). *J. R. Statist. Soc. A*, **130**, 232–238.
50. Ord, J. K. (1975). *Families of Frequency Distributions*. *Griffin's Statist. Monogr. No. 30*.
51. Patil, G. P. (1965). *Classical and Contagious Discrete Distributions* (Proc. Int. Symp., Montreal). Pergamon Press, New York.
52. Patil, G. P. ed. (1970). *Random Counts in Physical Science, Geological Science, and Business*, Vols. 1–3. The Penn State Statistics Series. Pennsylvania State University Press, University Park, Pa.
53. Patil, G. P. and Joshi, S. W. (1968). *A Dictionary and Bibliography of Discrete Distributions*. Oliver & Boyd, Edinburgh.
54. Pearson, E. S. (1963). *Biometrika*, **50**, 95–111.
55. Pearson, E. S. (1965). *Biometrika*, **52**, 282–285.
56. Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press, Cambridge.
57. Pearson, K. (1894). *Philos. Trans. R. Soc. Lond. A*, **185**, 71–110.
58. Pearson, K. (1895). *Philos. Trans. R. Soc. Lond. A*, **186**, 343–414.
59. Pearson, K. (1901). *Philos. Trans. R. Soc. Lond. A*, **197**, 443–459.
60. Pearson, K. (1925). *Biometrika*, **17**, 268–313.
61. Pearson, K. (1933). *Biometrika*, **25**, 210–213.
62. Romanowsky, V. (1925). *Biometrika*, **16**, 106–116.
63. Särndal, C.-E. (1971). *Biometrika*, **58**, 375–391.
64. Shenton, L. R. (1951). *Biometrika*, **38**, 58–73.
65. Shenton, L. R. and Bowman, K. O. (1975). *J. Amer. Statist. Ass.*, **70**, 220–229, 349.
66. Shohat, J. A. and Tamarkin, J. E. (1943). *The Problem of Moments*. American Mathematical Society, New York.
67. Simpson, J. A. and Welch, B. L. (1960). *Biometrika*, **47**, 399–410.
68. Solomon, H. and Stephens, M. A. (1978). *J. Amer. Statist. Ass.*, **73**, 153–160.
69. Tiku, M. L. (1964). *Biometrika*, **51**, 83–95.
70. Tiku, M. L. (1971). *Aust. J. Statist.*, **13**(3), 142–148.
71. Wicksell, S. D. (1917). *Ark. Mat. Astron. Fys.*, **12**(20).

See also ASYMPTOTIC EXPANSIONS—I; CORNISH–FISHER AND EDGEWORTH EXPANSIONS; FREQUENCY CURVES, SYSTEMS OF; FREQUENCY SURFACES, SYSTEMS OF; GRAM–CHARLIER SERIES; JOHNSON'S SYSTEM OF DISTRIBUTIONS; and PEARSON SYSTEM OF DISTRIBUTIONS.

K. O. BOWMAN
L. R. SHENTON

APPROXIMATIONS TO FUNCTIONS.

See FUNCTIONS, APPROXIMATIONS TO

A PRIORI DISTRIBUTION

The term *a priori distribution* is used to describe a distribution ascribed to a parameter in a model. It occurs most commonly in the application of Bayesian methods. The a priori distribution is usually supposed to be known exactly—and not to depend on unknown parameters of its own.

See also BAYESIAN INFERENCE.

ARBITRARY ORIGIN.

See CODED DATA

ARBITRARY SCALE.

See CODED DATA

ARBUTHNOT, JOHN

Born: April 29, 1667, in Arbuthnott, Kincardineshire, Scotland.

Died: February 27, 1735, in London, England.

Contributed to: early applications of probability, foundations of statistical inference, eighteenth-century political satire, maintaining the health of Queen Anne.

John Arbuthnot was the eldest son of Alexander, parson of the village of Arbuthnott, near the east coast of Scotland. They were part of the Aberdeenshire branch of the Arbuthnot family, whose precise relationship to the Viscounts Arbuthnott of Kincardineshire, and to the ancient lairds of the family, is not known. In later years, John invariably signed his name "Arbuthnott," yet "Arbuthnot" consistently appears on his printed works. The latter is evidently the more ancient form, to which the Kincardineshire branch added a second "t" some time in the seventeenth century.

John entered Marischal College, Aberdeen, at the age of fourteen, and graduated with an M.A. in medicine in 1685. Following the Glorious Revolution of 1688–1689, John's father was deprived of his living in 1689, because he would not accept Presbyterianism. He died two years later, and soon after that John left Scotland to settle in London, where initially he made his living teaching mathematics. It was during this period that he produced one of the two pieces of work that have merited him a place in the history of probability and statistics. He translated, from Latin into English, Huygens' *De Ratiociniis in Ludo Aleae*, the first probability text [7]. Arbuthnot's English edition of 1692, *Of the Laws of Chance* [3], was not simply a translation. He began with an introduction written in his usual witty and robust style, gave solutions to problems Huygens* had posed, and added further sections of his own about gaming with dice and cards.

In 1694, he was enrolled at University College, Oxford as a "fellow commoner," acting as companion to Edward Jeffreys, the eldest son of the Member of Parliament for Brecon. Arbuthnot developed lasting friendships with Arthur Charlett, the Master of University College, and with David Gregory, the Savilian Professor of Astronomy at Oxford. It seems, though, that Edward Jeffreys did not use his time at Oxford well, and in 1696, Arbuthnot became "resolv'd on some other course of life." He moved briefly to St. Andrews and there presented theses which earned him a doctor's degree in medicine in September 1696.

Having returned to London, he quickly earned himself a reputation as a skilled physician and a man of learning. Two of his early scientific publications were *An Examination of Dr. Woodward's Account of the Deluge* (1697; John Woodward was Professor of Physic at Gresham College) [3] and *An Essay on the Usefulness of Mathematical Learning* (1701) [1]. He became a Fellow of both the Royal Society (1704) and the Royal College of Physicians (1710), and served on two prestigious committees of the former. The first, which included Sir Isaac Newton* (then President of the Royal Society), Sir Christopher Wren, and David Gregory, was set up in 1705 to oversee the publication of the astronomical observations of the Astronomer Royal, John Flamsteed. Arbuthnot was directly involved in trying to secure Flamsteed's cooperation in carrying the venture through to completion, in the face of hostility between Flamsteed and Newton, and accusations on each side that the other was obstructing the project. After long delays, the observations were finally published in 1712. Arbuthnot was less active in the second of the committees, appointed in 1712 to deliberate on the rival claims of Newton and Leibniz to invention of the "method of fluxions" (differential calculus).

In 1705, Arbuthnot became Physician Extraordinary to Queen Anne, and in 1709 Physician in Ordinary. Until Queen Anne's death in 1714, he enjoyed a favored position at the Court. In 1711, he held office in the Customs, and in 1713 he was appointed Physician at Chelsea College.

Outside the statistical community, he is most widely remembered for his comic satirical writings. He was the creator of the figure John Bull, who has become a symbol of the English character. John Bull featured in a series of pamphlets written by Arbuthnot in 1713; they contain a witty allegorical account of the negotiations taking place towards the settlement of the War of the Spanish Succession and were, in due course, put together as *The History of John Bull* [1]. Arbuthnot was a close colleague of Jonathan Swift, Alexander Pope, and John Gay, and a founder member of the Scriblerus Club, an association of scholars formed around 1714 with the aim of ridiculing pedantry and poor scholarship. It is thought that much of the *Memoirs of the Extraordinary Life, Works, and Discoveries of Martinus Scriblerus* [1] came from Arbuthnot's pen. His writings ranged widely over aspects of science, mathematics, medicine, politics and philosophy. Often he was modest or careless in claiming authorship, particularly of his political satire.

The paper he presented to the Royal Society of London on April 19, 1711 [2] has attracted most attention from historians of statistics and probability (it appeared in a volume of the *Philosophical Transactions* dated 1710, but published late). Arbuthnot's paper was "An Argument for Divine Providence, taken from the constant Regularity observ'd in the Births of both Sexes." In it, he maintained that the guiding hand of a divine being was to be discerned in the nearly constant ratio of male to female christenings recorded annually in London over the years 1629 to 1710. Part of his reasoning is recognizable as what we would now call a sign test, so Arbuthnot has gone down in statistical history as a progenitor of significance testing*.

The data he presented showed that in each of the 82 years 1629–1710, the annual number of male christenings had been consistently higher than the number of female christenings, but never very much higher. Arbuthnot argued that this remarkable regularity could not be attributed to chance*, and must therefore be an indication of divine providence. It was an example of the "argument from design," a thesis of considerable theological and scientific influence during the closing decades of the seventeenth century and much of the next. Its supporters held that natural phenomena of many kinds showed evidence of careful and beneficent design, and were therefore indicative of the existence of a supreme being.

Arbuthnot's representation of "chance" determination of sex at birth was the toss of a fair two-sided die, with one face marked M and the other marked F. From there, he argued on two fronts: "chance" could not explain the very close limits within which the annual ratios of male to female christenings had been observed to fall, neither could it explain the numerical dominance, year after year, of male over female christenings.

He pursued the first argument by indicating how the middle term of the binomial expansion, for even values of the size parameter n, becomes very small as n gets large. Though he acknowledged that in practice the balance between male and female births in any one year was not exact, he regarded his mathematical demonstration as evidence that, "if mere Chance govern'd," there would be years when the balance was not well maintained.

The second strand of his argument, concerning the persistent yearly excess of male over female christenings, was the one which ultimately caught the attention of historians of statistics. He calculated that the probability of 82 consecutive years in which male exceeded female christenings in number, under the supposition that "chance" determined sex, was very small indeed. This he took as weighty evidence against the hypothesis of chance, and in favor of his alternative of divine providence. He argued that if births were generated according to his representation of chance, as a fair two-sided die, the probability of observing an excess of male over female births in any one year would be no higher than one-half. Therefore the probability of observing 82 successive "male years" was no higher than $(\frac{1}{2})^{82}$ (a number of the order of 10^{-25} or 10^{-26}). The probability of observing the data given the "model," as we might now say, was very small indeed, casting severe doubt on the notion that chance determined sex at birth. Arbuthnot proceeded to a number of conclusions of a religious or philosophical nature, including the observation that his arguments vindicated the undesirability of polygamy in a civilized society.

We can see in Arbuthnot's probabilistic reasoning some of the features of the modern hypothesis test. He defined a null hypothesis ("chance" determination of sex at birth) and an alternative (divine providence). He calculated, under the assumption that the null hypothesis was true, a probability defined by reference to the observed data. Finally, he argued that the extremely low probability he obtained cast doubt on the null hypothesis and offered support for his alternative.

Arbuthnot's reasoning has been thoroughly examined by modern statisticians and logicians, most notably by Hacking [5, 6]. We have, of course, the benefit of more than 250 years of hindsight and statistical development. The probability of $(\frac{1}{2})^{82}$, on which hinged Arbuthnot's dismissal of the "chance" hypothesis, was one of a well-defined reference set, the binomial distribution with parameters 82 and one-half. It was the lowest and most extreme probability in this reference set, and hence also in effect a tail-area probability. And it was an *extremely low* probability. Arbuthnot made only the last of these points explicit.

Arbuthnot's advancement of an argument from design did not single him out from his contemporaries. Nor were his observations on the relative constancy of the male to female birth ratio radical. What was novel was his attempt to provide a statistical "proof" of his assertions, based on a quantitative concept of chance, explicitly expressed and concluded in numerical terms.

An unpublished manuscript in the Gregory collection at the University of Edinburgh indicates that Arbuthnot had been flirting with ideas of probabilistic proof well before 1711, possibly as early as 1694 [4]. In his 10-page "treatise on chance" is an anticipation of his 1711 argument concerning the middle term of the binomial as n gets large, as well as two other statistical "proto-tests" concerning the lengths of reign of the Roman and Scottish kings. The chronology of the first seven kings of Rome was suspect, he suggested, because they appeared to have survived far longer on average than might reasonably be expected from Edmund Halley's life table, based on the mortality bills of Breslau. In the case of the Scottish kings, on the other hand, the evidence seemed to indicate that mortality amongst them was higher than might be expected from Halley's table. However, neither of the calculations Arbuthnot outlined had the clarity of statistical modeling evident in his 1711 paper, nor did they culminate in a specific probability level quantifying the evidence.

Arbuthnot's 1711 paper sparked off a debate which involved, at various times, William 'sGravesande* (a Dutch scientist who later became Professor of Mathematics, Astronomy and Philosophy at the University of Leiden), Bernard Nieuwentijt (a Dutch physician and mathematician), Nicholas Bernoulli*, and Abraham de Moivre*. 'sGravesande developed Arbuthnot's test further, attempting to take into account the close limits within which the male-to-female birth ratio fell year after year. Bernoulli, on the other hand, questioned Arbuthnot's interpretation of "chance." He proposed that the fair two-sided die could be replaced by a multifaceted die, with 18 sides marked M and 17 marked F. If tossed a large number of times, Bernoulli maintained, such a die would yield ratios of M's to F's with similar variability to the London christenings data. Certain aspects of the exchanges between the participants in the debate can be seen as attempts to emulate and develop Arbuthnot's mode of statistical reasoning, but have not proved as amenable to reinterpretation within modern frameworks of statistical logic.

Though Arbuthnot's 1711 argument tends now to be regarded as the first recognizable statistical significance test, it is doubtful whether his contribution, and the debate it provoked, provided any immediate stimulus to ideas of statistical significance testing. The obvious impact was to fuel interest in the "argument from design," in the stability of statistical ratios, and in the interplay of one with the other.

An oil painting of Arbuthnot hangs in the Scottish National Portrait Gallery, Edinburgh. By all accounts, he was a charitable and benevolent man. In a letter to Pope, Swift said of him: "Our doctor hath every quality in the world that can make a man amiable and useful; but alas! he hath a sort of slouch in his walk."

REFERENCES

1. Aitken, G. A. (1892). *The Life and Works of John Arbuthnot*. Clarendon Press, Oxford.
2. Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Phil. R. Soc. London, Trans.*, 27, 186–190. Reprinted in (1977). *Studies in the History of Statistics and Probability*, M. G. Kendall and R. L. Plackett, eds., Griffin, London, **vol. 2**, pp. 30–34.
3. Arbuthnot, J. (1751). *Miscellaneous works of the late Dr. Arbuthnot, with Supplement*. Glasgow. [Arbuthnot's *Of the Laws of Chance* (1692) was included as *Huygens' de Ratiociniis in Ludo Aleae*: translated into English by Dr. Arbuthnot.]
4. Bellhouse, D. R. (1989). A manuscript on chance written by John Arbuthnot. *Int. Statist. Rev.*, **57**, 249–259.

5. Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, pp. 75–81.

6. Hacking, I. (1975). *The Emergence of Probability*. Cambridge University Press, pp. 166–171.

7. Huygens, C. (1657). De ratiociniis in ludo aleae. In *Exercitationum mathematicarum libri quinque*, F van Schooten, ed. Amsterdam.

BIBLIOGRAPHY

Beattie, L. M. (1935). *John Arbuthnot: Mathematician and Satirist.* Harvard Studies in English, vol. XVI. Harvard University Press, Cambridge, Massachusetts.

Hald, A. (1990). *A History of Probability and Statistics and their Applications before 1750*. Wiley, New York.

Shoesmith, E. (1987). The continental controversy over Arbuthnot's argument for divine providence. *Historia Math.* **14**, 133–146.

Stephen, L. and Lee, S. (eds.) (1917). John Arbuthnot. In *Dictionary of National Biography*, vol. I. Oxford University Press.

E. SHOESMITH

ARCH AND GARCH MODELS

Many time series* display time-varying dispersion, or uncertainty, in the sense that large (small) absolute innovations* tend to be followed by other large (small) absolute innovations. A natural way to model this phenomenon is to allow the variance to change through time in response to current developments of the system. Specifically, let $\{y_t\}$ denote the observable univariate discrete-time stochastic process of interest. Denote the corresponding innovation process by $\{\epsilon_t\}$, where $\epsilon_t \equiv y_t - E_{t-1}(y_t), \text{and} E_{t-1}(\cdot)$ refers to the expectation conditional on time-$(t-1)$ information. A general specification for the innovation process that takes account of the time-varying uncertainty would then be given by

$$\epsilon_t = z_t \sigma_t, \tag{1}$$

where $\{z_t\}$ is an i.i.d. mean-zero, unit-variance stochastic process*, and σ_t represents the time-t latent volatility; i.e., $E(\epsilon_t^2 | \sigma_t) = \sigma_t^2$. Model specifications in which σ_t in (1) depends non-trivially on the past innovations and/or some other latent variables are referred to as *stochastic volatility* (SV) models. The historically first, and often most convenient, SV representations are the *autoregressive conditionally heteroscedastic* (ARCH) models pioneered by Engle [21]. Formally the ARCH class of models are defined by (1), with the additional restriction that σ_t must be measurable with respect to the time-$(t-1)$ observable information set. Thus, in the ARCH class of models $\text{var}_{t-1}(y_t) \equiv E_{t-1}(\epsilon_t^2) = \sigma_t^2$ is predetermined as of time $t-1$.

VOLATILITY CLUSTERING

The ARCH model was originally introduced for modeling inflationary uncertainty, but has subsequently found especially wide use in the analysis of financial time series. To illustrate, consider the plots in Figs. 1 and 2 for the daily Deutsche-mark—U.S. dollar (DM/$) exchange rate and the Standard and Poor's 500 composite stock-market index (S&P 500) from October 1, 1979, through September 30, 1993. It is evident from panel (a) of the figures that both series display the long-run swings or trending behavior that are characteristic of unit-root*, or $I(1)$, nonstationary processes. On the other hand, the two return series, $r_t = 100 \ln(P_t/P_{t-1})$, in panel (b) appear to be covariance-stationary. However, the tendency for large (and for small) absolute returns to cluster in time is clear.

Many other economic and financial time series exhibit analogous volatility clustering features. This observation, together with the fact that modern theories of price determination typically rely on some form of a risk—reward tradeoff relationship, underlies the very widespread applications of the ARCH class of time series models in economics and finance* over the past decade. Simply treating the temporal dependencies in σ_t as a nuisance would be inconsistent with the trust of the pertinent theories. Similarly, when evaluating economic and financial time series forecasts it is equally important that the temporal variation in the forecast error uncertainty be taken into account.

The next section details some of the most important developments along these lines.

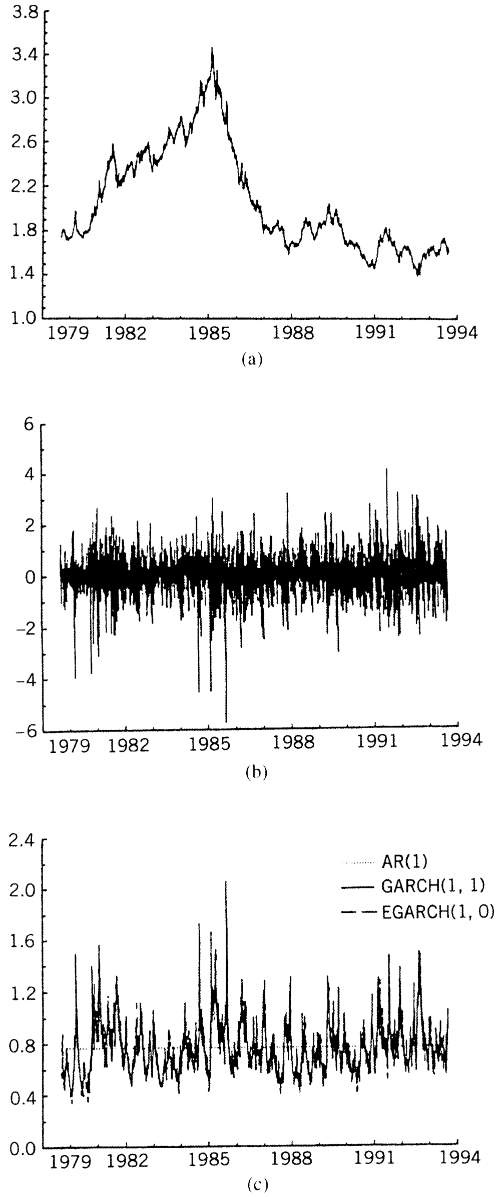


Figure 1. Daily deutsche-mark—U.S. dollar exchange rate. Panel (a) displays daily observations on the DM/U.S. \$ exchange rate, s_t, over the sample period October 1, 1979 through September 30, 1993. Panel (b) graphs the associated daily percentage appreciation of the U.S. dollar, calculated as $r_t \equiv 100 \ln(s_t/s_{t-1})$. Panel (c) depicts the conditional standard-deviation estimates of the daily percentage appreciation rate for the U.S. dollar implied by each of the three volatility model estimates reported in Table 1.

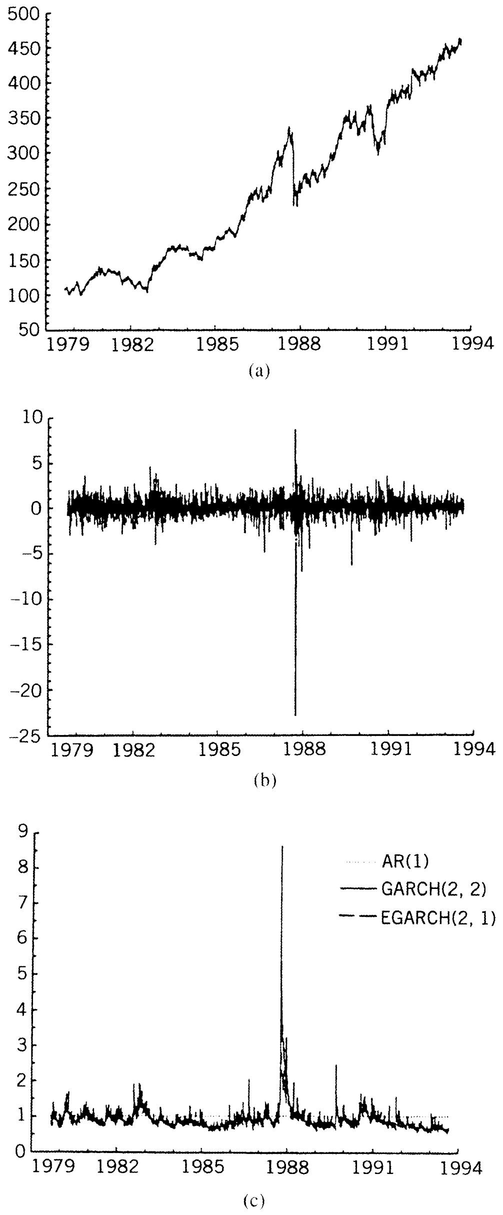


Figure 2. Daily S&P 500 stock-market index. Panel (a) displays daily observations on the value of the S&P 500 stock-market index, P_t, over the sample period October 1, 1979 through September 30, 1993. Panel (b) graphs the associated daily percentage appreciation of the S&P 500 stock index excluding dividends, calculated as $r_t \equiv 100 \ln(P_t/P_{t-1})$. Panel (c) depicts the conditional standard-deviation estimates of the daily percentage appreciation rate for the S&P 500 stock-market index implied by each of the three volatility-model estimates reported in Table 2.

For notational convenience, we shall assume that the $\{\epsilon_t\}$ process is directly observable. However, all of the main ideas extend directly to the empirically more relevant situation in which ϵ_t denotes the time-t innovation of another stochastic process, y_t, as defined above. We shall restrict discussion to the univariate case; most multivariate generalizations follow by straightforward analogy.

GARCH

The definition of the ARCH class of models in (1) is extremely general, and does not lend itself to empirical investigation without additional assumptions on the functional form, or smoothness, of σ_t. Arguably, the two most successful parameterizations have been the *generalized ARCH*, or GARCH (pq), model of Bollerslev [7] and the *exponential GARCH*, or EGARCH (p, q), model of Nelson [46]. In the GARCH (p, q) model, the conditional variance is parametrized as a distributed lag of past squared innovations and past conditional variances,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$
$$\equiv \omega + \alpha(B)\epsilon_t^2 + \beta(B)\sigma_t^2, \qquad (2)$$

where B denotes the backshift (lag) operator; i.e., $B^i y_t \equiv y_{t-i}$. For $\alpha_i > 0$, this parametrization directly captures the tendency for large (small) ϵ_{t-i}^2's to be followed by other large (small) squared innovations. Of course, for the conditional variance in (2) to be positive almost surely, and the process well defined, the coefficients in the corresponding infinite ARCH representation for σ_t^2, expressed in terms of $\{\epsilon_{t-i}^2\}_{i=1}^{\infty}$, must all be nonnegative, i.e., $[1 - \beta(B)]^{-1}\alpha(B)$, where all of the roots of $1 - \beta(x) = 0$ are assumed to be outside the unit circle.

On rearranging the terms in (2), we obtain

$$[1 - \alpha(B) - \beta(B)]\epsilon_t^2 = \omega + [1 - \beta(B)]\nu_t, \quad (3)$$

where $\nu_t \equiv \epsilon_t^2 - \sigma_t^2$. Since $E_{t-1}(\nu_t) = 0$, the GARCH(p, q) formulation in (3) is readily interpreted as an ARMA($\max\{p, q\}, p$) model for the squared innovation process $\{\epsilon_t^2\}$; see Milhøj [43] and Bollerslev [9]. Thus, if the roots of $1 - \alpha(x) - \beta(x) = 0$ lie outside the unit circle, then the GARCH(p, q) process for $\{\epsilon_t\}$ is covariance-stationary, and the unconditional variance equals $\sigma^2 = \omega[1 - \alpha(1) - \beta(1)]^{-1}$. Furthermore, standard ARMA*-based identification and inference procedures may be directly applied to the process in (3), although the heteroscedasticity in the innovations, $\{\nu_t\}$, renders such an approach inefficient.

In analogy to the improved forecast accuracy obtained in traditional time-series analysis by utilizing the conditional as opposed to the unconditional mean of the process, ARCH models allow for similar improvements when modeling second moments. To illustrate, consider the s-step-ahead ($s \geqslant 2$) minimum mean square error* forecast for the conditional variance* in the simple GARCH(1, 1) model,

$$E_t(\epsilon_{t+s}^2) = E_t(\sigma_{t+s}^2)$$
$$= \omega \sum_{i=0}^{s-2} (\alpha_1 + \beta_1)^i + (\alpha_1 + \beta_1)^{s-1}\sigma_{t+1}^2. \qquad (4)$$

If the process is covariance-stationary, i.e. $\alpha_1 + \beta_1 < 1$, it follows that $E_t(\sigma_{t+s}^2) = \sigma^2 + (\alpha_1 + \beta_1)^{s-1}(\sigma_{t+1}^2 - \sigma^2)$. Thus, if the current conditional variance is large (small) relative to the unconditional variance, the multistep forecast is also predicted to be above (below) σ^2, but converges to σ^2 at an exponential rate as the forecast horizon lengthens. Higher-order covariance-stationary models display more complicated decay patterns [3].

IGARCH

The assumption of covariance stationarity has been questioned by numerous studies which find that the largest root in the estimated lag polynomial $1 - \hat{\alpha}(x) - \hat{\beta}(x) = 0$ is statistically indistinguishable from unity. Motivated by this stylized fact, Engle and Bollerslev [22] proposed the so-called *integrated GARCH*, or IGARCH(p, q), process, in which the autoregressive polynomial in (3) has one unit root; i.e., $1 - \alpha(B) - \beta(B) \equiv (1 - B)\phi(B)$, where $\phi(x) \neq 0$ for $|x| \leqslant 1$. However, the notion of a unit root* is intrinsically a linear concept, and considerable care should

Table 1. Daily Deutsche-Mark—U.S. Dollar Exchange-Rate Appreciation

AR(1):(a)					
r_t	$=$	-0.002	$-0.033 \cdot r_{t-1} + \epsilon_t$		
		(0.013)	(0.017)		
		[0.013]	[0.019]		
σ_t^2	$=$	0.585			
		(0.014)			
		[0.022]			
$\text{Log1} = -4043.3, \quad b_3 = -0.25, \quad b_4 = 5.88, \quad Q_{20} = 19.69, \quad Q_{20}^2 = 231.17$					
AR(1)—GARCH(1, 1): (b)					
r_t	$=$	-0.001	$-0.035 \cdot r_{t-1}$	$+\epsilon_t$	
		(0.012)	(0.018)		
		[0.012]	[0.019]		
σ_t^2	$=$	0.019	$+0.103 \cdot \epsilon_{t-1}^2$	$+0.870 \cdot \sigma_{t-1}^2$	
		(0.004)	(0.011)	(0.012)	
		[0.004]	[0.015]	[0.015]	
$\text{Log1} = -3878.8, b_3 = -0.10, b_4 = 4.67, Q_{20} = 32.48, Q_{20}^2 = 22.45$					
AR(1)—EGARCH(1, 0): (c)					
r_t	$=$	0.005	$-0.034 \cdot r_{t-1} + \epsilon_t$		
		(0.012)	(0.018)		
		[0.012]	[0.017]		
$\ln \sigma_t^2$	$=$	-0.447	$+[0.030 \cdot z_{t-1}$	$+0.208 \cdot (\lvert z_{t-1}\rvert - \sqrt{2/\pi})]$	$+0.960 \cdot \ln(\sigma_{t-1}^2)$
		(0.081)	(0.009)	(0.019)	(0.004)
		[0.108]	[0.013]	[0.022]	[0.008]
$\text{Log1} = -3870.8, \quad b_3 = -0.16, \quad b_4 = 4.54, \quad Q_{20} = 33.61, \quad Q_{20}^2 = 24.22$					

Notes: All the model estimates are obtained under the assumption of conditional normality; i.e., $z_t \equiv \epsilon_t \sigma_t^{-1}$ i.i.d. $N(0, 1)$. Conventional asymptotic standard errors based on the inverse of Fisher's information matrix are given in parentheses, while the numbers in square brackets represent the corresponding robust standard errors as described in the text. The maximized value of the pseudo-log-likelihood function is denoted Log1. The skewness and kurtosis of the standardized residuals, $\hat{z}_t = \hat{\epsilon}_t \hat{\sigma}_t^{-1}$, are given by b_3 and b_4, respectively. Q_{20} and Q_{20}^2 refer to the Ljung—Box portmanteau test for up to 20th-order serial correlation in $\hat{z}_t$and $\hat{z}_t^2$, respectively.

be exercised in interpreting persistence in nonlinear models. For example, from (4), the IGARCH(1, 1) model with $\alpha_1 + \beta_1 = 1$ behaves like a random walk*, or an $I(1)$ process, for forecasting purposes. Nonetheless, by repeated substitution, the GARCH(1, 1) model may be written as

$$\sigma_t^2 = \sigma_0^2 \prod_{i=1}^{t} (\alpha_1 z_{t-i}^2 + \beta_1) + \omega \left(1 + \sum_{j=1}^{t-1} \prod_{i=1}^{j} (\alpha_1 z_{t-i}^2 + \beta_1) \right).$$

Thus, as Nelson [44] shows, strict stationarity and ergodicity of the GARCH(1, 1) model requires only geometric convergence of $\{\alpha_1 z_t^2 + \beta_1\}$, or$E[\ln(\alpha_1 z_t^2 + \beta_1)] < 0$, a weaker condition than arithmetic convergence, or $E(\alpha_1 z_t^2 + \beta_1) = \alpha_1 + \beta_1 < 1$, which is required for covariance stationarity. This also helps to explain why standard maximum likelihood* based inference procedures, discussed below, still apply in the IGARCH context [39,42,51].

EGARCH

While the GARCH(p, q) model conveniently captures the volatility clustering phenomenon, it does not allow for asymmetric effects in the evolution of the volatility process. In the EGARCH(p, q) model of Nelson [46], the logarithm of the conditional variance is given as an ARMA(p, q) model in both the absolute size and the sign of the

Table 2. Daily S&P 500 Stock-Market Index Returns

AR(1):(a)						
r_t	=	0.039	$+0.055 \cdot r_{t-1}$	$+\epsilon_t$		
		(0.017)	(0.017)			
		[0.018]	[0.056]			
σ_t^2	=	1.044				
		(0.025)				
		[0.151]				
Log1 $= -5126.7$, $b_3 = -3.20$, $b_4 = 75.37$, $Q_{20} = 37.12$, $Q_{20}^2 = 257.72$						
AR(1)—GARCH(2, 2): (b)						
r_t	=	0.049	$+0.058 \cdot r_{t-1}$	$+\epsilon_t$		
		(0.014)	(0.018)			
		[0.015]	[0.019]			
σ_t^2	=	0.014	$+0.143 \cdot \epsilon_{t-1}^2$	$- 0.103 \cdot \epsilon_{t-2}^2$	$+0.885 \cdot \sigma_{t-1}^2$	$+0.060 \cdot \sigma_{t-2}^2$
		(0.004)	(0.018)	(0.021)	(0.101)	(0.092)
		[0.008]	[0.078]	[0.077]	[0.098]	[0.084]
Log1 $= -4658.0, b_3 = -0.58, b_4 = 8.82, Q_{20} = 11.79, Q_{20}^2 = 8.45$						
AR(1)—EGARCH(2, 1): (c)						
r_t	=	0.023	$+0.059 \cdot r_{t-1} + \epsilon_t$			
		(0.014)	(0.018)			
		[0.015]	[0.017]			
$\ln \sigma_t^2$	=	$0.281 + (1$	$-0.927 \cdot B)[$	$-0.093 \cdot z_{t-1} +$	$0.173 \cdot (\lvert z_{t-1}\rvert - \sqrt{2/\pi})]$	
		(0.175)	(0.031)	(0.014)	(0.018)	
		[0.333]	[0.046]	[0.046]	[0.058]	
		$+1.813 \ln \sigma_{t-1}^2$	$-0.815 \ln \sigma_{t-2}^2$			
		(0.062)	(0.061)			
		[0.113]	[0.112]			
Log1 $= -4643.4$, $b_3 = -0.60$, $b_4 = 9.06$, $Q_{20} = 8.93$, $Q_{20}^2 = 9.37$						

Notes: See Table 1.

lagged innovations,

$$\ln \sigma_t^2 = \omega + \sum_{i=1}^{p} \varphi_i \ln \sigma_{t-i}^2 + \sum_{j=0}^{q} \psi_j g(z_{t-1-j})$$

$$\equiv \omega + \varphi(B) \ln \sigma_t^2 + \psi(B) g(z_t), \qquad (5)$$

$$g(z_t) = \theta z_t + \gamma\,[|z_t| - E(|z_t|)], \qquad (6)$$

along with the normalization $\psi_0 \equiv 1$. By definition, the news impact function $g(\cdot)$ satisfies $E_{t-1}[g(z_t)] = 0$. When actually estimating EGARCH models the numerical stability of the optimization procedure is often enhanced by approximating $g(z_t)$ by a smooth function that is differentiable at zero. Bollerslev et al. [12] also propose a richer parametrization for this function that downweighs the influence of large absolute innovations. Note that the EGARCH model still predicts that large (absolute) innovations follow other large innovations, but if $\theta < 0$ the effect is accentuated for negative ϵ_t's. Following Black [6], this stylized feature of equity returns is often referred to as the "leverage effect."

ALTERNATIVE PARAMETRIZATIONS

In addition to GARCH, IGARCH, and EGARCH, numerous alternative univariate parametrizations have been suggested. An incomplete listing includes: *ARCH-in-mean*, or ARCH-M [25], which allows the conditional variance to enter directly into the equation for the conditional mean of the process; *nonlinear augmented ARCH*, or NAARCH [37], *structural ARCH*, or STARCH [35]; *qualitative threshold ARCH*, or QTARCH [31]; *asymmetric power ARCH*, or AP-ARCH [19]; *switching ARCH*, or SWARCH [16,34]; *periodic GARCH*, or PGARCH [14]; and *fractionally integrated GARCH*, or FIGARCH [4]. Additionally, several authors have proposed the inclusion of various asymmetric terms in the conditional-variance equation to better capture the aforementioned leverage* effect; see e.g., refs. 17, 26, 30.

TIME-VARYING PARAMETER AND BILINEAR MODELS

There is a close relation between ARCH models and the widely-used time-varying parameter class of models. To illustrate, consider the simple ARCH(q) model in (2), i.e., $\sigma_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2$. This model is observationally equivalent to the process defined by

$$\epsilon_t = w_t + \sum_{i=1}^{q} a_i\epsilon_{t-i},$$

where $w_t, a_1, \ldots, a_q$ are i.i.d. random variables with mean zero and variances $\omega, \alpha_1, \ldots, \alpha_q$, respectively; see Tsay [54] and Bera et al. [5] for further discussion. Similarly, the class of bilinear time series models discussed by Granger and Anderson [32] provides an alternative approach for modeling nonlinearities; see Weiss [56] and Granger and Teräsvirta [33] for a more formal comparison of ARCH and bilinear models. However, while time-varying parameter and bilinear models may conveniently allow for heteroskedasticity* and/or nonlinear dependencies through a set of nuisance parameters, in applications in economics and finance the temporal dependencies in σ_t are often of primary interest. ARCH models have a distinct advantage in such situations by directly parametrizing this conditional variance.

ESTIMATION AND INFERENCE

ARCH models are most commonly estimated via maximum likelihood. Let the density for the i.i.d. process z_t be denoted by $f(z_t; v)$, where v represents a vector of nuisance parameters. Since σ_t is measurable with respect to the time-$(t-1)$ observable information set, it follows by a standard prediction-error decomposition argument that, apart from initial conditions, the log-likelihood* function for $\boldsymbol{\epsilon}_T \equiv \{\epsilon_1, \epsilon_2, \ldots, \epsilon_T\}$ equals

$$\log L(\boldsymbol{\epsilon}_T; \boldsymbol{\xi}, v) = \sum_{t=1}^{T} \left[\ln f(\epsilon_t\sigma_t^{-1}; v) - \tfrac{1}{2}\ln\sigma_t^2\right], \tag{7}$$

where ξ denotes the vector of unknown parameters in the parametrization for σ_t. Under conditional normality,

$$f(z_t; v) = (2\pi)^{-1/2}\exp\left(-\tfrac{1}{2}z_t^2\right). \tag{8}$$

By Jensen's inequality*, $E(\epsilon_t^4) = E(z_t^4) \times E(\sigma_t^4) \geqslant E(z_t^4)E(\sigma_t^2)^2 = E(z_t^4)E(\epsilon_t^2)^2$. Thus, even with conditionally normal* innovations, the unconditional distribution for ϵ_t is leptokurtic. Nonetheless, the conditional normal distribution often does not account for all the leptokurtosis in the data, so that alternative distributional assumptions have been employed; parametric examples include the $t -$ distribution* in Bollerslev [8] and the generalized error distribution (GED) in Nelson [46], while Engle and Gonz'alez-Rivera [23] suggest a nonparametric* approach. However, if the conditional variance is correctly specified, the normal quasiscore vector based on (7) and (8) is a martingale* difference sequence when evaluated at the true parameters, $\boldsymbol{\xi}_0$; i.e., $E_{t-1}[\frac{1}{2}(\nabla_\xi\sigma_t^2)\sigma_t^{-2}(\epsilon_t^2\sigma_t^{-2} - 1)] = 0$. Thus, the corresponding quasi-maximum-likelihood* estimate (*QMLE*), $\hat{\boldsymbol{\xi}}$, generally remains consistent, and asymptotically valid inference may be conducted using an estimate of a robustified version of the asymptotic covariance matrix, $\mathbf{A}(\boldsymbol{\xi}_0)^{-1}\mathbf{B}(\boldsymbol{\xi}_0)\mathbf{A}(\boldsymbol{\xi}_0)^{-1}$, where $\mathbf{A}(\boldsymbol{\xi}_0)$ and $\mathbf{B}(\boldsymbol{\xi}_0)$ denote the Hessian* and the outer product of the gradients respectively [55]. A convenient form of $\mathbf{A}(\hat{\boldsymbol{\xi}})$ with first derivatives only is provided in Bollerslev and Wooldridge [15].

Many of the standard mainframe and PC computer-based packages now contain ARCH estimation procedures. These include E-VIEW, RATS, SAS, TSP, and a special set of time series libraries for the GAUSS computer language.

TESTING

Conditional moment (CM) based misspecification tests are easily implemented in the ARCH context via simple auxiliary regressions [50, 53, 57, 58]. Specifically, following Wooldridge [58], the moment condition

$$E_{t-1}[(\boldsymbol{\lambda}_t\sigma_t^{-2})(\epsilon_t^2 - \sigma_t^2)\sigma_t^{-2}] = 0 \tag{9}$$

(evaluated at the true parameter $\boldsymbol{\xi}_0$) provides a robust test in the direction indicated by the vector $\boldsymbol{\lambda}_t$ of misspecification indicators. By selecting these indicators as appropriate functions of the time-$(t-1)$ information set, the test may be designed to have asymptotically optimal power* against a specific alternative; e.g., the conditional variance specification may be tested for goodness of fit over subsamples by letting λ_t be the relevant indicator function, or for asymmetric effects by letting $\lambda_t \equiv \epsilon_{t-1} I\{\epsilon_{t-1} < 0\}$, where $I\{\cdot\}$ denotes the indicator function for $\epsilon_{t-1} < 0$. Lagrange-multiplier-type* tests that explicitly recognize the one-sided nature of the alternative when testing for the presence of ARCH have been developed by Lee and King [40].

EMPIRICAL EXAMPLE

As previously discussed, the two time-series plots for the DM/$ exchange rate and the S&P 500 stock market index in Figs. 1 and 2 both show a clear tendency for large (and for small) absolute returns to cluster in time. This is also borne out by the highly significant Ljung—Box [41] portmanteau tests* for up to 20th-order serial correlation* in the squared residuals from the estimated AR(1) models, denoted by Q_{20}^2 in panel (a) of Tables 1 and 2. To accommodate this effect for the DM/$ returns, Panel (b) of Table 1 reports the estimates from an AR(1)—GARCH(1, 1) model. The estimated ARCH coefficients are overwhelmingly significant, and, judged by the Ljung—Box test, this simple model captures the serial dependence in the squared returns remarkably well. Note also that $\hat{\alpha}_1 + \hat{\beta}_1$ is close to unity, indicative of IGARCH-type behavior. Although the estimates for the corresponding AR(1)—EGARCH(1, 0) model in panel (c) show that the asymmetry coefficient θ is significant at the 5% level, the fit of the EGARCH model is comparable to that of the GARCH specification. This is also evident from the plot of the estimated volatility processes in panel (c) of Fig. 1.

The results of the symmetric AR(1)—GARCH(2, 2) specification for the S&P 500 series reported in Table 2 again suggest a very high degree of volatility persistence. The largest inverse root of the autoregressive polynomial in (3) equals $\frac{1}{2}\{\hat{\alpha}_1 + \hat{\beta}_1 + [(\hat{\alpha}_1 + \hat{\beta}_1)^2 + 4(\hat{\alpha}_2 + \hat{\beta}_2)]^{1/2}\} = 0.984$, which corresponds to a half-life of 43.0, or approximately two months. The large differences between the conventional standard errors* reported in parentheses and their robust counterparts in square brackets highlight the importance of the robust inference procedures with conditionally nonnormal innovations. The two individual robust standard errors for $\alpha_2$andβ_2 suggest that a GARCH(1, 1) specification may be sufficient, although previous studies covering longer time spans have argued for higher-order models [27,52]. This is consistent with the results for the EGARCH(2, 1) model reported in panel (c), where both lags of $g(z_t)$ and $\ln \sigma_1^2$ are highly significant. On factorizing the autoregressive polynomial for $\ln \sigma_t^2$, the two inverse roots equal 0.989 and 0.824. Also, the EGARCH model points to potentially important asymmetric effects in the volatility process. In summary, the GARCH and EGARCH volatility estimates depicted in panel (c) of Fig. 2 both do a good job of tracking and identifying periods of high and low volatility in the U.S. equity market.

FUTURE DEVELOPMENTS

We have provided a very partial introduction to the vast ARCH literature. In many applications a multivariate extension is called for; see refs. 13, 18, 10, 11, 24, 48 for various parsimonious multivariate parametrizations. Important issues related to the temporal aggregation of ARCH models are addressed by Drost and Nijman [20]. Rather than directly parametrizing the functional form for σ_t in (1), Gallant and Tauchen [29], and Gallant et al. [28] have developed flexible nonparametric techniques for analysis of data with ARCH features. Much recent research has focused on the estimation of stochastic volatility models in which the process for σ_t is treated as a latent variable* [1,36,38]. For a more detailed discussion of all of these ideas, see the many surveys listed in the Bibliography below.

A conceptually important issue concerns the rationale behind the widespread empirical findings of IGARCH-type behavior, as exemplified by the two time series analyzed above. One possible explanation is provided by the continuous record asymptotics developed in a series of papers by Nelson [45,47] and Nelson and Foster [49]. Specifically, suppose that the discretely sampled observed process is generated by a continuous-time diffusion, so that the sample path for the latent instantaneous volatility process $\{\sigma_t^2\}$ is continuous almost surely. Then one can show that any consistent ARCH filter must approach an IGARCH model in the limit as the sampling frequency increases. The empirical implications of these theoretical results should not be carried too far, however. For instance, while daily GARCH(1, 1) estimates typically suggest $\hat{\alpha}_1 + \hat{\beta}_1 \approx 1$, on estimating GARCH models for financial returns at intraday frequencies, Andersen and Bollerslev [2] document large and systematic deviations from the theoretical predictions of approximate IGARCH behavior.

This breakdown of the most popular ARCH parametrizations at the very high intraday frequencies has a parallel at the lowest frequencies. Recent evidence suggests that the exponential decay of volatility shocks in covariance-stationary GARCH and EGARCH parametrizations results in too high a dissipation rate at long horizons, whereas the infinite persistence implied by IGARCH-type formulations is too restrictive. The fractionally integrated GRACH, or FIGARCH, class of models [4] explicitly recognizes this by allowing for a low hyperbolic rate of decay in the conditional variance function. However, a reconciliation of the empirical findings at the very high and low sampling frequencies within a single consistent modeling framework remains an important challenge for future work in the ARCH area.

REFERENCES

1. Andersen, T. G. (1996). Return volatility and trading volume: an information flow interpretation of stochastic volatility. *J. Finance*, **51**, 169–204.
2. Andersen, T. G. and Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *J. Empirical Finance*, **4**, 2–3.
3. Baillie, R. T. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *J. Econometrics*, **52**, 91–113.
4. Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996). Fractional integrated generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, **74**, 3–30.
5. Bera, A. K., Higgins, M. L., and Lee, S. (1993). Interaction between autocorrelation and conditional heteroskedasticity: a random coefficients approach. *J. Bus. and Econ. Statist.*, **10**, 133–142.
6. Black, F. (1976). Studies of stock market volatility changes. *Proc. Amer. Statist. Assoc.*, Business and Economic Statistics Section, 177–181.
7. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, **31**, 307–327.
8. Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return. *Rev. Econ. and Statist.*, **69**, 542–547.
9. Bollerslev, T. (1988). On the correlation structure for the generalized autoregressive conditional heteroskedastic process. *J. Time Ser. Anal.*, **9**, 121–131.
10. Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH approach. *Rev. Econ. and Statist.*, **72**, 498–505.
11. Bollerslev, T. and Engle, R. F. (1993). Common persistence in conditional variance. *Econometrica*, **61**, 166–187.
12. Bollerslev, T., Engle, R. F., and Nelson, D. B., (1994). ARCH models. In *Handbook of Econometrics*, vol. **4**, R. F. Engle and D. McFadden, eds., Elsevier Science–North Holland, Amsterdam.
13. Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A capital asset pricing model with time varying covariances. *J. Polit. Econ.*, **96**, 116–131.
14. Bollerslev, T. and Ghysels, E. (1996). Periodic autoregressive conditional heteroskedasticity. *J. Bus. and Econ. Statist.*, **14**, 139–151.
15. Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum-likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Rev.*, **11**, 143–172.
16. Cai. J. (1994). A Markov model switching regime ARCH. *J. Bus. and Econ. Statist.*, **12**, 309–316.

17. Campbell, J. Y. and Hentschel, L. (1992). No news is good news: an asymmetric model of changing volatility in stock returns. *J. Financial Econ.*, **31**, 281–318.

18. Diebold, F. X. and Nerlove, M. (1989). The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *J. Appl. Econometrics*, **4**, 1–21.

19. Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *J. Empirical Finance*, **1**, 83–106.

20. Drost, F. C. and Nijman, T. E. (1993). Temporal aggregation of GARCH processes. *Econometrica*, **61**, 909–928.

21. Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.

22. Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Rev.*, **5**, 1–50, 8187.

23. Engle, R. F. and Gonzalez-Rivera, G. (1991). Semiparametric ARCH models, *J. Bus. and Econ. Statist.*, **9**, 345–359.

24. Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, **11**, 122–150.

25. Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). Estimating time varying risk premia in the term structure: the ARCH-M model. *Econmetrica*, **55**, 391–407.

26. Engle, R. F., and Ng, V. (1993). Measuring and testing the impact of news on volatility. *J. Finance*, **48**, 1749–1778.

27. French, K. R., Schwert, G. W., and Stambaugh, R. F. (1987). Expected stock returns and volatility. *J. Financial Econ.*, **19**, 3–30.

28. Gallant, A. R., Rossi, P. E. and Tauchen, G. (1993). Nonlinear dynamic structures. *Econometrica*, **61**, 871–907.

29. Gallant, A. R. and Tauchen, G. (1989). Semi non-parametric estimation of conditionally constrained heterogeneous processes: asset pricing applications. *Econometrica*, **57**, 1091–1120.

30. Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance*, **48**, 1779–1801.

31. Gourieroux, C. and Monfort, A. (1992). Qualitative threshold ARCH models. *J. Econometrics*, **52**, 159–199.

32. Granger, C. W. J. and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoech and Ruprecht, Göttingen.

33. Granger, C. W. J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford, England.

34. Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *J. Econometrics*, **64** 307–333.

35. Harvey, A. C., Ruiz, E., and Sentana, E. (1992). Unobserved component time series models with ARCH disturbances. *J. Econometrics*, **52**, 129–158.

36. Harvey, A. C., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *Rev. Econ. Stud.*, **61**, 247–264.

37. Higgins, M. L. and Bera, A. K. (1992). A class of nonlinear ARCH models. *Int. Econ. Rev.*, **33**, 137–158.

38. Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models. *J. Bus. and Econ. Statist.*, **12**, 371–417.

39. Lee, S. W. and Hansen, B. E. (1994). Asymptotic theory for the GARCH(1, 1) quasi-maximum-likelihood estimator. *Econometric Theory*, **10**, 29–52.

40. Lee, J. H. H. and King, M. L. (1993). A locally most mean powerful based score test for ARCH and GARCH regression disturbances. *J. Bus. and Econ. Statist.*, **7**, 259–279.

41. Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

42. Lumsdaine, R. L. (1996). Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1, 1) and covariance stationary GARCH(1, 1) models. *Econometrica*.

43. Milhøj, A. (1985). The moment structure of ARCH processes. *Scand. J. Statist.*, **12**, 281–292.

44. Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory*, **6**, 318–334.

45. Nelson, D. B. (1990). ARCH models as diffusion approximations. *J. Econometrics*, **45**, 7–38.

46. Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, **59**, 347–370.

47. Nelson, D. B. (1992). Filtering and forecasting with misspecified ARCH models I: getting

the right variance with the wrong model. *J. Econometrics*, **52**, 61–90.

48. Nelson, D. B., Braun, P. A., and Sunier, A. M. (1995). Good news, bad news, volatility, and betas, *J. Finance*, **50**, 1575–1603.
49. Nelson, D. B. and Foster, D. (1995). Filtering and forecasting with misspecified ARCH models II: making the right forecast with the wrong model. *J. Econometrics*, **67**, 303–335.
50. Newey, W. K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica*, **53**, 1047–1070.
51. Pagan, A. (1996). The econometrics of financial markets. *J. Empirical Finance*, **3**, 15–102.
52. Pagan, A. R. and Schwert, G. W., (1990). Alternative models for conditional stock volatility. *J. Econometrics*, **45**, 267–290.
53. Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *J. Econometrics*, **30**, 415–443.
54. Tsay, R. S. (1987). Conditional heteroskedastic time series models. *J. Amer. Statist. Ass.*, **82**, 590–604.
55. Weiss, A. A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory*, **2**, 107–131.
56. Weiss, A. A. (1986). ARCH and bilinear time series models: comparison and combination. *J. Bus. and Econ. Statist.*, **4**, 59–70.
57. White, H. (1987). Specification testing in dynamic models. In *Advances in Econometrics: Fifth World Congress*, vol. I, T. F. Bewley, ed. Cambridge University Press, Cambridge, England.
58. Wooldridge, J. M. (1990). A unified approach to robust regression based specification tests. *Econometric Theory*, **6**, 17–43.

BIBLIOGRAPHY

Bera, A. K. and Higgins, M. L. (1993). ARCH models: properties, estimation and testing. *J. Econ. Surv.*, **7**, 305–366. (This article discusses the properties of the most commonly used univariate and multivariate ARCH and GARCH models.)

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). ARCH modelling in finance: a review of theory and empirical evidence. *J. Econometrics*, **52**, 5–59. (Chronicles more than two hundred of the earliest empirical applications in the ARCH literature.)

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). ARCH models. In *Handbook of Econometrics*, vol. 4, R. F. Engle and D. McFadden, eds. Elsevier Science–North Holland, Amsterdam. (This chapter contains a comprehensive discussion of the most important theoretical developments in the ARCH literature.)

Brock, W. A., Hsieh, D. A., and LeBaron, B. (1991). *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence*. MIT Press, Cambridge, MA. (Explores the use of ARCH models for explaining nonlinear, possibly deterministic, dependencies in economic data, along with the implications for statistical tests for chaos.)

Diebold, F. X. and Lopez, J. A. (1996). Modeling volatility dynamics. In *Macroeconomics: Developments, Tensions and Prospects*, K. Hoover, ed. Kluwer, Amsterdam. (An easily accessible survey, including some recent results on volatility prediction.)

Enders, W. (1995). *Applied Econometric Time Series*. Wiley, New York. (An easy accessible review of some of the most recent advances in time-series analysis, including ARCH models, and their empirical applications in economics.)

Engle, R. F. (1995). *ARCH: Selected Readings*. Oxford University Press, Oxford, England. (A collection of some of the most important readings on ARCH models, along with introductory comments putting the various papers in perspective.)

Granger, C. W. J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford, England. (Explores recent theoretical and practical developments in the econometric modeling of nonlinear relationships among economic time series.)

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton. (A lucid and very comprehensive treatment of the most important developments in applied time-series analysis over the past decade, including a separate chapter on ARCH models.)

Mills, T. C. (1993). *The Econometric Modelling of Financial Time Series*. Cambridge University Press, Cambridge, England. (Discusses a wide variety of the statistical models that are currently being used in the empirical analysis of financial time series, with much of the coverage devoted to ARCH models.)

Nijman, T. E. and Palm, F. C. (1993). GARCH modelling of volatility: an introduction to theory and applications. In *Advanced Lectures in Economics*, vol. II, A. J. De Zeeuw, ed. Academic Press, San Diego, California. (Discusses the statistical properties and estimation of GARCH models motivated by actual empirical applications in finance and international economics.)

Pagan, A. (1996). The econometrics of financial markets, J. Empirical Finance. (Contains a discussion of the time-series techniques most commonly applied in the analysis of financial time series, including ARCH models.)

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Likelihood, Time Series with Econometric and Other Applications*. D. R. Cox, D. V. Hinkley, and O. E. Barndorff-Nielsen, eds. Chapman and Hall, London. (A discussion and comparison of the statistical properties of ARCH and stochastic volatility models, with an emphasis on recently developed computational procedures.)

Taylor, S. J. (1986). *Modelling Financial Time Series*. Wiley, Chichester, England. (Early treatment of both ARCH and stochastic-volatility-type models for financial time series, with an extensive discussion of diagnostics.)

Taylor, S. J. (1994). Modeling stochastic volatility. *Math. Finance*, **4**, 183–204. (An empirical driven comparison of ARCH and stochastic volatility models.)

See also AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS; DTARCH MODELS; FINANCE, STATISTICS IN; HETEROSCEDASTICITY; and TIME SERIES.

TORBEN G. ANDERSEN
TIM BOLLERSLEV

ARCHAEOLOGY, STATISTICS IN—I

The application of statistical thought to archaeology has been a slow process. This reluctance arises because (1) archaeological data rarely can be gathered in a well-designed statistical experiment; (2) describing empirical findings requires an expertise that is not easily modeled.

In recent years, however, the central problem of archaeology, generally labeled "typology," and the important related problem of "seriation" have received considerable mathematical and statistical attention, which we discuss herewith. The advent of the high-speed computer has made feasible analyses of large sets of archaeological data which were previously impracticable. The application of routine statistical methodology has been infrequent and nonsystematic (see ref. 5, Chap. 13). A recent article by Mueller [23] is noteworthy, illustrating the use of sampling schemes in archaeological survey. The important question of how to locate artifact sites in a region that cannot be totally surveyed is examined. Simple random sampling and stratified sampling are compared in conjunction with an empirical study.

The artifact provides the class of entities with which archaeology is concerned. Typology is concerned with the definition of artifact types. Since mode of usage is unobservable, definition arises from an assortment of qualitative and quantitative variables (e.g., shape, color, weight, length) yielding a list of attributes for each artifact. Artifact types are then defined in terms of "tight clusters of attributes" [19]. The definition of types is usually called taxonomy* and the methods of numerical taxonomy have come to be employed in archaeological typology. (One does find in the literature references to taxonomy for archaeological sites, an isomorphic problem to that of typology for artifacts.)

"Typological debate" [11] has run several decades, resulting in a voluminous literature (see ref. 5). The issues of contention include such matters as whether types are "real" or "invented" to suit the researcher's purposes, whether there is a "best" classification of a body of materials, whether types can be standardized, whether types represent "basic" data, and whether there is a need for more or fewer types. Statistical issues arise in the construction of a typology.

Krieger's effort is a benchmark in unifying the typological concept. An earlier article in this spirit by Gozodrov [10] is deservedly characterized by Krieger as "tentative and fumbling" [19, p. 271]. Krieger reveals the variance in published thought on the classification* issue through examination of the work on pottery description and on projectile style. He articulates the "typological method" and cites Rouse [28] as a good illustration. The earliest quantitative work is by Spaulding [30]. A previous paper by Kroeber [20] concerned itself solely with relating pairs of attributes. The usual χ^2 statistic* (*see* CHI-SQUARE TEST—I) as well as other measures of association* were studied for 2×2 presence–absence attribute tables. From this lead Spaulding suggests, given attribute lists for each artifact in the collection, the preparation of cross-tabulations of all attributes (attribute categories are no longer restricted to presence-absence). Two-way tables are χ^2-tested, leading to the clustering of nonrandomly associated attributes. Artifact types are then defined by identifying classes of artifacts that exhibit sets of associated attributes.

Attempting to formalize Spaulding's technique led archaeologists into the realm of cluster analysis*. A basic decision in cluster analysis is whether the items or the components of the item data vectors are to be clustered. Specifically, are we clustering artifacts (Q-mode* of analysis, as it has been called) or attributes (R-mode* of analysis)? In defining a typology a Q-mode of analysis is appropriate, but the unfortunate use of the phrase "cluster of attributes" by both Krieger and Spaulding has resulted in a persistent confusion in the literature. Factor* and principal components* analyses have occasionally been employed as R-mode analyses to group attributes by significant dimensions interpretable as underlying features of the data.

Typology, then, involves discerning clusters of artifacts on the basis of similarity of their attributes. Similarity between artifacts is measured by similarity functions that arise in other (e.g., psychological and sociological) settings as well. Nonmathematically, a similarity function between two vectors reflects the closeness between components of the vectors as an inverse "distance." For a set of vectors the similarities between all pairs are arranged in a "similarity" matrix*. Beginning with such a matrix, clustering procedures are usually effected with the assistance of a computer.

The earliest computer typologies were done by single-link (nearest-neighbor) clustering [29, p. 180]. Links are assigned from largest similarities and clusters are derived from linked units. Unwelcome "chaining" often occurs, whence average linkage (weighted and unweighted) and complete linkage procedures have been suggested [29, p. 181]. Their sensitivity to spurious large similarities led Jardine and Sibson [15] to formulate double-link cluster analysis, but now "chaining" returns. Thus Hodson [13] proposes a K-means cluster analysis approach. The total collection is partitioned into a predetermined number of clusters. Rules are defined for transferring artifacts from one cluster to another until a "best" clustering is obtained. The procedure is repeated for differing initial numbers, with expertise determining the final number of types defined. The approach can accommodate very large collections.

Similarities implicitly treat attributes in a hierarchical manner in defining types. Whallon [34] suggests that often a hierarchy of importance among attributes exists and that this is how archaeologists feel their way to defining types. Attribute trees are defined where presence or absence of an attribute creates a branch. Employing χ^2 values computed over appropriate 2×2 tables, the sequence of attributes forming the tree is achieved and also the definition of types from such sequences. Cell frequency problems plague many of the χ^2 values. Read [25] formalizes hierarchical classification in terms of partitions of a set of items and allows both discrete and continuous attributes. Clark [3] extends these ideas, assuming discrete attributes and setting them in appropriate higher-order contingency tables* to which log-linear models* are fitted.

In summary, then, a typology is usually obtained by clustering* (through an appropriate procedure) artifacts having similar attributes and defining types through these attributes. Recent hierarchical classification approaches show promise for data sets exhibiting weak clustering.

The next natural step in archaeological enterprise is the comparison of artifact collections, the process called seriation. In broadest terms, seriation consists of arranging a set of collections in a series with respect to similarity of the component artifacts to infer ordering in some nonobservable (usually time) dimension. The collections will typically be grave lots or assemblages. The chronological inference is drawn by the assumption that the degree of similarity between two collections varies inversely with separation in time. Such an assumption implicitly requires a "good" typology for the collections. Of course, other dimensions (e.g., geographic, cultural) may also affect the degree of similarity between collections and confound a "time" ordering. Techniques such as stratigraphy*, dated inscriptions, crossties with established sequences, or radiocarbon dating, if available, would thus preempt seriation. If, in addition to order, one wishes relative distance (in units of time) between the collections, we have a scaling problem as well. The seriation literature is quite extensive. Sterud [33] and Cowgill [6] provide good bibliographies.

The general principles of sequence dating originate with Flinders Petrie [24]. Brainerd [2] and Robinson [27] in companion papers set forth the first formalized mathematical seriation procedure. Robinson offers the methodology with examples and is credited as first to have linked similarities with sequencing. Brainerd provides the archaeological support for the method as well as the interpretation of its results. Some earlier formal attempts in the literature include Spier [31], Driver and Kroeber [8], and Rouse [28]. An assortment of earlier ad hoc seriations are noted by Brainerd [2, p. 304], who comments that they were often qualified as provisional pending stratigraphic support. Kendall [17], making more rigorous the ideas of Petrie, sets the problem as one of estimation. The observed collections Y_i are assumed independent with independent components Y_{ij} indicating the number of occurrences of the jth artifact type in collection i. Each Y_{ij} is assumed Poisson* distributed with mean μ_{ij}, a function of parameters reflecting abundance, centrality, and dispersion. P, the permutation of the Y_i's yielding the true temporal order, is also a parameter. A maximum likelihood* approach enables maximization over the μ_{ij}'s independently of P and yields a scoring function $S(P)$ to be maximized over all permutations. But for as few as 15 collections, exhaustive search for the maximum is not feasible.

Similarities help again the similarity matrix now being between collections, described by vectors, with components noting incidence or abundance of artifact types. Using similarities, an order is specified up to reversibility, with expertise then directing it. Labeling the similarity matrix by F, the objective of a seriation is to find a permutation of the rows and columns to achieve a matrix A with elements a_{ij} such that a a_{ij} increases in j for $j < i$; a_{ij} decreases in j for $j > i$. A similarity matrix having this form has been called a Robinson matrix; the process of manipulating F to this form has been called petrifying. A permutation achieving this form must be taken as ideal under our assumptions but one need not exist. Practically, the goal is to get "close" (in some sense) to a Robinson form.

Taking Robinson's lead, archaeologists such as Hole and Shaw [14], Kuzara et al. [21], Ascher and Ascher [1], and Craytor and Johnson [7], studying large numbers of collections, develop orderings with elaborate computer search procedures (e.g., rules to restrict the search, sampling from all permutations, trial-and-error manipulations). Kendall [18], making these approaches more sophisticated, develops the "horseshoe method," based upon a multidimensional scaling* program in two dimensions. Both theory and examples suggest that with repeated iterations of such a program, a two-dimensional figure in the shape of a horseshoe may be expected if the data are amenable to a seriation. The horseshoe is then unfolded to give a one-dimensional order. Kadane [16] also suggests a computer-based approach by relating the problem of finding a "best" permutation to the traveling salesman problem*. In both cases one seeks a minimum-path-length permutation for a set of points, a problem for which effective computer solutions exist.

Sternin [32] takes a more mathematical tact. He sets the model $F = PAP^T + E$ where P is an unknown permutation matrix and E an error matrix accounting for the possible inability to restore F to exactly a Robinson form. Sternin argues that for certain types of Robinson matrices (e.g., exponential, Green's, and Toeplitz matrices*), the components of eigenvectors corresponding to the two largest eigenvalues will exhibit recognizable patterns. With $E = 0$, F and A have the same eigenvalues*, so Sternin suggests rearranging the components of the corresponding eigenvectors* of F to these patterns.

Gelfand [9] presents two "quick and dirty" techniques. Both methods guarantee the obtaining of the ideal P, if one exists. The better method takes each collection in turn as a reference unit, sequencing all other collections about it. After orienting each of these sequences in the same direction, a final order is obtained by "averaging" these sequences. The averaging should reduce the effect of E and yield a sequence invariant to the original order. An index of fit for a permutation similar to the stress measure used in multidimensional scaling is given, enabling comparison of orders. If an ideal permutation exists, it will minimize this index. Renfrew and Sterud [26] describe a "double-link" method analogous to double-link clustering.

In summary, if an ideal seriation exists, it can be found. If not, but if the data are sufficiently "one-dimensional," the foregoing techniques yield orders from which, with minor modifications suggested by expertise or index of fit, a "best" sequence can be produced.

We now turn to brief discussion of an example. The La Tène Cemetery at Munsingen-Rain near Berne, Switzerland, has proved a rich source of archaeological evidence and has been discussed in numerous articles over the past 15 years. (Hodson [12] provides the definitive work.) The excavation consists of 59 "closed-find" graves. Within these graves were found considerable numbers of fibulae, anklets, bracelets, etc. These ornamental items are typical of the more complex kinds of archaeological material in providing a wide range of detail that allows almost infinite variation within the basic range. A typology for these items was developed employing single-link cluster analysis, average-link cluster analysis, and a principal components analysis*. As a result, some 70 varieties or "types" were defined. A 59 × 70 incidence matrix of types within graves was created and converted to a 59 × 59 similarity matrix between graves. This matrix has been seriated using both the Kendall horseshoe method and Gelfand's technique. The unusual, almost linear form of the cemetery implies a geographical sequencing, which enabled Hodson to establish a very satisfactory seriation. The serial orders obtained by Kendall and by Gelfand in the absence of this information are both in good agreement with Hodson's.

In conclusion, the two books by Clarke [4,5] provide the best current picture of quantitative work in archaeology. Specifically, the articles by Hill and Evans [11] and by Cowgill [6] in the earlier book present excellent synopses on typology and seriation, respectively. The article by Hodson [13] is delightful in bringing some very sophisticated statistical thought to these problems. Finally, the volume from the conference in Mamaia [22] documents a very significant dialogue between archaeologists and statisticians and mathematicians. It bodes well for future analytic work in archaeology.

REFERENCES

1. Ascher, M. and Ascher, R. (1963). *Amer. Anthropol.*, **65**, 1045–1052.
2. Brainerd, G. W. (1951). *Amer. Antiq.*, **16**, 301–313.
3. Clark, G. A. (1976). *Amer. Antiq.*, **41**, 259–273.
4. Clarke, D. L. (1972). *Models in Archaeology*. Methuen, London.
5. Clarke, D. L. (1978). *Analytical Archaeology*, 2nd ed. Methuen, London.
6. Cowgill, G. L. (1972). In *Models for Archaeology*, D. L. Clarke, ed. Methuen, London, pp. 381–424.
7. Craytor, W. B. and Johnson, L. (1968). *Refinements in Computerized Item Seriation. Mus. Nat. History, Univ. Oreg. Bull*. 10.
8. Driver, H. E. and Kroeber, A. L. (1932). *Univ. Calif. Publ. Amer. Archaeol. Ethnol.*, **31**, 211–256.
9. Gelfand, A. E. (1971). *Amer. Antiq.*, **36**, 263–274.
10. Gozodrov, V. A. (1933). *Amer. Anthropol.*, **35**, 95–103.
11. Hill, J. N. and Evans, R. K. (1972). In *Models in Archaeology*, D. L. Clarke, ed. Methuen, London, pp. 231–274.
12. Hodson, F. R. (1968). *The LaTène Cemetery at Munsingen-Rain*. Stämpfi, Berne.
13. Hodson, F. R. (1970). *World Archaeol.*, **1**, 299–320.
14. Hole, F. and Shaw, M. (1967). *Computer Analysis of Chronological Seriation. Rice Uuiv. Stud. No. 53(3)*, Houston.
15. Jardine, N. and Sibson, R. (1968). *Computer J.*, **11**, 177.
16. Kadane, J. B. (1972). *Chronological Ordering of Archaeological Deposits by the Minimum Path Length Method. Carnegie-Mellon Univ. Rep. No. 58* Carnegie-Mellon University, Dept. of Statistics, Pittsburgh, Pa.

17. Kendall, D. G. (1963). *Bull. I.S.I.*, **40**, 657–680.
18. Kendall, D. G. (1971). In *Mathematics in the Archaeological and Historical Sciences*. Edinburgh Press, Edinburgh, pp. 215–252.
19. Krieger, A. D. (1944). *Amer. Antiq.*, **9**, 271–288.
20. Kroeber, A. L. (1940). *Amer. Antiq.*, **6**, 29–44.
21. Kuzara, R. S., Mead, G. R., and Dixon, K. A. (1966). *Amer. Anthropol.*, **68**, 1442–1455.
22. *Mathematics in the Archaeological and Historical Sciences* (1971). Edinburgh Press, Edinburgh.
23. Mueller, J. W. (1974). *The Use of Sampling in Archaeology Survey*. *Amer. Antiq. Mem. No. 28*.
24. Petrie, W. M. Flinders (1899). *J. Anthropol. Inst.*, **29**, 295–301.
25. Read, D. W. (1974). *Amer. Antiq.*, **39**, 216–242.
26. Renfrew, C. and Sterud, G. (1969). *Amer. Antiq.*, **34**, 265–277.
27. Robinson, W. S. (1951). *Amer. Antiq.*, **16**, 293–301.
28. Rouse, I. (1939). *Prehistory in Haiti, A Study in Method*. *Yale Univ. Publ. Anthropol. No. 21*.
29. Sokal, R. R. and Sneath, H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco.
30. Spaulding, A. C. (1953). *Amer. Antiq.*, **18**, 305–313.
31. Spier, L. (1917). *An Outline for a Chronology of Zuni Ruins*. *Anthropol. Papers Amer. Mus. Nat. History*, **18**, Pt. 3.
32. Sternin, H. (1965). *Statistical Methods of Time Sequencing*. *Stanford Univ. Rep. No. 112*, Dept. of Statistics, Stanford University, Stanford, Calif.
33. Sterud, G. (1967). *Seriation Techniques in Archaeology*. Unpublished M.S. thesis, University of California at Los Angeles.
34. Whallon, R. (1972). *Amer. Antiq.*, **37**, 13–33.

See also CLUSTER ANALYSIS; MULTIDIMENSIONAL SCALING; SIMILARITY, DISSIMILARITY AND DISTANCE, MEASURES OF; and TRAVELING-SALESMAN PROBLEM.

ALAN E. GELFAND

ARCHAEOLOGY, STATISTICS IN—II

Applications of statistics to archaeological data interpretation are widespread and can be divided broadly into two groups: those which are descriptive in nature (used primarily to reduce large and/or complex data sets to a more manageable size) and those which are model-based (used to make inferences about the underlying processes that gave rise to the data we observe). Approaches of the first type are most commonly adopted and, in general, are appropriately used and well understood by members of the archaeological profession. Model-based approaches are less widely used and usually rely upon collaboration with a professional statistician.

In the preceding entry Gelfand has provided an excellent survey of the application of statistics to archaeology up to and including the late 1970s. This entry supplements the earlier one, and the emphasis is on work undertaken since that time. Even so, this entry is not exhaustive, and readers are also encouraged to consult the review article of Fieller [15]. Statistics forms an increasingly important part of both undergraduate and graduate courses in archaeology, and several modern textbooks exist. At an introductory level, Shennan [28] assumes very little background knowledge and introduces the reader to both descriptive and model-based approaches. Baxter [1] concentrates on the interpretation of multivariate data in archaeology. The focus is on exploratory rather than model-based approaches, since this has been the primary approach to multivariate data adopted by the archaeological community. Buck et al. [4] take up where the other authors leave off. Their work uses a range of case studies that require model-based approaches and advocates a Bayesian approach so that all prior information is included in the data interpretation process.

We turn first to the uses archaeologists make of simple descriptive statistics. Most modern archaeological field work (and much undertaken in the laboratory) results in the collection of enormous quantities of numeric data. These might take the form of length and breadth measurements used to characterize particular types of artifacts (for example, human and animal bones, pottery vessels, or metal artifacts such as swords or knives) or counts of finds in particular locations (for example, pottery fragments observed on the surface of a site prior to excavation, grave goods deposited with the body at time of burial, or different types of pollen grains obtained by coring different parts of an archaeological landscape). Although such data can represent an enormous range of different archaeological phenomena, the same kinds of statistical approaches are likely to be used to compress the information to a manageable size for presentation and interpretation. Most common are means and standard deviations, percentages, scatter plots, bar charts (2-D and 3-D), and line graphs.

Most archaeological site reports contain a selection of these types of data presentation. In a recent example Cunliffe [14] reports on 25 years of excavation of the Iron Age hill-fort at Danebury in Hampshire, UK. This report provides examples of all the descriptive statistical techniques outlined above and some model-based ones too.

Model-based approaches to archaeological data interpretation have been rather slow to take off, since very few "off the peg" approaches are suitable. Nonetheless, some professional statisticians have shown an interest in helping to interpret archaeological data, and a range of subject-specific model-based approaches have been developed; the most famous is probably the approach used in an attempt to order chronologically (or seriate) archaeological deposits on the basis of the artifact types found within them. A good example might be the desire to establish the order in which bodies were placed in a cemetery on the basis of the grave goods found with them. The basic model is that objects come into use (or "fashion") and then after a period of time go out of use again, but never come back. This model is used not because archaeologists believe that it completely represents the past, but because it adequately reflects the nature of human activity and is not so sophisticated that it cannot be easily adopted in practice. (Gelfand made it the center of the preceding entry.) Development and formalization of the basic model can be attributed to Robinson [27], but see also refs. 6, 19, 21. The early works assumed that archaeological data would conform to the model exactly; Laxton and Restorick [21] noted that there was great potential for stochastic components in archaeological data and modeled this into their approach; and Buck and Litton [6] adopted the same model, but suggested a Bayesian approach so that prior information could also be explicitly modeled into the interpretation process.

Other areas of archaeology that have benefited from adoption of model-based approaches include:

interpretation of soil particle size data in an attempt to understand the nature of the climate and landscape that gave rise to currently observed deposits [16],

consideration of the minimum numbers of individuals represented within assemblages of archaeological artifacts such as disarticulated skeletons [17,18,32],

interpretation of radiocarbon dates in the light of any available relative chronological information from excavation or literary sources [22,5,9], interpretation of data from site surveys; for example soil phosphate analysis or soil resistance measurements [10,3],

identifying the optimum duration and digging strategies for archaeological excavations [23], formalizing descriptions of the shapes and structural mechanics of prehistoric vaulted structures in Europe [11,12,13,20], and interpretation of multivariate chemical compositional data from archaeological artifacts (such as ceramics or glass) collected in an attempt to identify the geological source or site of manufacture [1,8].

Many of these applications are very new and have arisen from recent developments in statistics rather than from any specific changes in archaeological practice. Let us consider the interpretation of radiocarbon dates (for detailed information on radiocarbon dating in archaeology see Bowman [2]) and in so doing return to the report of the Danebury excavations [14]. The first discussion of radiocarbon dating at Danebury (Orton [25]) suggested a mechanism whereby radiocarbon data and archaeological information could be combined within an explicit mathematical framework. The work was completed, however, before the radiocarbon community had adopted a suitable calibration* curve from the many that had been published and before suitable statistical procedures had been developed to allow anything more than point estimates (rather than full distributional information) to be computed. But between 1983 and 1990 dramatic changes took place which we briefly document here.

First, we develop some notation. Suppose that we wish to obtain an estimate for the date of death (θ) of an organic sample found during excavation of the hill-fort at Danebury. This sample is sent for analysis at a radiocarbon dating laboratory, which returns an estimate of the radiocarbon age and an associated estimate of the laboratory error, represented by $y \pm \sigma$. Now the amount of radioactive carbon in the atmosphere has not been constant over time—indeed, it has varied considerably—and as a result a calibration curve is required to map radiocarbon age onto the calendar time scale. The first internationally agreed version of such a curve was published in 1986 [26,29]. It takes the form of bidecadal data that provide a nonmonotonic piecewise-linear calibration curve, which we represent by $\mu(\theta)$. By convention y is then modeled as normally distributed with mean $\mu(\theta)$ and standard deviation σ. This means that for any single radiocarbon determination $y \pm \sigma$ the (posterior) probability distribution of the calendar date θ can fairly readily be computed. Several specialist computer programs exist to do this (for example, CALIB [30,31]). However, because the calibration curve is nonmonotonic and because, in practice, the laboratory errors are often quite large, one radiocarbon determination often does not provide us with much information about the calendar date of interest. Indeed, posterior distributions* commonly have a range of several hundred years.

In an attempt to improve on this, archaeologists soon realized that groups of related determinations would be much more likely to provide precise information than would single ones. This was the approach adopted at Danebury. Cunliffe [14, Table 40, p. 132] reports a total of 60 determinations, all collected with the aim of refining the chronology at Danebury. At the outset Cunliffe realized that sophisticated statistical investigation would be required to make the most of the data available, and his collaboration with Orton began. Between them Orton and Cunliffe developed a model that reflected Cunliffe's beliefs about the relative chronological information at the site.

Naylor and Smith [24] took the story one step further by determining not only that there was a model to be built and that the calibration curve must be allowed for in the interpretation process, but also that the relative chronological information (provided by archaeological stratigraphy) represented prior information, and that the whole problem could be represented using extremely elegant mathematical models.

On the basis of pottery evidence, Cunliffe divided the chronology of the site into four distinct phases. Initially (but see below) he saw these phases as following one another in a strict sequence. A major reason for taking so many radiocarbon samples at Danebury was to learn about the calendar dates of the phase boundaries of the four abutting phases. Consequently, the archaeologists carefully ascribed each of the organic samples to one (and only one) of the four ceramic phases.

In order to explain the statistical approach, we label the calendar dates associated with the 60 samples $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_{60}\}$. We then label the calendar dates of the phase boundaries so that the calendar date of the start of ceramic phase 1 is event 1 and the calendar date of the end of the same phase as event 2. In the same manner the calendar date of the start of phase 2 is event 3 and the calendar date of the end of the phase is event 4. Thus, in order to bound four ceramic phases, we need to define eight events. We then represent these events using the notation $\boldsymbol{\Psi} = \{\Psi_1, \Psi_2, \ldots, \Psi_8\}$. Since the four phases are modeled as abutting, however, we have $\Psi_2 = \Psi_3$, $\Psi_4 = \Psi_5$, and $\Psi_6 = \Psi_7$, and [since calibrated radiocarbon determinations are conventionally reported "before present" (BP), where "present" is 1950]

$$\Psi_1 > \Psi_3 > \Psi_5 > \Psi_7 > \Psi_8. \tag{1}$$

Since the beginning of each phase is always before its end, it can also be stated that

$$\Psi_{2j-1} > \Psi_{2j} \quad (j = 1, 2, 3, 4). \tag{2}$$

It was $\Psi_1, \Psi_3, \Psi_5, \Psi_7$, and Ψ_8 for which Naylor and Smith [24] provided estimates, using computer software based on quadrature methods. Their methodology was coherent, concise, and elegant, but unfortunately they did not work directly with archaeologists. As a result they made one or two fundamental archaeological errors.

With the benefit of hindsight, it may be better that things happened this way, since dramatic changes were taking place in applied statistics that altered the whole approach to calculating Bayesian posterior estimates and revolutionized the way archaeologists think about radiocarbon calibration. The advances in the modeling of archaeological phenomena had taken place against the development of Markov chain Monte Carlo* methods for computing Bayesian posteriors. Realistic models could be used for both likelihoods and priors, since posteriors would simply be simulated from their conditionals. All that is required is large amounts of computer power. Since applied statisticians usually have access to powerful computer workstations, fruitful collaborations were undertaken; Buck et al. [9] is one example.

To understand the approach, note first that, if D is taken to represent the set of all dates for the phase boundaries that satisfy the two sets of inequalities in (1) and (2), the joint prior density for $\boldsymbol{\Psi}$ is

$$\Pr(\boldsymbol{\Psi}) = \begin{cases} c, & \boldsymbol{\Psi} \in D, \\ 0, & \text{otherwise}, \end{cases}$$

where c is a constant. We also need to model the distribution of the θ's within each phase. In the absence of firm archaeological information on deposition rates of the relevant organic samples, it was assumed that if the ith radiocarbon sample is associated with phase j, then its calendar date θ_{ij} (where $j = 1, 2, 3, 4$) is uniformly distributed over the interval Ψ_{2j-1}, Ψ_{2j}, i.e., a uniform deposition rate is assumed:

$$\Pr(\theta_{ij}|\Psi_{2j-1}, \Psi_{2j}) \\ = \begin{cases} (\Psi_{2j-1} - \Psi_{2j})^{-1}, & \Psi_{2j-1} > \theta_{ij} > \Psi_{2j}, \\ 0, & \text{otherwise}. \end{cases}$$

Then assuming that, conditional on the Ψ_k's ($k = 1, 2, \ldots, 8$), the θ_{ij}'s are independent, we have

$$\Pr(\boldsymbol{\theta}|\boldsymbol{\Psi}) = \prod_{j=1}^{4} \prod_{i=1}^{n_j} \Pr(\theta_{ij}|\Psi_{2j-1}, \Psi_{2j}),$$

where n_j is the number of samples in phase j. If x_{ij} represents the radiocarbon determination with standard deviation σ_{ij} which corresponds to the sample with calendar date θ_{ij}, then x_{ij} is a realization of a random variable X_{ij} with mean $\mu(\theta_{ij})$ and variance σ_{ij}^2. Since the calibration curve is piecewise linear,

$$\mu(\theta) = \begin{cases} a_1 + b_1\theta & (\theta \leqslant t_0), \\ a_l + b_l\theta & (t_{l-1} < \theta \leqslant t_l, l = 1, 2, \ldots, L), \\ a_L + b_L\theta & (\theta > t_L), \end{cases}$$

where the t_l are the knots of the calibration curve, $L + 1$ is the number of knots, and a_l and b_l are known constants.

Using Stuiver and Pearson's [29,26] calibration curves and the relative chronological information described above, Buck et al. [9] calibrated the 60 radiocarbon determinations to obtain posterior probability distributions for $\Psi_1, \Psi_3, \Psi_5, \Psi_7$ and Ψ_8 via Markov chain Monte Carlo simulation. They took the joint posterior density to be

$$\Pr(\boldsymbol{\theta}, \boldsymbol{\Psi}|x, \boldsymbol{\sigma}^2) \sim \Pr(x|\boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\sigma}^2) \Pr(\boldsymbol{\theta}|\boldsymbol{\Psi}) \Pr(\boldsymbol{\Psi}),$$

where

$$\Pr(x|\boldsymbol{\Psi}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{4} \prod_{i=1}^{n_j} \Pr(x_{ij}|\theta_{ij}, \sigma_{ij}^2),$$

and the likelihood is given by

$$\Pr(x_{ij}|\theta_{ij}, \sigma_{ij}^2) \\ = (2\pi\sigma_{ij}^2)^{-1/2} \exp\left(-\frac{[x_{ij} - \mu(\theta_{ij})]^2}{2\sigma_{ij}^2}\right);$$

the priors, $\Pr(\boldsymbol{\theta}|\boldsymbol{\Psi})$ and $\Pr(\boldsymbol{\Psi})$, are given above. Originally developed to solve a particular problem, this methodology is fairly general and was published in an archaeological journal. As a result, after about ten years of developmental work and collaboration between more than half a dozen individuals, a model-based, fully coherent approach was available in a forum accessible to archaeological researchers.

This was not the end of the story. Between 1992 and 1995 Cunliffe undertook a reassessment of the data from Danebury and decided that his initial interpretation of the ceramic phases was not entirely appropriate; a more appropriate assessment was that phases 1, 2 and 3 abut one another, but that phase 4 may overlap phase 3. In addition, phase 4 definitely began after the end of phase 2. This required a restatement of relationships between boundary parameters. The new information is that $\Psi_1 > \Psi_2 = \Psi_3 > \Psi_4 = \Psi_5 > \Psi_6$ and $\Psi_5 \geqslant \Psi_7 > \Psi_8$. Having developed a general approach to the work reported in Buck et al. [9], it was possible to alter the model restrictions and recompute the posterior probability distributions for the six parameters now of interest [7].

A large number of researchers (archaeologists and statisticians) have been involved in aiding the interpretation of the radiocarbon determinations from Danebury. As a result there is now a well-developed and widely tested model-based framework in which radiocarbon calibration and interpretation can take place. In general, in order to obtain access to powerful computers, archaeologists currently need to collaborate with statisticians; some see this as a great drawback. In the foreseeable future this is likely to change, but for the moment there are some benefits: applied model-based statistics is not treated as a black box technology, and the applied statistician working on the project ensures that each application is approached afresh, the models are tailor-made for the problem under study, and no presumptions are made that might color judgment about the available prior information.

In summary, statistics is an essential tool for the investigation and interpretation of a wide range of archaeological data types. Descriptive statistics are widely used to allow archaeologists to summarize and display large amounts of otherwise uninterpretable data. Model-based statistics are increasingly widely used, but continue to be most commonly adopted by teams of archaeologists and statisticians working in collaboration. Most model-based statistical interpretations are still seen to be at the cutting edge of archaeological research.

REFERENCES

1. Baxter, M. J. (1994). *Exploratory Multivariate Analysis in Archaeology*. Edinburgh University Press, Edinburgh.
2. Bowman, S. (1990). *Interpreting the Past: Radiocarbon Dating*. British Museum Publications, London.
3. Buck, C. E., Cavanagh, W. G., and Litton, C. D. (1988). The spatial analysis of site phosphate data. In *Computer Applications and Quantitative Methods in Archaeology*, S. P. Q. Rahtz, ed. British Archaeological Reports, Oxford, pp. 151–160.
4. Buck, C. E., Cavanagh, W. G., and Litton, C. D. (1996). *The Bayesian Approach to Interpreting Archaeological Data*. Wiley, Chichester.
5. Buck, C. E., Kenworthy, J. B., Litton, C. D., and Smith, A. F. M. (1991). Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity*, **65**(249), 808–821.
6. Buck, C. E. and Litton, C. D. (1991). A computational Bayes approach to some common archaeological problems. In *Computer Applications and Quantitative Methods in Archaeology 1990*, K. Lockyear and S. P. Q. Rahtz, eds. Tempus Reparatum, British Archaeological Reports, Oxford, pp. 93–99.
7. Buck, C. E. and Litton, C. D. (1995). The radio-carbon chronology: further consideration of the Danebury dataset. In *Danebury: An Iron Age Hill-fort in Hampshire, vol. 6*, a *Hill-fort Community in Hampshire*, Report 102. Council for British Archaeology, pp. 130–136.
8. Buck, C. E. and Litton, C. D. (1996). Mixtures, Bayes and archaeology. In *Bayesian Statistics 5*, A. P. Dawid, J. M. Bernardo, J. Berger, and A. F. M. Smith, eds. Clarendon Press, Oxford, pp. 499–506.
9. Buck, C. E., Litton, C. D., and Smith, A. F. M. (1992). Calibration of radiocarbon results pertaining to related archaeological events. *J. Archaeol. Sci.*, **19**, 497–512.
10. Cavanagh, W. G., Hirst, S., and Litton, C. D. (1988). Soil phosphate, site boundaries and change-point analysis. *J. Field Archaeol.*, **15**(1), 67–83.
11. Cavanagh, W. G. and Laxton, R. R. (1981). The structural mechanics of the Mycenaean

tholos tombs. *Ann. Brit. School Archaeol. Athens*, **76**, 109–140.

12. Cavanagh, W. G. and Laxton, R. R. (1990). Vaulted construction in French Megalithic tombs. *Oxford J. Archaeol.*, **9**, 141–167.
13. Cavanagh, W. G., Laxton, R. R., and Litton, C. D. (1985). An application of change-point analysis to the shape of prehistoric corbelled domes; i. The maximum likelihood method. *PACT*, **11**(III.4), 191–203.
14. Cunliffe, B. (1995). *Danebury: an Iron Age Hill-fort in Hampshire, vol. 6*, a Hill-fort Community in Hampshire, Report 102. Council for British Archaeology.
15. Fieller, N. R. J. (1993). Archaeostatistics: old statistics in ancient contexts. *Statistician*, **42**, 279–295.
16. Fieller, N. R. J. and Flenley, E. (1988). Statistical analysis of particle size and sediments. In *Computer Applications and Quantitative Methods in Archaeology*, C. L. N. Ruggles and S. P. Q. Rahtz, eds. British Archaeological Reports, Oxford, pp. 79–94.
17. Fieller, N. R. J. and Turner, A. (1982). Number estimation in vertebrate samples. *J. Archaeol. Sci.*, **9**(1), 49–62.
18. Horton, D. R. (1984). Minimum numbers: a consideration. *J. Archaeol. Sci.*, **11**(3), 255–271.
19. Kendall, D. G. (1971). Abundance matrices and seriation in archaeology. *Z. Wahrsch. Verw. Geb.*, **17**, 104–112.
20. Laxton, R. R. Cavanagh, W. G., Litton, C. D., Buck, R., and Blair, C. E. (1994). The Bayesian approach to archaeological data analysis: an application of change-point analysis for prehistoric domes. *Archeol. e Calcolatori*, **5**, 53–68.
21. Laxton, R. R. and Restorick, J. (1989). Seriation by similarity and consistency. In *Computer Applications and Quantitative Methods in Archaeology 1989*. British Archaeological Reports, Oxford, pp. 229–240.
22. Litton, C. D. and Leese, M. N. (1991). Some statistical problems arising in radiocarbon calibration. In *Computer Applications and Quantitative Methods in Archaeology 1990*, K. Lockyear and S. P. Q. Rahtz, eds. Tempus Reparatum, Oxford, pp. 101–109.
23. Nakai, T. (1991). An optimal stopping problem in the excavation of archaeological remains. *J. Appl. Statist.*, **28**, 924–929.
24. Naylor, J. C. and Smith, A. F. M. (1988). An archaeological inference problem. *J. Amer. Statist. Ass.*, **83**(403), 588–595.
25. Orton, C. R. (1983). A statistical technique for integrating C-14 dates with other forms of dating evidence. In *Computer Applications and Quantitative Methods in Archaeology*. J. Haigh, ed. School of Archaeological Science, University of Bradford, pp. 115–124.
26. Pearson, G. W. and Stuiver, M. (1986). High-precision calibration of the radiocarbon time scale, 500–2500 BC. *Radiocarbon*, **28**(2B), 839–862.
27. Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *Amer. Antiquity*, **16**, 293–301.
28. Shennan, S. (1997). *Quantifying Archaeology*, 2nd edition. Edinburgh University Press, Edinburgh.
29. Stuiver, M. and Pearson, G. W. (1986). High-precision calibration of the radiocarbon time scale, AD 1950–500 BC. *Radiocarbon*, **28**(2B), 805–838.
30. Stuiver, M. and Reimer, P. (1986). A computer program for radiocarbon age calibration. *Radiocarbon*, **28**(2B), 1022–1030.
31. Stuiver, M. and Reimer, P. (1993). Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon*, **35**(1), 215–230.
32. Turner, A. and Fieller, N. R. J. (1985). Considerations of minimum numbers: a response to Horton. *J. Archaeol. Sci.*, **12**(6), 477–483.

See also ALGORITHMS, STATISTICAL; CALIBRATION—I; and MARKOV CHAIN MONTE CARLO ALGORITHMS.

CAITLIN E. BUCK

ARC-SINE DISTRIBUTION

The arc-sine distribution is a name attributed to a discrete and several continuous probability distributions. The discrete and one of the continuous distributions are principally noted for their applications to fluctuations in random walks*. In particular, the discrete distribution describes the percentage of time spent "ahead of the game" in a fair coin tossing contest, while one of the continuous distributions has applications in the study of waiting times*. The distribution most appropriately termed "arc-sine" describes the location, velocity, and related attributes at random time of a particle in simple harmonic motion. Here "random time" means that the time of observation is independent of the initial phase angle, $0 \leqslant \theta_0 < 2\pi$.

The arc-sine distribution with parameter $b > 0$ has support $[-b, b]$ and PDF $\pi^{-1}(b^2 - x^2)^{-1/2}$ for $-b < x < b$. The position at random time of a particle engaged in simple harmonic motion with amplitude $b > 0$ has the arc-sine (b) distribution.

If X is an arc-sine (1) random variable (RV) and $b \neq 0$, then the RV $Y = bX$ has arc-sine $(|b|)$ distribution. Salient features of this distribution are:

$$\text{moments: } \begin{cases} EX^{2k} = 2^{-2k}\dbinom{2k}{k} \\ EX^{2k+1} = 0 \end{cases}$$

$$(k = 0, 1, 2, \ldots)$$

CDF: $(\sin^{-1} x + \pi/2)/\pi \quad (-1 < x < 1)$
characteristic function: $Ee^{itX} = J_0(t)$,

where $J_0(t)$ is the Bessel function* of the first kind, of order 0, $\sum_{k=0}^{\infty}(-1)^k(t/2)^{2k}/(k!)^2$.

Let $\sim$ denote "is distributed as." In ref. 6, Norton showed that if X_1 and X_2 are independent arc-sine(b) RVs, then $b(X_1 + X_2)/2 \sim X_1X_2$, and in ref. 7 made the following conjecture. Let X_1 and X_2 be independent identically distributed RV's having all moments, and let F denote the common CDF. Then the only nondiscrete F for which $b(X_1 + X_2)/2 \sim X_1X_2$ is the arc-sine(b) distribution. This conjecture was proved by Shantaram [8].

Arnold and Groeneveld [1] proved several results. Let X be a symmetric RV. Then $X^2 \sim (1 + X)/2$ if and only if $X \sim$ arc-sine(1). If X is symmetric and $X^2 \sim 1 - X^2$, then $X \sim 2X\sqrt{1 - X^2}$ if and only if $X \sim$ arc-sine(1). If X_1 and X_2 are symmetric independent identically distributed RVs with $X_i^2 \sim 1 - X_i^2$, then $X_1^2 - X_2^2 \sim X_1X_2$ if and only if $X_i \sim$ arc-sine(1).

Feller [3] discusses distributions that have acquired the arc-sine name. Set $u_{2k} = \dbinom{2k}{k} 2^{-2k}$, $k = 0, 1, 2, \ldots (u_0 = 1)$. Let X_k equal ±1 according to the kth outcome in a fair coin tossing game, and let $S_n = \sum_{k=1}^{n} X_k$ denote the net winnings of a player through epoch $n(S_0 = 0)$. From epochs 0 through $2n$, let Z_{2n} denote that epoch at which the last visit to the origin occurs. Then Z_{2n} necessarily assumes only even values and $\Pr[Z_{2n} = 2k] = u_{2k}u_{2n-2k} = \dbinom{2k}{k}\dbinom{2n-2k}{n-k}2^{-2n}$, $k = 0, 1, \ldots, n$. The probability distribution of the RV Z_{2n} is called the *discrete arc-sine distribution of order n*. Set $\Pr[Z_{2n} = 2k] = p_{2k,2n}$. The probability that in the time interval from 0 to $2n$ the S_j's are positive (the player is ahead) during exactly $2k$ epochs is $p_{2k,2n}$. This result is readily rephrased in terms of $x = k/n$, the proportion of the time the player is ahead. If $0 < x < 1$, the probability that at most $x(100)\%$ of the S_j's are positive tends to $2\pi^{-1}\sin^{-1}\sqrt{x}$ as $n \to \infty$. The corresponding PDF is $\pi^{-1}[x(1 - x)]^{-1/2}$, $0 < x < 1$, which has acquired the name "arc-sine density." Consideration of $p_{2k,2n}$ or the PDF shows that in a fair coin tossing game, being ahead one-half the time is the least likely possibility, and being ahead 0% or 100% of the time are the most likely possibilities. The probability that the first visit to the terminal value S_{2n} occurs at epoch $2k$ (or $2n - 2k$) is $p_{2k,2n}$. In a game of $2n$ tosses the probability that a player's maximum net gain occurs for the first time at epoch k, where $k = 2r$ or $k = 2r + 1$, is $\frac{1}{2}p_{2r,2n}$ for $0 < k < 2n$, u_{2n} for $k = 0$, and $\frac{1}{2}u_{2n}$ for $k = 2n$.

Feller also notes related results in other settings. Let $X_1, X_2, \ldots$ be independent symmetric RVs with common continuous CDF F. Let K_n denote the epoch (index) at which the maximum of $S_0, S_1, \ldots, S_n$ is first attained. Then $\Pr[K_n = k] = p_{2k,2n}$ and, for fixed $0 < \alpha < 1$, as $n \to \infty \Pr[K_n < n\alpha] \to 2\pi^{-1}\sin^{-1}\sqrt{\alpha}$. The number of strictly positive terms among $S_1, \ldots, S_n$ has the same distribution as K_n.

Standard beta* densities with support [0, 1] and having form $f_\alpha(x) = [B(1 - \alpha, \alpha)]^{-1}x^{-\alpha}(1 - x)^{\alpha-1}$, $0 < \alpha < 1$, are called *generalized arc-sine densities*. When $\alpha = \frac{1}{2}$, f_α is the "arc-sine density" $\pi^{-1}[x(1 - x)]^{-1/2}$, $0 < x < 1$, mentioned earlier. Such PDFs play a role in the study of waiting times*. For example, let $X_1, X_2, \ldots$ be positive independent RVs with common CDF F and $S_n = \sum_{k=1}^{n} X_k$. Let N_t denote the random index for which $S_{N_t} \leqslant t < S_{N_t+1}$. Define $Y_t = t - S_{N_t}$. A result of Dynkin [2] is that if $0 < \alpha < 1$ and $1 - F(x) = x^{-\alpha}L(x)$, where $L(tx)/L(t) \to 1$ as $t \to \infty$, then the variable Y_t/t has limiting distribution with PDF f_α. Horowitz [4] extended Dynkin's result to semilinear Markov processes*. Imhof [5] considers the case in which t denotes time and

$\{X(t) : 0 \leqslant t \leqslant T\}$ is a stochastic process* satisfying certain conditions. If V denotes the elapsed time until the process reaches a maximum, then

$$\Pr[V < \alpha T] = 2\pi^{-1} \sin^{-1} \sqrt{\alpha}.$$

REFERENCES

1. Arnold, B. and Groeneveld, R. (1980). *J. Amer. Statist. Ass.*, **75**, 173–175. (Treats some characterizations of the distribution.)
2. Dynkin, E. B. (1961). *Select. Transl. Math. Statist. Probl.*, **1**, 171–189. (Requires probability theory.)
3. Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*. Wiley, New York. (Provides a good survey of the arc-sine distributions.)
4. Horowitz, J. (1971). *Ann. Math. Statist.*, **42**, 1068–1074. (Requires elementary stochastic processes.)
5. Imhof, J. P. (1968). *Ann. Math. Statist.*, **39**, 258–260. (Requires probability theory.)
6. Norton, R. M. (1975). *Sankhyā, A*, **37**, 306–308. (Gives a characterization.)
7. Norton, R. M. (1978). *Sankhyā, A*, **40**, 192–198. (Treats primarily moment properties of discrete RVs.)
8. Shantaram, R. (1978). *Sankhyā, A*, **40**, 199–207. (Uses combinatorial identities.)

See also CHARACTERIZATIONS OF DISTRIBUTIONS; RANDOM WALKS; and TAKÁCS PROCESS.

R. M. NORTON

ARC-SINE TRANSFORMATION. See VARIANCE STABILIZATION

AREA SAMPLING

An area sample is a sample with primary sampling units* that are well-defined fractions of the earth's surface. The sampling frame* can be visualized as a map that has been subdivided into N nonoverlapping subareas that exhaust the total area of interest. The N distinct subareas are the primary sampling units. *See* SURVEY SAMPLING for a discussion of the meanings of primary sampling unit and sampling frame. The sampling units in area sampling are often called *segments* or *area segments*. The list of all area segments is the *area frame*.

Area samples are used to study characteristics of the land, such as the number of acres in specific crops, the number of acres under urban development, the number of acres covered with forest, or the fraction of cultivated acres subject to severe water erosion. Area sampling is an integral part of the U.S. Department of Agriculture's method of estimating acreages and yields of farm crops. *See* AGRICULTURE, STATISTICS IN.

An example of a recent large-scale area sample is the study of the potential for increasing the cropland of the United States conducted by the U.S. Soil Conservation Service. See Dideriksen et al. [1] and Goebel [2]. Area sampling is used heavily in forestry*. See Husch et al. [6] and Labau [13].

Area samples are also used when the observation units are persons or institutions for which a list is not available. For example, area frames are used in studies of the general population in the United States and in other countries where current lists of residents are not maintained. The Current Population Survey of the U.S. Bureau of the Census*, from which statistics on unemployment are obtained, is an area sample of the population of the United States.

Area sampling developed apace with probability sampling*. Mahalanobis* [14] in a discussion of area sampling described the contribution of Hubbock [5], who was responsible for a 1923 study that specified methods of locating a random sample* of areas used in estimating the yield of rice. King [12] cites a number of European studies of the 1920s and 1930s that used a type of area sampling.

In 1943, a large project was undertaken by the Statistical Laboratory of Iowa State College in cooperation with the Bureau of Agricultural Economics, U.S. Department of Agriculture, to design a national area sample of farms in the United States. The name *Master Sample* was applied to the project. The Bureau of the Census* also cooperated in the project and developed an area sample of cities, which, together with the Master Sample of rural areas, was used as a sample of the entire population. The materials developed in the Iowa State project, updated for changes in culture, are still used in the creation of area samples of the rural part of the United States. Stephan [18] provides an excellent review of sampling history.

The basic idea of area sampling is relatively simple, but the efficient implementation of the method requires some sophistication. It must be possible for the field worker (enumerator) to identify the boundaries of each area segment. Thus roads, streets, streams, fences, and other “natural boundaries” are used to define the segments, whenever possible. Aerial photographs, county highway maps, and street maps are materials commonly used in area sampling. Aerial photographs are particularly useful for studies outside heavily urbanized areas. *see* CENSUS.

A precise set of rules associating the elements of the population with the area segments must be developed when the population of interest is not a characteristic of the land itself. For a study of households, the households whose residence is on a given area segment are associated with that segment. Even for this relatively simple rule, problems arise. One problem is the definition of a household. There is also the problem of defining the primary residence in the case of multiple residences. For samples of farms or other businesses, it is common practice to associate the business with the segment on which the “headquarters” is located. See, for example, Jabine [7] for a set of rules of association.

In studies of the characteristics of the land itself, the definition of boundaries of the area segments is very important. A phenomenon called “edge bias” has been identified in empirical studies of crop yields. It has been observed that field workers tend to include plants near the boundary of a plot. Therefore, yields based on small areas are often biased upward. See Sukhatme [19] and Masuyama [16].

Any subdivision of the study area into segments will, theoretically, provide an unbiased estimator* of the population total. But the variances of the estimator obtained for two different partitions of the study area may be very different. Therefore, the design of efficient area samples requires that the area segments be as nearly equal in “size” as is possible. In this context, size is a measure assigned to the area segments that is correlated with the characteristic of interest. An example of a measure of size is the number of households reported in the most recent census. The measure of size used in the Master Sample of Agriculture was the number of dwelling units indicated on county highway maps. This number was correlated with the number of farm headquarters.

In area samples used to study populations over time the size of the area segment will change. Gray and Platek [3] discuss methods of modifying the design in such situations.

Part of the cost of designing area samples is the cost of obtaining information on the estimated size of the area units from recent photos, maps, censuses, city directories, field visits, etc. In designing an area sample, the practitioner must balance the cost of obtaining additional information, the smaller variance of estimates obtained from units of nearly equal estimated size, and the practical requirement for boundaries that can be identified by the field worker.

REFERENCES

1. Dideriksen, R. I., Hidlebaugh, A. R., and Schmude, K. O. (1977). *Potential Cropland Study. Statist. Bull. No. 578*, U.S. Dept. of Agriculture.
2. Goebel, J. J. (1967). *Proc. Social Statist. Sect. Amer. Statist. Ass.*, 1976, pp. 350–354.
3. Gray, C. B. and Platek, R. (1968). *J. Amer. Statist. Ass.*, **63**, 1280–1297.
4. Houseman, E. E. (1975). *Area Frame Sampling in Agriculture. U.S. Dept. Agric. Bull. SRS No. 20*.
5. Hubbock, J. A. (1927). *Sampling for Rice Yield in Bihar and Orissa. Bull. No. 166*, Agricultural Research Institute, Pusa, Government of India. Reprinted in *Sankhyā*, **7**, 281–294 (1947).
6. Husch, B., Miller, C. I., and Beers, T. W. (1972). *Forest Mensuration*. Ronald Press, New York.
7. Jabine, T. B. (1965). In *Estimation of Areas in Agricultural Statistics*. Food and Agriculture Organization of the United Nations, Rome, pp. 136–175.
8. Jessen, R. J. (1942). *Statistical Investigation of a Sample Survey for Obtaining Farm Facts*.

Res. Bull. 304, Iowa Agric. Exper. Stat., Iowa State College, Ames, Iowa.

9. Jessen, R. J. (1945). *J. Amer. Statist. Ass.*, **40**, 45–56.
10. Jessen, R. J. (1947). *J. Farm Econ.*, **29**, 531–540.
11. Jessen, R. J. (1978). *Statistical Survey Techniques*. Wiley, New York.
12. King, A. J. (1945). *J. Amer. Statist. Ass.*, **40**, 38–45.
13. Labau, V. J. (1967). *Literature on the Bitterlich Method of Forest Cruising. U.S. Forest Serv. Res. Paper PNW-47*, U.S. Dept. of Agriculture.
14. Mahalanobis, P. C. (1944). *Philos. Trans. Ser. B*, **231**, 329–451.
15. Mahalanobis, P. C. (1947). *Sankhyā*, **7**, 269–280.
16. Masuyama, M. (1954). *Sankhyā*, **14**, 181–186.
17. Monroe, J. and Finkner, A. L. (1959). *Handbook of Area Sampling*. Chilton, New York.
18. Stephan, F. F. (1948). *J. Amer. Statist. Ass.*, **43**, 12–39.
19. Sukhatme, P. V. (1947). *J. Amer. Statist. Ass.*, **42**, 297–310.

See also Agriculture, Statistics in; Forestry, Statistics in; Small Area Estimation; Survey Sampling; and U.S. Bureau of the Census.

Wayne A. Fuller

ARES PLOTS

A procedure suggested by Cook and Weisberg [1] and originally called "an animated plot for adding regression variables smoothly" was designed to show the impact of adding a set of predictors to a linear model by providing an "animated" plot. The term ARES is an acronym for "adding regressors smoothly" [1]. The basic idea is to display a smooth transition between the fit of a smaller model and that of a larger one. In the case of the general linear model* we could start with the fit of the subset model

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \qquad (1)$$

and then smoothly add X_2 according to some control parameter $\lambda \in [0,1]$ with $\lambda = 0$ corresponding to (1) and $\lambda = 1$ corresponding to the full model:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \qquad (2)$$

The procedure consists in plotting

$$\{\hat{\boldsymbol{Y}}(\lambda), e(\lambda)\},$$

where $\hat{\boldsymbol{Y}}(\lambda)$ are the fitted values and $e(\lambda)$ are the corresponding residuals obtained when the control parameter is equal to λ. A similar device of plotting $\{\boldsymbol{e}(\lambda), \lambda\}$ and $\{\hat{\boldsymbol{Y}}(\lambda), \lambda)\}$ was suggested by Pregibon [4], who call it a *traceplot* or λ-*trace*.

Cook and Weisberg [2] extend ARES for generalized linear models* (see, e.g., McCullagh and Nelder [3]) and provide details on the available software in the LISP-STAT code.

REFERENCES

1. Cook, R. D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphs, Response. *Technometrics*, **31**, 309–311. (Original text, *ibid.*, 273–290.)
2. Cook, R. D. and Weisberg, S. (1994). ARES plots for generalized linear models. *Comput. Statist. and Data Anal.*, **17**, 303–315.
3. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.
4. Pregibon, D. (1989). Discussion of Cook, D. R. and Weisberg, S. (1989). *Technometrics*, **31**, 297–301.

See also Graphical Representation of Data and Regression Diagnostics.

ARFWEDSON DISTRIBUTION

This is the distribution of the number (M_0, say) of zero values among k random variables $N_1, N_2, \ldots, N_k$ having a joint equiprobable multinomial distribution*. If the sum of the N's is n, then

$$\begin{aligned}\Pr[M_0 = m] &= \binom{k}{m}\sum_{i=0}^{m}(-1)^i\binom{m}{i}\left(\frac{m-i}{k}\right)^n \\ &= k^{-n}\binom{k}{m}\Delta^m\mathbf{0}^n \\ &\qquad (m = 0, 1, \ldots, k-1).\end{aligned}$$

It is a special occupancy distribution*, being the number of cells remaining empty after n balls have been distributed randomly among k equiprobable cells.

FURTHER READING

Arfwedson, G. (1951). *Skand. Aktuarietidskr*, **34**, 121–132.

Johnson, N. L., Kotz, S. (1969). *Discrete Distributions*. Wiley, New York, p. 251.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Discrete Distributions*. (2nd ed.) Wiley, New York, p. 414.

See also Difference of Zero; Multinomial Distributions; and Occupancy Problems.

ARIMA MODELS. See Autoregressive–Moving Average (ARMA) Models; Box–Jenkins Model

ARITHMETIC MEAN

The arithmetic mean of n quantities $X_1, X_2, \ldots, X_n$ is the sum of the quantities divided by their number n. It is commonly denoted by $\overline{X}$, or when it is desirable to make the dependence upon the sample size explicit, by $\overline{X}_n$. Symbolically,

$$\overline{X} = (1/n)\sum_{i=1}^{n} X_i.$$

An alternative form of the definition, useful for iterative calculations, is

$$\overline{X}_{n+1} = \overline{X}_n + (X_{n+1} - \overline{X}_n)/(n+1);$$
$$\overline{X}_0 = 0.$$

Historically, the arithmetic mean is one of the oldest algorithmic methods for combining discordant measurements in order to produce a single value, although even so, few well-documented uses date back to before the seventeenth century [3,6,8]. Today it is the most widely used and best understood data summary in all of statistics. It is included as a standard function on all but the simplest hand-held calculators, and it enjoys the dual distinction of being the optimal method of combining measurements from several points of view, and being the least robust* such method according to others. Our discussion begins with the consideration of its distributional properties.

DISTRIBUTION

Suppose that the X_i are independent, identically distributed (i.i.d.) with CDF $F_X(x)$, mean μ, and variance σ^2. The distribution of $\overline{X}$ may be quite complicated, depending upon F_X, but it will always be true that

$$E[\overline{X}] = \mu, \tag{1}$$

$$\operatorname{var}(\overline{X}) = \sigma^2/n, \tag{2}$$

whenever these moments exist. For some distributions $F_X, \overline{X}$ possesses a distribution of simple form. If F_X is $N(\mu, \sigma^2)$, then $\overline{X}$ has a normal $N(\mu, \sigma^2/n)$ distribution*. If F_X is a Bernoulli (p) distribution*, then $n\overline{X}$ has a binomial bin(n, p) distribution*. If F_X is a Poisson (λ) distribution*, $n\overline{X}$ has a Poisson ($n\lambda$) distribution. If F_X is a Cauchy distribution*, then $\overline{X}$ has the same Cauchy distribution. If F_X is a gamma distribution* or a chi-squared distribution*, then $\overline{X}$ has a gamma distribution*. The exact density of $\overline{X}$ when F_X is a uniform distribution* was derived as long ago as 1757 by T. Simpson [6,8]; it is a complicated case of a B-spline*. For further information about the distribution of $\overline{X}$ for different parametric families F_X, see the entries under those distributions.

Since $\overline{X}$ is a sum of independent random variables, many aspects of its distribution are amenable to study by using generating functions*. In particular, the characteristic function* of $\overline{X}$ can be given in terms of the characteristic function $\phi(t)$ of F_X as $\phi_{\overline{X}}(t) = [\phi(n^{-1}t)]^n$.

Much is known about the asymptotic behavior of $\overline{X}_n$ for large n. *See* Laws of Large Numbers; Limit Theorem, Central. For example, the Kolmogorov strong law of large numbers* states that $\overline{X}_n \overset{\text{a.s.}}{\to} \mu$ as $n \to \infty$ if and only if $E|X_i| < \infty$ and $EX_i = \mu$. The classical central limit theorem states that if $\sigma^2 < \infty$, $n^{1/2}(\overline{X}_n - \mu)$ is asymptotically distributed as $N(0, \sigma^2)$. The degree of approximation that can be expected from this asymptotic result has received considerable study, although even the strongest available results are usually too pessimistic to be practically useful. *See* e.g., ASYMPTOTIC NORMALITY.

Various refinements to the normal approximation are also available. *See* CORNISH–FISHER AND EDGEWORTH EXPANSIONS.

If the measurements X_i are not independent, then of course the distribution of $\overline{X}$ may be more complicated. If the X_i form a stationary* sequence with $E[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2$, then $E[\overline{X}] = \mu$, but $\text{var}(\overline{X})$ may be either larger or smaller than in the independent case. For example, if the X_i follow a first-order moving average process [with $X_i = \mu + a_i + \theta a_{i-1}$, where the a_i are i.i.d. with $E[a_i] = 0$, $\text{var}(a_i) = \sigma_a^2$, and $\sigma^2 = (1+\theta^2)\sigma_a^2$], then $\rho = \text{corr}(X_i, X_{i+1}) = \theta/(1+\theta^2)$ varies from -0.5 to 0.5, and

$$\text{var}(\overline{X}) = (\sigma^2/n)[1 + 2(1-(1/n))\rho], \tag{3}$$

which varies from σ^2/n^2 to $2\sigma^2/n - \sigma^2/n^2$. *See* TIME SERIES. If the measurements X_i are determined by sampling at random* from a finite population* of size N without replacement, then

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \tag{4}$$

where σ^2 is the population variance. *See* FINITE POPULATIONS, SAMPLING FROM.

STATISTICAL PROPERTIES

The arithmetic mean $\overline{X}$ is usually considered as an estimator of a population mean μ: if the X_i are i.i.d. with CDF $F_X(x)$ and finite mean μ, then $\overline{X}$ is an unbiased estimator* of μ regardless of F_X (in this sense it is a nonparametric estimator of μ). The same is true if the X_i are sampled at random without replacement from a finite population with mean μ. Chebyshev's inequality* tells us $\Pr[|\overline{X} - \mu| > \in] \leqslant \text{var}(\overline{X})/\in^2$, so $\overline{X}$ will in addition be a consistent estimator* of μ as long as $\text{var}(\overline{X}) \to 0$ as $n \to \infty$, which will hold, for example, if $\sigma^2 < \infty$ in (2), (3), and (4). Laws of large numbers* provide several stronger consistency results. For the case of i.i.d. X_i with finite variances, (2) can be interpreted as stating that the precision of $\overline{X}$ as an estimator of μ increases as the square root of the sample size. In the i.i.d. case, the nonparametric unbiased estimator of $\text{var}(\overline{X})$ is s^2/n, where s^2 is the sample variance* $(n-1)^{-1}\sum(X_i - \overline{X})^2$.

The arithmetic mean enjoys several optimality properties beyond unbiasedness and consistency. It is a special case of a least-squares estimator*; $\sum(X_i - c)^2$ is minimized by $c = \overline{X}$. As such, $\overline{X}$ has all the properties of least-squares estimators: The Gauss–Markov theorem* ensures that $\delta = \overline{X}$ minimizes $E(\delta - \mu)^2$ within the class of all linear unbiased estimators; when F_X is normal $N(\mu, \sigma^2)$, $\overline{X}$ is the maximum-likelihood estimator of μ; and from a Bayesian* perspective, $\overline{X}$ is at the maximum of the posterior distribution* of μ for a uniform prior distribution*. (In fact, C. F. Gauss* proved in 1809 that this later property of $\overline{X}$ characterized the normal within the class of all location parameter families.)

The optimality of $\overline{X}$ as an estimator of a parameter of course depends upon the parametric family in question, but in the i.i.d. case there are several examples of F_X [including $N(\mu, \sigma_0^2)$, σ_0^2 known; Poisson (μ); Bernoulli (μ); the one-parameter exponential (μ)*], where $\overline{X}$ is the maximum-likelihood estimator and a minimal sufficient* statistic. For a simple example of how the optimality of $\overline{X}$ depends upon the distribution and the criterion, however, see ref. 10.

Much attention to the arithmetic mean in recent years has focused upon its lack of robustness*, in particular, its sensitivity to aberrant measurements such as are likely to occur when sampling from heavy-tailed distributions. The most commonly noted example of this was noted as early as 1824 by Poisson: if F_X is a Cauchy distribution*, then $\overline{X}$ has the same Cauchy distribution (and thus no mean or variance no matter how large n is). Estimators such as the sample median* perform much more efficiently in this case. Indeed, even a small amount of heavy-tailed contamination* can in principle drastically effect the efficiency of $\overline{X}$ as an estimator of μ.

Opinion is divided on the question of whether such contamination occurs in practice with a severity or frequency to dictate drastic remedy; see refs. 1, 5, and 8 for an airing of these and related issues. Meanwhile, a vast array of estimators have been devised that are less sensitive than $\overline{X}$ to extreme measurements; the simplest of these (the Winsorized mean* and the trimmed mean*) are, in fact, equivalent to the calculation of the arithmetic mean of a modified sample.

The arithmetic mean is frequently used as a test statistic for testing hypotheses about the mean μ, often in the form of Student's t statistic*, $t = \sqrt{n}(\overline{X} - \mu_0)/s$, and as the basis of confidence intervals* for μ.

RELATIONSHIP TO OTHER MEANS

Two classical means, the geometric* and the harmonic*, are related simply to the arithmetic mean. If $Y_i = \ln X_i$ and $Z_i = X_i^{-1}$, the geometric mean of the X_i is given by $(\prod X_i)^{1/n} = \exp(\overline{Y})$ and the harmonic mean is $(\overline{Z})^{-1}$. Hardy et al. [5] discuss inequality relationships between these and more general mean functions, the simplest being that if all X_i are positive, then $(\overline{Z})^{-1} \leqslant \exp(\overline{Y}) \leqslant \overline{X}$. *See* GEOMETRIC MEAN; HARMONIC MEAN. Many other means have been related to the arithmetic mean in less mathematically precise ways. The best known such relationship is that between the arithmetic mean $\overline{X}$, the median m, and the mode* M for empirical distributions that are unimodal and skewed to the right; it is frequently true that $M \leqslant m \leqslant \overline{X}$. Furthermore, a rough rule of thumb that goes back at least to 1917 (see refs. 2, 4, 7 and 12) observes that these means often satisfy, approximately, the relationship $(\overline{X} - M) = 3(\overline{X} - m)$. The arithmetic mean may also be viewed as the expected valued or mean of the empirical distribution* which places mass 1/n at each X_i, a fact that points to several other characterizations of $\overline{X}$: it is the center of gravity of the X_i; it is value such that the sum of the residuals about that value is zero $[\sum(X_i - \overline{X}) = 0]$; it is a functional of the empirical* cumulative distribution function F_n,

$$\overline{X} = \int x dF_n(x).$$

For other definitions for which the arithmetic mean is a special case, *see* INDEX NUMBERS; ORDER STATISTICS; and ROBUST ESTIMATION.

REFERENCES

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J. (This book presents the results of a Monte Carlo study of a large number of estimates, including the arithmetic mean.)
2. Doodson, A. T. (1917). *Biometrika*, **11**, 425–429.
3. Eisenhart, C. *The Development of the Concept of the Best Mean of a Set of Measurements from Antiquity to the Present Day*. ASA Presidential Address, 1971, to appear. (The early history of the arithmetic mean.)
4. Groeneveld, R. A. and Meeden, G. (1977). *Amer. Statist.*, **31**, 120–121.
5. Hardy, G. H., Littlewood, J. E. and Polya, G. (1964). *Inequalities*. Cambridge University Press, Cambridge.
6. Plackett, R. L. (1958). *Biometrika*, 45, 130–135. Reprinted in *Studies in the History of Statistics and Probability*, E. S. Pearson and M. G. Kendall, eds. Charles Griffin, London, 1970. (The early history of the arithmetic mean.)
7. Runnenburg, J. T. (1978). *Statist. Neerlandica*, **32**, 73–79.
8. Seal, H. L. (1977). *Studies in the History of Statistics and Probability*, Vol. 2, M. G. Kendall and R. L. Plackett, eds. Charles Griffin, London, Chap. 10. (Originally published in 1949.)
9. Stigler, S. M. (1977). *Ann. Statist.*, **5**, 1055–1098. (A Study of the performance of several estimates, including the arithmetic mean, on real data sets. Much of the discussion focuses on the characteristics of real data.)
10. Stigler, S. M. (1980). *Ann. Statist.*, **8**, 931–934. (An early example is discussed which shows that the arithmetic mean need not be optimal, even for samples of size 2, and even if all moments exist.)
11. Tukey, J. W. (1960). In *Contributions to Probability and Statistics*, I. Olkin et al., eds. Stanford University Press, Stanford, Calif., pp. 448–485.

12. Zwet W. R. van (1979). *Statist. Neerlandica*, **33**, 1–5.

STEPHEN M. STIGLER

ARITHMETIC PROGRESSION

This is a sequence of numbers with constant difference between successive numbers. The mth member of the sequence a_m can be expressed as

$$a_m = a_1 + (m - 1)d,$$

where d is the constant difference.

See also GEOMETRIC DISTRIBUTION.

ARITHMETIC TRIANGLE. See COMBINATORICS

ARMA MODELS. See AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS

ARRAY

This term is applied to the distribution of sample values of a variable Y, for a fixed value of another variable X. It refers especially to the frequency distribution (*see* GRAPHICAL REPRESENTATION OF DATA) formed by such values when set out in the form of a contingency table*. Such an array is formed only when the data are discrete or grouped.

See also ORTHOGONAL ARRAYS AND APPLICATIONS.

ARRAY MEAN

An array mean is the arithmetic mean* of the values of a variable Y in a group defined by limits on the values of variables $X_1, \ldots, X_k$ (an array*). It is an estimate of the regression function* of Y on $X_1, \ldots, X_k$.

See also ARRAY; LOG-LINEAR MODELS IN CONTINGENCY TABLES; and REGRESSION, POLYCHOTOMOUS.

ARS CONJECTANDI. See BERNOULLIS, THE

ARTIFICIAL INTELLIGENCE. See STATISTICS AND ARTIFICIAL INTELLIGENCE

ASCERTAINMENT SAMPLING

Ascertainment sampling is used frequently by scientists interested in establishing the genetic basis of some disease. Because most genetically based diseases are rare, simple random sampling* will usually not provide sample sizes of affected individuals sufficiently large for a productive statistical analysis. Under ascertainment sampling, the entire family (or some other well-defined set of relatives) of a proband (i.e., an individual reporting with the disease) is sampled: we say that the family has been *ascertained through a proband*. If the disease does have a genetic basis, other family members have a much higher probability of being affected than individuals taken at random, so that by using data from such families we may obtain samples of a size sufficient to draw useful inferences. Additional benefits of sampling from families are that it yields linkage* information not available from unrelated individuals and that an allocation of genetic and environmental effects can be attempted.

The fact that only families with at least one affected individual can enter the sample implies that a conditional probability must be used in the data analysis, a fact recognized as early as 1912 by Weinberg [5]. However, the conditioning event is not that there is at least one affected individual in the family, but that the family is ascertained. These two are usually quite different, essentially because of the nature of the ascertainment process.

If the data in any family are denoted by D and the event that the family is ascertained by A, the ascertainment sampling likelihood* contribution from this family is the conditional likelihood P(DA)/P(A), where both numerator and denominator probabilities depend on the number of children in the family. The major problem in practice with ascertainment sampling is that the precise nature of the sampling procedure must be known in order to calculate both numerator and denominator probabilities. In practice this procedure is often not well known, and this leads to potential biases in the estimation of genetic parameters, since while the numerator in the above probability can be written as the product $\mathrm{P}(D)\mathrm{P}(A|D)$ of genetic and ascertainment parameters, the denominator cannot, thus confounding estimation of the genetic parameters with the properties of the ascertainment process.

This is illustrated by considering two commonly discussed ascertainment sampling procedures. The first is that of *complete ascertainment*, arising for example from the use of a registry of families, and sampling only from those families in the registry with at least one affected child. Here the probability of ascertainment is independent of the number of affected children. The second procedure is that of *single ascertainment*; here the probability that a family is sampled is proportional to the number of affected children. There are many practical situations where this second form of sampling arises—for example, if we ascertain families by sampling all eighth-grade students in a certain city, a family with three affected children is essentially three times as likely to be ascertained as a family with only one affected child.

These two procedures require different ascertainment corrections. For example, if the children in a family are independently affected with a certain disease, each child having probability p of being affected, the probability of ascertainment of a family with s children under complete ascertainment is $1-(1-p)^s$ and is proportional to p under single ascertainment. The difference between the two likelihoods caused by these different denominators can lead to significant bias in estimation of p if one form of ascertainment is assumed when the other is appropriate.

In practice the description of the ascertainment process is usually far more difficult than this, since the actual form of sampling used is seldom clear-cut (and may well be neither complete nor single ascertainment). For example, age effects are often important (an older child is more likely to exhibit a disease than a younger child, and thus more likely to lead to ascertainment of the family), different population groups may have different social customs with respect to disease reporting, the relative role of parents and children in disease reporting is often not clear-cut (and depends on the age of onset of the disease), and the most frequent method of obtaining data (from families using a clinic in which a physician happens to be collecting disease data) may not be described well by any obvious sampling procedure.

Fisher [3] attempted to overcome these problems by introducing a model in which complete and single ascertainment are special cases. In his model it is assumed that any affected child is a proband with probability π and that children act independently with respect to reporting behavior. Here π is taken as an unknown parameter; the probability that a family with i affected children is ascertained is $1-(1-\pi)^i$, and the two respective limits $\pi = 1$ and $\pi \to 0$ correspond to complete and single ascertainment, respectively. However, the assumptions made in this model are often unrealistic: children in the same family will seldom act independently in reporting a disease, and the value of π will vary from family to family and will usually depend on the birth order of the child. Further, under the model, estimation of genetic parameters cannot be separated from estimation of π, so that any error in the ascertainment model will imply biases in the estimation of genetic parameters.

Given these difficulties, estimation of genetic parameters using data from an ascertainment sampling procedure can become a significant practical problem. An *ascertainment-assumption-free* approach which largely minimizes these difficulties is the following. No specific assumption is made about the probability $\alpha(i)$ of ascertaining a family having i affected children. [For complete ascertainment, $\alpha(i)$ is assumed to be independent of i and, for single ascertainment, to be proportional to i—but we now regard $\alpha(i)$ as an unknown parameter.] The denominator in the likelihood* contribution from any ascertained family is thus $\sum_i p_i\alpha(i)$, where p_i, the probability that a family has i affected children, is a function only of genetic parameters. The numerator in the likelihood contribution is $\mathrm{P}(D) = \alpha(i)$. The likelihood is now maximized jointly with respect to the genetic parameters and the $\alpha(i)$.

When this is done, it is found that estimation of the genetic parameters separates out from estimation of the $\alpha(i)$, and that the former can be estimated directly by using as the likelihood contribution P(D)/P(i) from a family having i affected children. Estimation of ascertainment parameters is not necessary, and the procedure focuses entirely on genetic parameters, being unaffected by the nature of the ascertainment process.

More generally, the data D in any family can be written in the form $D = \{D_1, D_2\}$, where it is assumed that only D_1 affects the probability of ascertainment. Then the likelihood contribution used for such a family is $\mathrm{P}(D_1, D_2)/\mathrm{P}(D_1)$. This procedure [1] gives asymptotically unbiased parameter estimators no matter what the ascertainment process—all that is required is that the data D_1 that is “relevant to ascertainment” can be correctly defined.

These estimators have higher standard error* than those arising if the true ascertainment procedure were known and used in the likelihood leading to the estimate, since when the true ascertainment procedure is known, this procedure conditions on more data than necessary, leaving less data available for estimation. The increase in standard error can be quantified using information concepts. In practice, the geneticist must choose between a procedure giving potentially biased estimators by using an incorrect ascertainment assumption and the increase in standard error in using the ascertainment-assumption-free method.

Conditioning not only on D_1 but on further parts of the data does not lead to bias in the estimation procedure, but will lead to increased standard errors of the estimate by an amount which can be again quantified by information concepts. Further conditioning of this type sometimes arises with continuous data. For such data the parallel with the dichotomy “affected/not affected” might be “blood pressure not exceeding T/blood pressure exceeding T,” for some well-defined threshold T, so that only individuals having blood pressure exceeding T can be probands. To simplify the discussion, suppose that only the oldest child in any family can be a proband. Should the likelihood be conditioned by the probability element $f(x)$ of the observed blood pressure x of this child, or should it be conditioned by the probability $\mathrm{P}(X \geqslant T)$ that his blood pressure exceeds T? The correct ascertainment correction is always the probability that the family is ascertained, so the latter probability is correct. In using the former, one conditions not only on the event that the family is ascertained, but on the further event that the blood pressure is x. Thus no bias arises in this case from using the probability element $f(x)$, but conditioning on further information (the actual value x) will increase the standard error of parameter estimates.

The above example is unrealistically simple. In more realistic cases conditioning on the observed value (or values) often introduces a bias, since when any affected child can be a proband, $f(x)/\mathrm{P}(X \geqslant T)$ is a density function only under single ascertainment. Thus both to eliminate bias and to decrease standard errors, the appropriate conditioning event is that the family is ascertained.

A further range of problems frequently arises when large pedigrees, rather than families, are ascertained. For example if, as above, sampling is through all eighth-grade students in a certain city, and if any such student in the pedigree (affected or otherwise) is not observed, usually by being in a part of the pedigree remote from the proband(s), then bias in parameter estimation will, in general, occur. In theory this problem can be overcome by an exhaustive sampling procedure, but in practice this is usually impossible. This matter is discussed in detail in ref. 4. A general description of ascertainment sampling procedures is given in ref. 2.

REFERENCES

1. Ewens, W. J. and Shute, N. C. E. (1986). A resolution of the ascertainment sampling problem I: theory. *Theor. Pop. Biol.*, **30**, 388–412.
2. Ewens, W. J. (1991). Ascertainment biases and their resolution in biological surveys. In *Handbook of Statistics*, vol. 8, C. R. Rao and R. Chakraborty, eds. North-Holland, New York, pp. 29–61.
3. Fisher, R. A. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.*, **6**, 13–25.

4. Vieland, V. J. and Hodge, S. E. (1994). Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *Amer. J. Human Genet.*, **56**, 33–43.
5. Weinberg, W. (1912). Further contributions to the theory of heredity. Part 4: on methods and sources of error in studies on Mendelian ratios in man. (In German.) *Arch. Rassen- u. Gesellschaftsbiol.*, **9**, 165–174.

See also Human Genetics, Statistics in—I; Likelihood; Linkage, Genetic; and Statistical Genetics.

W. J. E. Wens

ASIMOV'S GRAND TOUR

This is a representation of multivariate data by showing a sequence of bivariate projections of the data.

It is a technique of exploratory projection pursuit*, based on the Cramér–Wold* theorem, which asserts that the distribution of a p-dimensional random vector $\boldsymbol{X}$ is determined by the set of all one-dimensional distributions of the linear combinations $\boldsymbol{\alpha}'\boldsymbol{X}$; here $\boldsymbol{\alpha}\epsilon R^p$ ranges through all "fixed" p-dimensional vectors. The underlying setting for construction of the Asimov grand tour [1] is the Grassmannian manifold $G_{2,p}$. This is the space of all unoriented planes in p-dimensional space. The sequence of projections should become rapidly and uniformly dense in $G_{2,p}$.

Three methods for choosing a path through $G_{2,p}$ are advocated by Asimov. These are the torus method, the at-random method, and the mixture of these two methods called the at-random walk*. The disadvantage of the grand tour is that it necessitates reviewing a large number of planes in order to find any structure. See also ref. [2].

REFERENCES

1. Asimov, D. (1985). The grand tour: A tool for viewing multi-dimensional data. *SIAM J. Sci. Stat. Comput.*, **6**(1), 128–143.
2. Buja, A., Cook, D., Asimov, D., and Harley, C. (1996). Theory and computational methods for dynamic projections in high-dimensional data visualization. *Tech. Memorandum*, Bellcore, NJ.

ASSESSMENT BIAS

LACK OF BLINDING

One of the most important and most obvious causes of assessment bias is lack of blinding. In empirical studies, lack of blinding has been shown to exaggerate the estimated effect by 14%, on average, measured as odds ratio [10]. These studies have dealt with a variety of outcomes, some of which are objective and would not be expected to be influenced by lack of blinding, for example, total mortality.

When patient reported outcomes are assessed, lack of blinding can lead to far greater bias than the empirical average. An example of a highly subjective outcome is the duration of an episode of the common cold. A cold doesn't stop suddenly and awareness of the treatment received could therefore bias the evaluation. In a placebo-controlled trial of vitamin C, the duration seemed to be shorter when active drug was given, but many participants had guessed they received the vitamin because of its taste [12]. When the analysis was restricted to those who could not guess what they had received, the duration was not shorter in the active group.

Assessments by physicians are also vulnerable to bias. In a trial in multiple sclerosis, neurologists found an effect of the treatment when they assessed the effect openly but not when they assessed the effect blindly in the same patients [14].

Some outcomes can only be meaningfully evaluated by the patients, for example, pain and well being. Unfortunately, blinding patients effectively can be very difficult, which is why active placebos are sometimes used. The idea behind an active placebo is that patients should experience side effects of a similar nature as when they receive the active drug, while it contains so little of a drug that it can hardly cause any therapeutic effect.

Since lack of blinding can lead to substantial bias, it is important in blinded trials to test whether the blinding has been compromised. Unfortunately, this is rarely done (Asbjørn Hróbjartsson, unpublished observations), and in many cases, double-blinding is little more than window dressing.

Some outcome assessments are not made until the analysis stage of the trial (see below). Blinding should, therefore, be used also during data analysis, and it should ideally be preserved until two versions of the manuscript—written under different assumptions, which of the treatments is experimental and which is control—have been approved by all the authors [8].

HARMLESS OR FALSE POSITIVE CASES OF DISEASE

Assessment bias can occur if increased diagnostic activity leads to increased diagnosis of true but harmless cases of disease. Many stomach ulcers are silent, that is, they come and go and give no symptoms. Such cases could be detected more frequently in patients who receive a drug that causes unspecific discomfort in the stomach.

Similarly, if a drug causes diarrhoea, this could lead to more digital, rectal examinations, and, therefore, also to the detection of more cases of prostatic cancer, most of which would be harmless, since many people die *with* prostatic cancer but rather few die *from* prostatic cancer.

Assessment bias can also be caused by differential detection of false positive cases of disease. There is often considerable observer variation with common diagnostic tests. For gastroscopy, for example, a kappa value of 0.54 has been reported for the interobserver variation in the diagnosis of duodenal ulcers [5]. This usually means that there are rather high rates of both false positive and false negative findings. If treatment with a drug leads to more gastroscopies because ulcers are suspected, one would therefore expect to find more (false) ulcers in patients receiving that drug. A drug that causes unspecific, nonulcer discomfort in the stomach could, therefore, falsely be described as an ulcer-inducing drug.

The risk of bias can be reduced by limiting the analysis to serious cases that would almost always become known, for example, cases of severely bleeding ulcers requiring hospital admission or leading to death.

DISEASE SPECIFIC MORTALITY

Disease specific mortality is very often used as the main outcome in trials without any discussion how reliable it is, even in trials of severely ill patients where it can be difficult to ascribe particular causes for the deaths with acceptable error.

Disease specific mortality can be highly misleading if a treatment has adverse effects that increases mortality from other causes. It is only to be expected that aggressive treatments can have such effects. Complications to cancer treatment, for example, cause mortality that is often ascribed to other causes although these deaths should have been added to the cancer deaths. A study found that deaths from other causes than cancer were 37% higher than expected and that most this excess occurred shortly after diagnosis, suggesting that many of the deaths were attritutable to treatment [1].

The use of blinded endpoint committees can reduce the magnitude of misclassification bias, but cannot be expected to remove it. Radiotherapy for breast cancer, for example, continues to cause cardiovascular deaths even 20 years after treatment [2], and it is not possible to distinguish these deaths from cardiovascular deaths from other causes. Furthermore, to work in an unbiased way, death certificates and other important documents must have been completed and patients and documents selected for review without awareness of status, and it should not be possible to break the masking during any of these processes, including review of causes of death. This seems difficult to obtain, in particular, since those who prepare excerpts of the data should be kept blind to the research hypothesis [3].

Fungal infections in cancer patients with neutropenia after chemotherapy or bone-marrow transplantation is another example of bias in severely ill patients. Not only is it difficult to establish with certainty that a patient has a fungal infection and what was the cause of death; there is also evidence that some of the drugs (azole antifungal agents) may increase the incidence of bacteriaemias [9]. In the largest placebo-controlled trial of fluconazole, more deaths were reported on drug than on placebo (55 vs 46 deaths), but the authors also reported that fewer deaths were ascribed to acute systemic fungal infections (1 vs 10 patients, $P = 0.01$) [6]. However, if this subgroup result is to be believed, it would mean that fluconazole increased mortality from other causes (54 vs 36 patients, $P = 0.04$).

Bias related to classification of deaths can also occur within the same disease. After publication of positive results from a trial in patients with myocardial infarction [16], researchers at the US Food and Drug Administration found that the cause-of-death classification was "hopelessly unreliable" [15]. Cardiac deaths were classified into three groups: sudden deaths, myocardial infarction, or other cardiac event. The errors in assigning cause of death, nearly all, favoured the conclusion that sulfinpyrazone decreased sudden death, the major finding of the trial.

COMPOSITE OUTCOMES

Composite outcomes are vulnerable to bias when they contain a mix of objective and subjective components. A survey of trials with composite outcomes found that when they included clinician-driven outcomes, such as hospitalization and initiation of new antibiotics, in addition to objective outcomes such as death, it was twice as likely that the trial reported a statistically significant effect [4].

COMPETING RISKS

Composite outcomes can also lead to bias because of competing risks [13], for example, if an outcome includes death as well as hospital admission. A patient who dies cannot later be admitted to hospital. This bias can also occur in trials with simple outcomes. If one of the outcomes is length of hospital stay, a treatment that increases mortality among the weakest patients who would have had long hospital stays may spuriously appear to be beneficial.

TIMING OF OUTCOMES

Timing of outcomes can have profound effects on the estimated result, and the selection of time points for reporting of the results is often not made until the analysis stage of the trials, when possible treatment codes have been broken. A trial report of the antiarthritic drug, celecoxib, gave the impression that it was better tolerated than its comparators, but the published data referred to 6 months of follow-up, and not to 12 and 15 months, as planned, when there was little difference; in addition, the definition of the outcome had changed, compared to what was stated in the trial protocol [11].

Trials conducted in intensive care units are vulnerable to this type of bias. For example, the main outcome in such trials can be total mortality during the stay in the unit, but if the surviving patients die later, during their subsequent stay at the referring department, little may be gained by a proven mortality reduction while the patients were sedated. A more relevant outcome would be the fraction of patients who leave the hospital alive.

ASSESSMENT OF HARMS

Bias in assessment of harms is common. Even when elaborate, pretested forms have been used for registration of harms during a trial, and guidelines for their reporting have been given in the protocol, the conversion of these data into publishable bits of information can be difficult and often involves subjective judgments.

Particularly vulnerable to assessment bias is exclusion of reported effects because they are not felt to be important, or not felt to be related to the treatment. Trials that have been published more than once illustrate how subjective and biased assessment of harms can be. Both number of adverse effects and number of patients affected can vary from report to report, although no additional inclusion of patients or follow-up have occurred, and these reinterpretations or reclassifications sometimes change a nonsignificant difference into a significant difference in favor of the new treatment [7].

REFERENCES

1. Brown, B. W., Brauner, C., and Minnotte, M. C. (1993). Noncancer deaths in white adult cancer patients. *J. Natl. Cancer Inst.*, **85**, 979–987.
2. Early Breast Cancer Trialists' Collaborative Group. (2000). Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials. *Lancet*, **355**, 1757–1770.

3. Feinstein, A. R. (1985). *Clinical Epidemiology*. W. B. Saunders, Philadelphia, Pa.
4. Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., and Griffin, C. (2003). Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA*, **289**, 2554–2559.
5. Gjørup, T., Agner, E., Jensen, L. B, Jensen, A. M., and Møllmann, K. M. (1986). The endoscopic diagnosis of duodenal ulcer disease. A randomized clinical trial of bias and interobserver variation. *Scand. J. Gastroenterol.*, **21**, 561–567.
6. Goodman, J. L., Winston, D. J., Greenfield, R. A., Chandrasekar, P. H., Fox, B., Kaizer, H., et al. (1992). A controlled trial of fluconazole to prevent fungal infections in patients undergoing bone marrow transplantation. *N. Engl. J. Med.*, **326**, 845–851.
7. Gøtzsche, P. C. (1989). Multiple publication in reports of drug trials. *Eur. J. Clin. Pharmacol.*, **36**, 429–432.
8. Gøtzsche, P. C. (1996). Blinding during data analysis and writing of manuscripts. *Control. Clin. Trials*, **17**, 285–290.
9. Gøtzsche, P. C and Johansen, H. K. (2003). Routine versus selective antifungal administration for control of fungal infections in patients with cancer (Cochrane Review). *The Cochrane Library*, Issue 3. Update Software, Oxford.
10. Jüni, P., Altman, D. G., and Egger, M. (2001). Systematic reviews in health care: assessing the quality of controlled clinical trials. *Br. Med. J.*, **323**, 42–46.
11. Jüni, P., Rutjes, A. W., and Dieppe, P. A. (2002). Are selective COX 2 inhibitors superior to traditional non steroidal anti-inflammatory drugs? *Br. Med. J.*, **324**, 1287–1288.
12. Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Kapikian, A. Z., Lewis, T. L., and Lynch, J. M. (1975). Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *JAMA*, **231**, 1038–1042.
13. Lauer, M. S. and Topol, E. J. (2003). Clinical trials—multiple treatments, multiple end points, and multiple lessons. *JAMA*, **289**, 2575–2577.
14. Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R. E., Yetisir, E., and Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*, **44**, 16–20.
15. Temple, R. and Pledger, G. W. (1980). The FDA's critique of the anturane reinfarction trial. *N. Engl. J. Med.*, **303**, 1488–1492.
16. The Anturane Reinfarction Trial Research Group. (1980). Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N. Engl. J. Med.*, **302**, 250–256.

See also CLINICAL TRIALS and MEDICINE, STATISTICS IN.

PETER C. GØTZSCHE

ASSESSMENT OF PROBABILITIES

NORMATIVE AND DESCRIPTIVE VIEWS

Central to this entry is a person, conveniently referred to as 'you', who is contemplating a set of *propositions* $A, B, C, \ldots$. You are *uncertain* about some of them; that is, you do not know, in your current state of knowledge $\mathscr{K}$, whether they are true or false. (An alternative form of language is often employed, in which $A, B, C, \ldots$ are *events* and their *occurrence* is in doubt for you. The linkage between the two forms is provided by propositions of the form 'A has occurred'.) You are convinced that the only logical way to treat your uncertainty, when your knowledge is $\mathscr{K}$, is to assign to each proposition, or combination of propositions, a probability $\Pr[A|\mathscr{K}]$ that A is true, given $\mathscr{K}$. This probability measures the strength of your *belief* in the truth of A. The task for you is that of *assessing* your probabilities; that is, of providing numerical values. That task is the subject of this article but, before discussing it, some side issues need to be clarified.

When it is said that you think that uncertainty is properly described by probability, you do not merely contemplate assigning numbers lying between 0 and 1, the convexity rule. Rather, you wish to assign numbers that obey all three rules of the probability calculus; convexity, addition, and multiplication (*see* PROBABILITY, FOUNDATIONS OF—I). This is often expressed as saying that your beliefs must *cohere* (*see* COHERENCE—I). It is therefore clear that, in order to perform the assessment, you must understand the rules of probability and their implications. You must be familiar with the calculus. In a sense, you have set yourself a *standard*, of coherence, and wish to adhere to that standard, or norm. This is often called the *normative* view of probability. A surveyor uses the normative theory of Euclidean geometry.

In contrast to the normative is the *descriptive* view, which aims to provide a description of how people in general attempt to deal with their uncertainties. In other words, it studies how people assess the truth of propositions when they do not have the deep acquaintance with the probability calculus demanded of the normative approach, nor feel the necessity of using that calculus as the correct, logical tool. There are several types of people involved. At one extreme are children making the acquaintance of uncertainty for the first time. At the other extreme are sophisticated people who employ different rules from those of the probability calculus; for example, the rules of fuzzy logic. This entry is *not* concerned with the descriptive concept. The seminal work on that subject is Kahneman et al. [3]. A recent collection of essays is that of Wright and Ayton [8].

Knowledge gained in descriptive studies may be of value in the normative view. For example, the former have exposed a phenomenon called *anchoring*, where a subject, having assessed a probability in one state of knowledge, may remain unduly anchored to that value when additional knowledge is acquired, and not change sufficiently. The coherent subject will update probabilities by Bayes' rule of the probability calculus (*see* BAYES' THEOREM). Nevertheless, an appreciation of the dangers of anchoring may help in the avoidance of pitfalls in the assessment of the numerical values to use in the rule. In view of the central role played by Bayes' rule, the coherent approach is sometimes called *Bayesian*, at least in some contexts.

Texts on the probability calculus do not include material on the assignment of numerical values, just as texts on geometry do not discuss mensuration or surveying. Assessment is an adjunct to the calculus, as surveying is to Euclidean geometry. Yet the calculus loses a lot of its value without the numbers.

SUBJECTIVE PROBABILITY*

In the context adopted here of your contemplating uncertain propositions or events and adhering to the probability calculus, the form of probability employed is usually termed *subjective* or *personal*. Your appreciation of uncertainty may be different from mine. Subjective ideas have been around as long as probability, but it is only in the second half of the twentieth century that they have attained the force of logic. This is largely due to the work of Ramsey [6] (whose work lay for long unappreciated), Savage [7], and De Finetti [2], amongst others. In various ways, these writers showed that, starting from axioms held to be self-evident truths, the existence of numbers describing your uncertainty in a given state of knowledge could be established, and that these numbers obeyed the rules of the probability calculus. The coherent, normative view thus follows from the axioms. It is important that although the axioms adopted by these writers differ—some treating uncertainty directly, others incorporating action in the face of uncertainty—they all lead to the same conclusion of "the inevitability of probability." Logic deals with truth and falsity. Probability is an extension of logic to include uncertainty, truth and falsity being the extremes, 1 and 0.

With these preliminaries out of the way, let us turn to the central problem: how can you, adhering to the normative view, and wishing to be coherent, assign numerical values to your uncertainties?

FREQUENCY DATA

There is one situation, of common occurrence, where the assessment of a probability is rather simple. This arises when, in addition to the event A about which you are uncertain, there is a sequence $A_1, A_2, \ldots, A_n$ of similar events where the outcome is known to you. If just r of these events are true, then it seems reasonable to suppose your probability of A is about r/n. The classic example is that of n tosses of a coin, yielding r heads and $n - r$ tails, where you are uncertain about the outcome of the next toss. In these cases the frequency r/n is equated with belief*. The situation is of such common occurrence that probability has often been identified with frequency, leading to a *frequentist* view of probability. According to the logical, subjective view this is erroneous. It is the data which are frequentist, not the probability. For the near-identification of frequency and belief within the normative approach, some conditions have to be imposed on your belief structure. These have been given by De Finetti in the concept of exchangeability* amongst all the events considered. The complete identification of frequency and belief is unreasonable, as can be seen by taking the case $n = 1$, when the identification would only leave 0 and 1 as possible values of the probability for the second toss.

CLASSICAL VIEW

A second common situation where probabilities are easily assessed is where the possible outcomes of an uncertain situation, N in all, are believed by you to be equally probable. Thus your belief that a stranger was born on a specific date is about 1/365; here $N = 365$. This is the *classical* view of probability. Applied to the coin-tossing example of the last paragraph, the probability of heads is $\frac{1}{2}$. Again, the connection between the normative and classical views requires a belief judgment on your part, namely of equal belief in all N possibilities. Notice that in the birthday example, you, understanding the calculus, can evaluate your belief that, amongst 23 independent strangers, at least two share the same birthday; the probability is about $\frac{1}{2}$. Experience shows that this is typically not true in the description of people's behavior.

There remain many situations in which neither are frequency data available, nor is the classical approach applicable, and yet you have to express your belief in an uncertain proposition. An important example occurs in courts of law where the jury, acting as 'you', has to express its belief in the defendant's guilt. A lot of information is available, in the form of the evidence produced in court, but it is rarely of the type that leads to a frequentist or classical analysis. So we now turn to other methods of assessment.

SCORING RULES

One assessment tool is the scoring rule*. This is derived from a method used as an axiom system by De Finetti. If you express your belief in the occurrence of an event E by a number x, then when the status of E becomes known, the rule assigns a penalty score dependent on x and the status. A popular rule is the quadratic one $(E - x)^2$, where E is also the indicator function of the event: $E = 1(0)$ if E is true (false). With some pathological exceptions, if you attempt to minimize your total expected penalty score, you will make assertions x which are such that they, or some function of them, obey the laws of probability. With the quadratic rule, x obeys the laws. Another rule yields log odds, rather than probabilities. The classic exposition is by Savage [7]. It is possible to train people to make probability statements using scoring rules, and the method has been applied by meteorologists issuing probability weather forecasts.

COHERENCE

One thing is clear about probability assessment: it is best to study beliefs about several related events, rather than to study an event in isolation. The reason for this is that it gives an opportunity for the rules of probability to be invoked. For a single event, only the convexity rule, that probability lies in the unit interval, is relevant. The simplest case of two events, E and F, illustrates "the point." You might assess $\Pr[E], \Pr[F|E], \Pr[F|E^c]$ as three numbers, unrestricted except that they each lie in the unit interval. (E^c denotes the complement of E, and a fixed knowledge base is assumed.) The rules then imply values for $\Pr[F], \Pr[E|F]$, etc. You would then consider whether these implications are in reasonable agreement with your beliefs. If they are not, then some adjustment will need to be made to the three original values. With more events, more checks are available and therefore better assessments possible. The object is to reach a set of probabilities that are coherent. Guidance is also available on which probabilities are best to assess. For example, under some conditions, it is better to assess $\Pr[E|F]$ directly, rather than $\Pr[EF]$ and $\Pr[F]$, using their ratio as an indirect assessment [4]. Computer programs can be written wherein some probability assessments may be fed in and others either calculated or limits given to where they must lie. By De Finetti's fundamental theorem* [2], they must lie in an interval. If the interval is empty, the values fed in are incoherent and some reassessment is necessary. Since the normative view is founded on the concept of coherence, its use in numerical assessment seems essential. It is the preferred method for a jury, who should aim to be coherent with all the evidence presented to them.

ROBUSTNESS

There has been a great deal of study of the robustness of statistical procedures (*see*, e.g., ROBUST ESTIMATION). Suppose that you know that the procedure you are using is robust to one of the probabilities. That is, the final conclusions are not sensitive to the actual numerical value assigned to that probability. Then this helps in the assessment, because the evaluation need not be done with great care. On the other hand, if it is known that the procedure is sensitive to a value, great care should be devoted to its determination, for example, by checking its coherence with other numerical probabilities. Knowledge of robustness shows where the work on assessment should be concentrated.

BETTING

Another way to perform the assessment is for you to consider bets (*see* GAMBLING, STATISTICS IN). At what odds will you accept a bet on the event E? When this is repeated for several events, either coherence results, or a Dutch book*, in which you will lose money for sure, can be made against you. Implicit in this method is the assumption that you are invoking only monetary considerations and that your utility* for money is locally linear. The latter restriction might be achieved by making the amounts of money involved small. The former is more difficult to attain. People commonly overreact by perceiving even a modest loss as a disaster or an embarrassment, and likewise are elated by a small win. A gain of $100 in a gamble is viewed differently from a gain of the same amount resulting from honest toil. Decision analysis, based on personalistic ideas, demonstrates that, to make a decision, only products of a probability and a utility matter. No decision changes if one is multiplied by c and the other divided by c. Probability and utility seem inseparable. Yet belief, as a basic concept, seems genuinely to exist in your mind, irrespective of any action you might take as a result of that belief. This conundrum awaits resolution before betting, or similar devices based on action, can be used with complete confidence for assessment.

CALIBRATION*

Suppose that you provide probabilities for a long sequence of events and, for a selected number p, we select all the events to which you assigned probability p. For realism, suppose your probabilities are only to one place of decimals. It seems reasonable to expect that a proportion p of the selected events should subsequently turn out to be true and the remainder false. If this happens for all p, you are said to be *well calibrated*. Generally, a curve of the proportion true against the stated value p is called a calibration curve. An example is again provided by meteorologists forecasting tomorrow's weather. Some studies show they are well calibrated [5].

EXPERT OPINION

Calibration may be a consideration when you need to use an expert's probability for your own purposes. For example, you have to decide whether to go on a picnic tomorrow, and you consult an expert weather forecaster, who says that there is a 90% probability that tomorrow will be fine. If the expert is thought by you to be well calibrated, then you may reasonably assess your probability of fine weather also to be 90%. But you could equally well proceed if you knew the meteorologist's calibration curve. For instance, if the meteorologist is found to be an optimist and only 75% of their predicted days at 90% turn out to be fine, you might assess your probability for tomorrow being fine as $\frac{3}{4}$.

Expert opinion has been studied in a different way by other researchers. Suppose an expert provides probability p for an event E of interest to you. How should this provision affect your probability for E? Bayes' theorem* in odds form says

$$\mathrm{Od}[E|p] = \frac{\Pr[p|E]}{\Pr[p|E^c]}\mathrm{Od}[E],$$

where Od means odds on and $\Pr[p|E]$ is your probability that, if E is true, the expert will announce p, and similarly for E false, E^c true. In the usual way, your odds are updated by multiplying by your likelihood ratio for the data p. The latter necessitates your assessment of your probabilities that the expert will announce p when the event is true, and also when it is false. Notice that this is not the calibration curve, which fixes p and examines a frequency.

Unexpected results follow from this analysis of expert opinion. Suppose several experts announce that an event has, for them, probability 0.7. Suppose further (a severe assumption) you feel the experts are independent; this is a probability judgment by you. Then calculation, following the probability calculus, suggests that your probability for the event should exceed 0.7. This is supported by the intuitive consideration that if the event were false, it would be astonishing that every expert consulted should think it more likely to be true than not. Contrast this with the agreement amongst several experts that two cities are 70 miles apart. You would unhesitatingly accept 70 as also your opinion. Probability behaves differently. A convenient reference is Cooke [1].

FUTURE WORK

It is remarkable how little attention has been paid to the measurement of probabilities. The sophisticated calculus applies to numbers that, outside the limited frequency and classical domains, have not been assessed satisfactorily. And those that have, have rarely been tested. For example, many research studies throughout science quote a significance level, stating that the null hypothesis was "significant at 5%." How many of such hypotheses have subsequently turned out to be false? Calibration might require 95%, but the few such efforts to assess accuracy in this way have been recent, and mainly appear in the medical literature. The personal view of probability would condemn them as not being sound statements of belief in the null hypotheses.

Euclidean geometry predates good mensuration by several centuries. It is to be hoped that it will not be necessary to wait that long before good assessment of beliefs fits with the other branch of mathematics, probability.

REFERENCES

1. Cooke, R. M. (1991). *Experts in Uncertainty: Opinions and Subjective Probability in Science*. Oxford University Press, Oxford.
2. De Finetti, B. (1974/5). *Theory of Probability: A Critical Introductory Treatment*. Wiley, London. (A translation from the Italian of the seminal text on the personal view of probability. Difficult, but rewarding, reading.)
3. Kahneman, D., Slovic, P., Tversky, A., eds. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, United Kingdom.
4. Lindley, D. V. (1990). The 1988 Wald memorial lectures: the present position in Bayesian statistics. *Statist. Sci.*, **5**, 44–89. (With discussion.)
5. Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Ass.*, **79**, 489–500.
6. Ramsey, F. P. (1926). Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*. Routledge & Kegan Paul, London.
7. Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Ass.*, **66**, 783–801.
8. Wright, G. and Ayton, P., eds. (1994). *Subjective Probability*. Wiley, Chichester. (A valuable collection of essays on many aspects, normative and descriptive, of subjective probability.)

See also BAYESIAN INFERENCE; BAYES' THEOREM; BELIEF FUNCTIONS; CALIBRATION—I; COHERENCE—I; ELICITATION; SCORING RULES; SUBJECTIVE PROBABILITIES; and SUBJECTIVE RANDOMNESS.

D. V. LINDLEY

ASSIGNABLE CAUSE

In model building, effects of certain factors ("causes") are allowed for in construction of the model. Ideally, all causes likely to have noticeable effects should be so represented. Such causes are often called "assignable causes." A better term might be "recognized causes."

Usually, there are, in fact, effects arising from causes that are not allowed for ("assigned") in the model. It is hoped that these will not be seriously large; they are supposed to be represented by random variation* included in the model.

Note that not all assignable causes may be actually used ("assigned") in the model. In the interests of simplicity, causes with recognized potential for effect may be omitted if the magnitudes of the effects are judged likely to be small.

See also STATISTICAL MODELING.

ASSOCIATION, MEASURES OF

Measures of association are numerical assessments of the strength of the statistical dependence of two or more qualitative variables. The common measures can be divided into measures for nominal polytomous variables* and measures for ordinal polytomous variables*.

MEASURES FOR NOMINAL VARIABLES

The most common measures of association for nominal variables are measures of prediction analogous in concept to the multiple correlation coefficient* of regression analysis*.

Consider two polytomous random variables X and Y with respective finite ranges I and J. A measure of prediction $\phi_{Y\cdot X}$ of Y given X depends on a measure Δ of the dispersion of a polytomous random variable. Such a measure is always nonnegative, with $\Delta_Y = 0$ if and only if Y is essentially constant; i.e., $\Delta_Y = 0$ if and only if for some $j \in J$, $p_{\cdot j} = \Pr[Y = j] = 1$. The measure does not depend on the labeling* of elements. Formally, one may require that $\Delta_Y = \Delta_{\sigma(Y)}$ if σ is a one-to-one transformation from J into a finite set J'. The added requirement is imposed that the conditional dispersion* $\Delta_{Y\cdot X}$ of Y given X not to exceed the unconditional dispersion Δ_Y of Y. Here $\Delta_{Y\cdot X}$ is the expected value of $\Delta_{Y\cdot X}(X)$, and $\Delta_{Y\cdot X}(i)$ is the dispersion of Y given that $X = i \in I$. [The definition of $\Delta_{Y\cdot X}(i)$ when $p_{i\cdot} = \Pr[X = i] = 0$ does not matter.] The measure of prediction

$$\phi_{Y\cdot X} = 1 - \Delta_{Y\cdot X}/\Delta_Y$$

compares the conditional dispersion of Y given X to the unconditional dispersion of Y, just as the multiple correlation coefficient* compares the expected conditional variance of the dependent variable to its unconditional variance. The measure $\phi_{Y\cdot X}$ is well defined if Y is not essentially constant. When $\phi_{Y\cdot X}$ is defined, $0 \leqslant \phi_{Y\cdot X} \leqslant 1$, with $\phi_{Y\cdot X} = 0$ if X and Y are independently distributed and $\phi_{Y\cdot X} = 1$ if X is an essentially perfect predictor* of Y, i.e., if for some function k from I to J, $\Pr[Y = k(X)] = 1$.

Two common examples of such measures of prediction are the λ coefficient of Guttman* [6] and of Goodman and Kruskal [1] and the τ-coefficient of Goodman and Kruskal* [1]. Let $p_{ij} = \Pr[X = i, Y = j]$, $i \in I, j \in J$. Then

$$\lambda_{Y\cdot X} = \frac{\left(\sum_{i\in I} \max_{j\in J} p_{ij} - \max_{j\in J} p_{\cdot j}\right)}{(1 - \max_{j\in J} p_{\cdot j})},$$

$$\tau_{Y\cdot X} = \frac{\left(\sum_{i\in I} \sum_{j\in J} p_{ij}^2/p_{i\cdot} - \sum_{j\in J} p_{\cdot j}^2\right)}{(1 - \sum_{j\in J} p_{\cdot j}^2)}.$$

In the last formula, 0/0 is defined as 0 to ensure that $p_{ij}^2/p_{i\cdot}$ is always defined.

In the case of $\lambda_{Y\cdot X}$, the measure $\Delta_Y = 1 - \max_{j\in J} p_{\cdot j}$ is the minimum probability of error from a prediction that Y is a constant k, while $\Delta_{Y\cdot X} = 1 - \sum_{i\in I} \max_{j\in J} p_{ij}$ is the minimum probability of error from a prediction that Y is a function $k(X)$ of X. In the case of $\tau_{Y\cdot X}$, $\Delta_Y = 1 - \sum_{j\in J} p_{\cdot j}^2$ is the probability that the random variable Y' does not equal Y, where Y' and Y are independent and identically distributed (i.i.d.). Similarly, $\Delta_{Y\cdot X} = 1 - \sum_{i\in I}\sum_{j\in J} p_{ij}^2/p_{i\cdot}$ is the probability that $Y' \neq Y$, where given X, Y and Y' are conditionally i.i.d.

Other measures of this type are also available. For example, Theil [18] has considered the measure

$$\eta_{Y\cdot X}$$
$$= -\sum_{i\in I}\sum_{j\in J} p_{ij}\log\left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right) \Bigg/ \sum_{j\in J} p_{\cdot j}\log p_{\cdot j},$$

based on the entropy* measure $\Delta_Y = -\sum_{j\in J} p_{\cdot j}\log p_{\cdot j}$ and the conditional entropy measure

$$\Delta_{Y\cdot X} = -\sum_{i\in I}\sum_{j\in J} p_{ij}\log(p_{ij}/p_{i\cdot}).$$

In these measures, $0/0 = 0$ and $0\log 0 = 0$.

The coefficient $\lambda_{Y\cdot X}$ has an attractively simple interpretation in terms of prediction*; however, $\lambda_{Y\cdot X}$ has the possible disadvantage that $\lambda_{Y\cdot X}$ may be 0 even if X and Y are dependent. In contrast, $\eta_{Y\cdot X}$ and $\tau_{Y\cdot X}$ are only 0 if X and Y are independent.

Partial and Multiple Association

As in Goodman and Kruskal [1,3], generalizations to cases involving three or more polytomous variables are straightforward. Consider a new polytomous variable W with finite range H. If $\Delta_{Y\cdot WX}$ denotes the conditional dispersion of Y given the polytomous vector (W, X), then the multiple association coefficient $\phi_{Y\cdot WX}$ may be defined as $1 - \Delta_{Y\cdot WX}/\Delta_Y$. The partial association of Y and W given X may then be defined as

$$\phi_{Y\cdot W|X} = 1 - \Delta_{Y\cdot WX}/\Delta_{Y\cdot X}.$$

Thus $\phi_{Y\cdot W|X}$ measures the additional predictive power of W given that X has already been used as a predictor of Y. If W and Y are conditionally independent given X, then $\phi_{Y\cdot W|X} = 0$. If X is not an essentially perfect predictor of Y but (W, X) is an essentially perfect predictor of Y, then $\phi_{Y\cdot W|X} = 1$. In general, if X is not an essentially perfect predictor of Y, one has $0 \leqslant \phi_{Y\cdot W|X} \leqslant 1$ and

$$1 - \phi_{Y\cdot WX} = (1 - \phi_{Y\cdot X})(1 - \phi_{Y\cdot Z|X}).$$

Symmetric Measures

A measure of prediction $\phi_{Y\cdot X}$ of Y given X is not generally equal to the corresponding measure $\phi_{X\cdot Y}$ for prediction of X by Y. This behavior can be contrasted with the square ρ^2 of the correlation coefficient* of two continuous random variables U and V. In the continuous case, ρ^2 measures the power of U as a predictor of V and the power of V as a predictor of U. In cases in which a symmetric measure is desired, Goodman and Kruskal [1] propose measures of the form

$$\phi_{XY} = 1 - (\Delta_{Y\cdot X} + \Delta_{X\cdot Y})/(\Delta_Y + \Delta_X).$$

For example,

$$\lambda_{XY} = \Big(\sum\nolimits_{i\in I} \max\nolimits_{j\in J} p_{ij} + \sum\nolimits_{j\in J} \max\nolimits_{i\in I} p_{ij}$$
$$- \max\nolimits_{j\in J} p_{\cdot j} - \max\nolimits_{i\in I} p_{i\cdot}\Big)$$
$$\cdot \left(2 - \max\nolimits_{j\in J} p_{\cdot j} - \max\nolimits_{i\in I} p_{i\cdot}\right)^{-1}.$$

The measure ϕ_{XY} is defined if either X or Y is not essentially constant. The coefficient ϕ_{XY} ranges between 0 and 1, with $\phi_{XY} = 0$ if X and Y are independent and $\phi_{XY} = 1$ if and only if for some functions k from I to J and m from J to I, $\Pr[Y = k(X)] = \Pr[X = m(Y)] = 1$.

Standardization

In some cases, it is desirable to standardize the marginal distributions* of X and Y before computation of a measure of association. For example, one may wish to find $\phi_{Y'\cdot X'}$, where X' has some standard reference distribution such as the uniform distribution* of I and the conditional distribution of Y' given that X' is identical to the conditional distribution of Y given X. If $p'_{i\cdot} = \Pr[X' = i]$, $i \in I$, and $p'_{\cdot j} = \sum_{i\in I}(p'_{i\cdot}/p_{i\cdot})p_{ij}, j \in J$, where $p_{i\cdot} = 0$ only when $p'_{i\cdot} = 0$, then

$$\lambda_{Y'\cdot X'} = \frac{\left(\sum_{i\in I}(\max_{j\in J} p_{ij})p'_{i\cdot}/p_{i\cdot} - \max_{j\in J} p'_{\cdot j}\right)}{(1 - \max_{j\in J} p'_{\cdot j})}.$$

Similarly, one may consider a measure $\phi_{Y^*\cdot X^*}$, where Y^* has the same standard marginal distributions and the conditional distribution of X^* given Y^* is the same as the conditional distribution of X given Y. More thorough standardization is also possible, as in Mosteller [13]. One may consider $\phi_{U\cdot V}$, where U and V have standard marginal distributions and $\Pr[U = i, V = j] = s_i t_j p_{ij}$, $i \in I$, $j \in J$, for some s_i, $i \in I$, and $t_j, j \in J$.

Optimal Prediction

Measures such as $\phi_{Y\cdot X}$ always have interpretations in terms of optimal prediction in the following sense. Some nonnegative and possibly infinite function $A_j(j, \mathbf{q})$ is defined for $j \in J$ and $\mathbf{q}$ in the simplex* S_J of vectors $\mathbf{q} = \langle q_j : j \in J\rangle$ with nonnegative coordinates with sum $\sum_{j\in J} q_j = 1$. This function represents a loss incurred if a probabilistic prediction $\mathbf{q}$ is made for Y and $Y = j$. The function is such that

$$d_J(\mathbf{q}) = \sum_{j\in J} q_j A_J(j, \mathbf{q}) \leqslant \sum_{j\in J} q_j A_J(j, \mathbf{q}') \qquad (\mathbf{q}, \mathbf{q}' \in S). \tag{1}$$

The dispersion Δ_Y of Y is $d_J(\mathbf{p}_Y)$, where $\mathbf{P}_Y = \langle p_{\cdot j} : j \in J\rangle$. Thus Δ_Y is the minimum expected loss achievable by prediction of Y without knowledge of X. Similarly, $\Delta_{Y\cdot X}(i)$ is $d_J(\mathbf{p}_{Y\cdot X}(i))$, where $\mathbf{p}_{Y\cdot X}(i) = \langle p_{ij}/p_{i\cdot} : j \in J\rangle$. Thus $\Delta_{Y\cdot X}(i)$ is the minimum expected loss achievable in prediction of Y, given that it is known that $X = i$, and $\Delta_{Y\cdot X}$ is the minimum expected loss achievable in prediction of Y given that X is known.

In the case of $\lambda_{Y\cdot X}$, one may define

$$A_J(j, \mathbf{q}) = \begin{cases} 1, & j \notin B(\mathbf{q}), \\ 1 - 1/m(\mathbf{q}), & j \in B(\mathbf{q}), \end{cases}$$

where $j \in B(\mathbf{q})$ if $q_j = \max_{k\in J} q_k$ and $m(\mathbf{q})$ is the number of elements in $B(\mathbf{q})$. In the typical case in which q_j has a unique maximum at a coordinate $j = \rho(\mathbf{q})$, the penalty $A_J(j, \mathbf{q})$ is 1 for $j \neq \rho(\mathbf{q})$ and 0 otherwise. In the case of $\tau_{Y\cdot X}$, $A_J(j, \mathbf{q})$ is the squared distance $\sum_{k\in K}(\delta_{kj} - q_k)^2$, where δ_{kj} is 1 for $k = j$ and δ_{kj} is 0 for $k \neq j$, while for $\eta_{Y\cdot X}$, one has

$$A_J(j, \mathbf{q}) = -\log q_j.$$

The loss function $A_J(j, \mathbf{q})$ is almost uniquely determined by the dispersion measure d_J. If $d_J(\alpha\mathbf{q})$ is defined to be $\alpha d_J(\mathbf{q})$ for $\alpha \geqslant 0$ and $\mathbf{q} \in S_J$, then d_J is a concave function on the set O_J of vectors $\mathbf{q} = \langle q_j : j \in J\rangle$ with all coordinates nonnegative. As in Savage [16], it follows that (1) is satisfied by $A_J(j, \mathbf{q})$, $j \in J$, $\mathbf{q} \in S_J$, if and only if for all $\mathbf{q} \in S_J$ and $\mathbf{q}' \in O_J$,

$$d_J(\mathbf{q}') \leqslant d_J(\mathbf{q}) + \sum_{j\in J} A_J(j, \mathbf{q})(q_j' - q_j).$$

As in Rockafellar [15], the vector $A_J(\mathbf{q}) = \langle A_J(j, \mathbf{q}) : j \in J\rangle$ is called a supergradient of d_J at $\mathbf{q}$. Some $A_J(\mathbf{q})$ exists at each $\mathbf{q} \in S_J$. If $\mathbf{q}$ is an element in the simplex S_J with all coordinates positive and if d_J is differentiable at $\mathbf{q}$, then $A_J(j, \mathbf{q})$ must be the partial derivative at $\mathbf{q}$ of d_J with respect to q_j. Thus the $A_J(j, \mathbf{q})$, $j \in J$, are uniquely determined and continuous at almost every point $\mathbf{q}$ on the simplex S_J. For example, in the case of $\eta_{Y\cdot X}$,

$$d_j(\mathbf{q}) = -\sum_{j\in J} q_j \log q_j + \left(\sum_{j\in J} q_j\right) \log \left(\sum_{j\in J} q_j\right)$$

for $\mathbf{q} \in O_J$. If all q_j are positive and $\sum_{j\in J} q_j = 1$, then $A(j, \mathbf{q}) = -\log q_j$ is the partial derivative at $\mathbf{q}$ of d_J with respect to q_j.

Estimation of Measures of Prediction

Typically bivariate measures of prediction in which standardization is not involved are estimated on the basis of a contingency table* $\mathbf{n} = \langle n_{ij} : i \in I, j \in J\rangle$ with a multinomial distribution of sample size N and probabilities $\mathbf{p} = \langle p_{ij} : i \in I, j \in J\rangle$. The estimate $\hat{\mathbf{p}} = N^{-1}\mathbf{n}$ of $\mathbf{p}$ is substituted for $\mathbf{p}$ in the formulas for $\phi_{Y\cdot X}$ and ϕ_{XY}. If $n_{\cdot j} = \sum_{i\in I} n_{ij}, j \in J$, $n_{i\cdot} = \sum_{j\in J} n_{ij}, \quad i \in I, \quad \hat{\mathbf{p}}_X = \langle N^{-1} n_{i\cdot} : i \in I\rangle, \hat{\mathbf{p}}_Y = \langle N^{-1} n_{\cdot j} : j \in J\rangle, \hat{\mathbf{p}}_{Y\cdot X}(i) = \langle n_{ij}/n_{i\cdot} : j \in J\rangle, i \in I$, and $\hat{\mathbf{p}}_{X\cdot Y}(j) = \langle n_{ij}/n_{\cdot j} : i \in I\rangle, j \in J$, then

$$\hat{\phi}_{Y\cdot X} = 1 - N^{-1} \sum_{i\in I} n_{i\cdot} d_J(\hat{\mathbf{p}}_{Y\cdot X}(i))/d_J(\hat{\mathbf{p}}_Y),$$

$$\hat{\phi}_{XY} = 1 - N^{-1} \times \frac{\left[\sum_{i\in I} n_{i\cdot} d_J(\hat{\mathbf{p}}_{Y\cdot X}(i)) + \sum_{j\in J} n_{\cdot j} d_I(\hat{\mathbf{p}}_{X\cdot Y}(j))\right]}{d_J(\hat{\mathbf{p}}_Y) + d_I(\hat{\mathbf{p}}_X)}.$$

For example,

$$\hat{\lambda}_{Y\cdot X} = \left(\sum_{i\in I} \max_{j\in J} n_{ij} - \max_{j\in J} n_{\cdot j}\right) \Big/ \left(N - \max_{j\in J} n_{\cdot j}\right).$$

Extensions to multivariate measures are straightforward.

Normal Approximation

Normal approximations* for distributions of measures such as $\hat{\phi}_{Y\cdot X}$ and $\hat{\phi}_{XY}$ are readily obtained as in Goodman and Kruskal [3,4]. Assume that d_J is differentiable at $\mathbf{p}_Y$ and at $\mathbf{p}_{Y\cdot X}(i)$ for $i \in I' = \{i \in I : p_{i\cdot} > 0\}$. [Alternatively, it suffices if $A(j, \cdot)$ is continuous at $\mathbf{p}_Y$ whenever $p_{\cdot j} > 0$ and $A(j, \cdot)$ is continuous at $p_{Y\cdot X}(i)$, $i \in I$, whenever $p_{ij} > 0$.] Assume that Y is not essentially constant. Then $N^{1/2}(\hat{\phi}_{Y\cdot X} - \phi_{Y\cdot X})$ has the large-sample distribution $N(0, \sigma^2(\phi_{Y\cdot X}))$, where $\sigma^2(\phi_{Y\cdot X})$ is the variance of the random variable

$$H(Y|X) = [(1 - \phi_{Y\cdot X}) d_J(Y, \mathbf{p}_Y) - d_J(Y, \mathbf{p}_{Y\cdot X}(X))]/d_J(\mathbf{p}_Y).$$

Since

$$E[H(Y|X)] = 0, \ \sigma^2(\phi_{Y\cdot X}) = \sum_{i\in I} \sum_{j\in J} p_{ij} [H(j|i)]^2.$$

For examples of formulas, see Goodman and Kruskal [3,4]. Assume, in addition, that d_I is differentiable at $\mathbf{p}_X$ and at $\mathbf{p}_{X\cdot Y}(j)$ for j such that $p_{\cdot j} > 0$. Thus $N^{1/2}(\hat{\phi}_{XY} - \phi_{XY})$ has large-sample distribution $N(0, \sigma^2(\phi_{XY}))$, where $\sigma^2(\phi_{XY})$ is the variance of

$$\begin{aligned} H(X, Y) = \{&(1 - \phi_{XY})[d_I(X, \mathbf{p}_X) + d_J(Y, \mathbf{p}_Y)] \\ &- d_I(X, \mathbf{p}_{X\cdot Y}(J)) - d_J(Y, \mathbf{p}_{Y\cdot X}(\boldsymbol{I}))\} \\ &\cdot [d_I(\mathbf{p}_X) + d_J(\mathbf{p}_Y)]^{-1}. \end{aligned}$$

Again $E[H(X, Y)] = 0$ and $\sigma^2(\phi_{XY}) = \sum_{i\in I} \sum_{j\in J} p_{ij} [H(i,j)]^2$. Since differentiability implies continuous differentiability in concave functions, $\sigma^2(\phi_{Y\cdot X})$ and $\sigma^2(\phi_{XY})$ possess consistent* estimates $\hat{\sigma}^2_{Y\cdot X}$ and $\hat{\sigma}^2_{XY}$ obtained by replacing p_{ij} by $\hat{p}_{ij} = N^{-1} n_{ij}$ in the relevant formulas. If $\sigma^2(\phi_{Y\cdot X}) > 0$, $0 < \alpha < 1$, and z_α is the upper $(\alpha/2)$-point of the $N(0, 1)$ distribution, then an approximate confidence interval* for $\phi_{Y\cdot X}$ of level α is

$$[\hat{\phi}_{Y\cdot X} - z_\alpha \hat{\sigma}(\phi_{Y\cdot X})/N^{1/2}, \hat{\phi}_{Y\cdot X} + z_\alpha \hat{\sigma}(\phi_{Y\cdot X})/N^{1/2}].$$

A similar argument applies to ϕ_{XY}.

Since $0 \leqslant \hat{\phi}_{Y\cdot X} \leqslant 1$, $\sigma(\phi_{Y\cdot X})$ must be 0 if a normal approximation applies and $\phi_{Y\cdot X}$ is 0 or 1. If all p_{ij} are positive and $\sigma(\phi_{Y\cdot X})$ is 0, then $\phi_{Y\cdot X} = 0$, for $\sigma_{Y\cdot X} = 0$ implies that $H(j|i)$ is always 0, so that

$$\begin{aligned} \Delta_Y &= \sum_{j\in J} p_{\cdot j} A(j, \mathbf{p}_Y) \\ &\leqslant \sum_{j\in J} p_{\cdot j} A(j, \mathbf{p}_{Y\cdot X}(i)) \\ &= (1 - \phi_{Y\cdot X}) \sum_{j\in J} p_{\cdot j} A(j, p_{\cdot j}) \\ &= (1 - \phi_{Y\cdot X}) \Delta_Y. \end{aligned}$$

In the special case of $\hat{\lambda}_{Y\cdot X}$, Goodman and Kruskal [3,4] note that $\sigma^2(\lambda_{Y\cdot X})$ is defined whenever $m(\mathbf{p}_Y)$ and $m(\mathbf{p}_{Y\cdot X}(i))$, $i \in I$, are all 1, and $\sigma^2(\lambda_{Y\cdot X}) = 0$ only if $\lambda_{Y\cdot X}$ is 0 or 1. The normal approximation always applies to the estimate $\hat{\tau}_{Y\cdot X}$ and $\hat{\eta}_{Y\cdot X}$; however, a simple necessary and sufficient condition for $\sigma^2(\tau_{Y\cdot X})$ or $\sigma^2(\eta_{Y\cdot X})$ to be 0 appears difficult to find.

Sampling by Rows

An alternative sampling problem has also been considered in Goodman and Kruskal [3,4], which is particularly appropriate for a standardized measure $\phi_{Y'\cdot X'}$ in which the conditional distribution of Y' given $X' = i \in \boldsymbol{I}$ is the same as the conditional distribution of Y given X and X' has a known marginal distribution with $p'_i = \Pr[X' = i]$. Let each row $\langle n_{ij} : j \in \boldsymbol{J}\rangle$, $i \in \boldsymbol{I}$, have an independent multinomial* distribution with sample size $N_i > 0$ and probabilities $\langle p_{ij}/p_{i\cdot} : j \in \boldsymbol{J}\rangle$. For simplicity, assume that each $p_{i\cdot}$ is positive. Let $N = \sum N_i$, let N_i/N approach $p_{i\cdot}$, and let

N approach infinity. Consider the standardized estimate

$$\hat{\phi}_{Y'\cdot X'} = 1 - \sum_{i\in I} p'_{i\cdot} d_J(\hat{\mathbf{p}}_{Y\cdot X}(i)) \Big/ d_J(\hat{\mathbf{p}}'_Y),$$

where $\hat{\mathbf{p}}'_Y = \langle p'_{\cdot j} : j \in \boldsymbol{J}\rangle$, $\hat{p}'_{\cdot j} = \sum_{i\in I} \cdot p'_{i\cdot} n_{ij}/N_i$, $j \in \boldsymbol{J}$, and $\hat{\mathbf{p}}_{Y\cdot X}(i) = \langle n_{ij}/N_i : j \in \boldsymbol{J}\rangle$ for $i \in \boldsymbol{I}$. Assume that d_J is differentiable at $\mathbf{p}'_Y = \langle p'_{\cdot j} : j \in \boldsymbol{J}\rangle$ and at $\mathbf{p}_{Y\cdot X}(i)$, $i \in \boldsymbol{I}$. Then $N^{1/2}(\hat{\phi}_{Y'\cdot X'} - \phi_{Y'\cdot X'})$ has an approximate $N(0, \sigma^2(\phi_{Y'\cdot X'}))$ distribution, where $\sigma^2(\phi_{Y'\cdot X'})$ is the expected conditional variance

$$\sum_{i\in I}\sum_{j\in J} p_{ij}[H'(j|i)]^2 - \sum_{i\in I}\left[\sum_{j\in J} H'(j|i)p_{ij}\right]^2 \Big/ p_{i\cdot}$$

of $H'(Y|X)$ given X. Here for $i \in I, j \in J$,

$$H'(j|i) = (p'_{i\cdot}/p_{i\cdot})[(1 - \phi_{Y'\cdot X'})A_J(Y, \mathbf{p}'_Y) - A_J(Y, \mathbf{p}_{Y\cdot X}(i))].$$

In the special case $p'_{i\cdot} = p_{i\cdot}$, one has $\phi_{Y'\cdot X'}$ equal to the unstandardized coefficient $\phi_{Y\cdot X}$, $H'(j|i) = H(j, i)$, and $\sigma^2(\phi_{Y'\cdot X'}) \leqslant \sigma^2(\phi_{Y\cdot X})$.

Clearly, $\sigma^2(\phi_{Y'\cdot X'}) = 0$ if $\phi_{Y'\cdot X'}$ is 0 or 1. In the case of $\lambda_{Y'\cdot X'}$, $\sigma^2(\lambda_{Y'\cdot X'}) = 0$ if and only if $\phi_{Y'\cdot X'}$ is 0 or 1, as noted in Goodman and Kruskal [4]. More generally, $\sigma^2(\phi_{Y'\cdot X'}) = 0$ implies that $\phi_{Y'\cdot X'} = 0$ if all probabilities p_{ij} are positive. The proof is only slightly changed from the corresponding proof for $\sigma_{Y\cdot X}$.

Older Measures

Numerous older measures of association between nominal variables are reviewed by Goodman and Kruskal [1,2] and by Kendall and Stuart [10, pp. 556–561]. The most common are based on the chi-square* statistic. They include the mean square contingency

$$\phi^2 = \sum_{i\in I}\sum_{j\in J} (p_{ij} - p_{i\cdot}p_{\cdot j})^2/(p_{i\cdot}p_{\cdot j})$$

and the coefficient of contingency $C = [\phi^2/(1 + \phi^2)]^{1/2}$ of Pearson [14] and Tschuprow's [19] coefficient $T = [\phi^2/\nu^{1/2}]^{1/2}$. In the last expression $\nu = (r - 1)(s - 1)$, I has r elements, and J has s elements. These measures lack the functional interpretations available in the case of $\phi_{Y\cdot X}$ or ϕ_{XY}.

MEASURES OF ASSOCIATION FOR ORDINAL POLYTOMOUS VARIABLES

The most commonly used measure of association for ordinal polytomous variables is the γ coefficient of Goodman and Kruskal [1–5]. Assume that the ranges I of X and J of Y are well ordered, so that if i and i' are in I, then $i < i'$, $i = i'$, or $i > i'$ and if j and j' are in J, then $j < j'$, $j = j'$, or $j > j'$. Let (X_1, Y_1) and (X_2, Y_2) be independently distributed pairs with the same distribution as (X, Y). Let $C = 2\Pr[X_1 > X_2 \text{ and } Y_1 > Y_2]$ be the probability that either $X_1 > X_2$ and $Y_1 > Y_2$ or $X_1 < X_2$ and $Y_1 < Y_2$, so that (X_1, Y_1) and (X_2, Y_2) are concordant. Let $2D = 2\Pr[X_1 > X_2 \text{ and } Y_2 > Y_1]$ be the probability that either $X_1 > X_2$ and $Y_1 < Y_2$ or $X_1 < X_2$ and $Y_1 > Y_2$, so that (X_1, Y_1) and (X_2, Y_2) are discordant. Then

$$\gamma_{XY} = \frac{C - D}{C + D}.$$

The coefficient is defined if $p_{ij} > 0$ and $p_{i'j'} > 0$ for some $i, i' \in I$ and $j, j' \in J$ with $i \neq i'$ and $j \neq j'$. One has $-1 \leqslant \gamma_{XY} \leqslant 1$, with $\gamma_{XY} = 0$ under independence of X and Y. For γ_{XY} to be 1, the nonzero p_{ij} must have an ascending staircase pattern, so that if $i < i'$, $p_{ij} > 0$, and $p_{i'j'} > 0$, then $j \leqslant j'$, while if $j < j'$, $p_{ij} > 0$, and $p_{i'j'} > 0$, then $i \leqslant i'$. Similarly, γ_{XY} can only be -1 if the nonzero p_{ij} have a descending staircase pattern. In the special case in which I and J have the two elements 1 and 2, γ_{XY} is the *coefficient of association* $(p_{11}p_{22} - p_{12}p_{21})/(p_{11}p_{22} + p_{12}p_{21})$ of *Yule* [20]. In this special case, $\gamma_{XY} = 0$ if and only if X and Y are independent, $\gamma_{XY} = 1$ only if $\Pr[X = Y] = 1$, and $\gamma_{XY} = -1$ only if $\Pr[X = Y] = 0$.

The measure γ_{XY} only considers pairs (X_1, Y_1) and (X_2, Y_2) in which $X_1 \neq X_2$ and $Y_1 \neq Y_2$. An alternative approach by Somers [17] considers all pairs with just $X_1 \neq X_2$. One obtains the asymmetric coefficient

$$\gamma_{Y\cdot X} = (C - D)/\Pr[X_1 \neq X_2].$$

Again $-1 \leqslant \gamma_{Y\cdot X} \leqslant 1$, with $\gamma_{Y\cdot X} = 0$ if X and Y are independent. The coefficient $\gamma_{Y\cdot X}$ can only be -1 (or 1) if γ_{XY} is -1 (or 1) and if for each $j \in J$, p_{ij} is positive for no more than one $i \in \boldsymbol{I}$. For further variants on γ_{XY}, see Kendall and Stuart [10, pp. 561–565].

Estimation of γ_{XY} and $\gamma_{Y\cdot X}$ is straightforward given a table $\mathbf{n} = \langle n_{ij} : i \in I, j \in J\rangle$ with a multinomial distribution with sample size N and probabilities $\mathbf{p} = \langle p_{ij} : i \in I, j \in J\rangle$. Let $\hat{C}$ be the sum of all products $2n_{ij}n_{i'j'}/N^2$ such that $i < i'$ and $j < j'$, and let $\hat{D}$ be the sum of all products $2n_{ij}n_{i'j'}/N^2$ such that $i < i'$ and $j < j'$. Then $\hat{\gamma}_{XY} = (\hat{C} - \hat{D})/(\hat{C} + \hat{D})$ and $\hat{\gamma}_{Y\cdot X} = (\hat{C} - \hat{D})/[1 - \sum_{i\in I}(n_{i\cdot}/N)^2]$. As noted in Goodman and Kruskal [3,4] $N^{1/2}(\hat{\gamma}_{XY} - \gamma_{XY})$ has an approximate $N(0, \sigma^2(\phi_{XY}))$ distribution, with

$$\begin{aligned}
\sigma^2(\phi_{XY}) &= \frac{16}{(C+D)^4}\sum_{i\in I}\sum_{j\in J} p_{ij}(CS_{ij} - DR_{ij})^2 \\
&\leqslant 2(1 - \gamma_{XY}^2)/(C+D), \\
S_{ij} &= \Pr[X > i \text{ and } Y < j] \\
&\quad + \Pr[X < i \text{ and } Y > j], \\
R_{ij} &= \Pr[X > i \text{ and } Y > j] \\
&\quad + \Pr[X < i \text{ and } Y < j].
\end{aligned}$$

Similarly, $N^{1/2}(\hat{\gamma}_{Y\cdot X} - \gamma_{Y\cdot X})$ has an approximate $N(0, \sigma^2(\gamma_{Y\cdot X}))$ distribution with $E = \Pr[X_1 \neq X_2, Y_1 \neq Y_2]$ and

$$\begin{aligned}
\sigma^2(\gamma_{Y\cdot X}) &= \frac{4}{(C+D+E)^4}\sum_{i\in I}\sum_{j\in J} p_{ij} \\
&\quad \cdot [(C-D)(1-p_{i\cdot}) \\
&\quad - (C+D+E)(S_{ij} - R_{ij})]^2.
\end{aligned}$$

One has $\sigma^2(\gamma_{XY}) = 0$ if $|\gamma_{XY}| = 1$ and $\sigma^2(\gamma_{Y\cdot X}) = 0$ if $|\gamma_{Y\cdot X}| = 1$. If all p_{ij} are positive, then $\sigma^2(\gamma_{XY}) > 0$. In the special case of $\gamma_{XY} = 0$, one has $\gamma_{Y\cdot X} = 0$,

$$\begin{aligned}
\sigma^2(\gamma_{XY}) &= \frac{4}{(C+D)^2}\sum_{i\in I}\sum_{j\in J} p_{ij}(S_{ij} - R_{ij})^2, \\
\sigma^2(\gamma_{Y\cdot X}) &= \frac{4}{(C+D+E)^2}\sum_{i\in I}\sum_{j\in J} p_{ij} \\
&\quad \times (S_{ij} - R_{ij})^2.
\end{aligned}$$

Kendall's τ*, Spearman's ρ_S*, and Goodman and Kruskal's γ*

In contrast to the nominal measures of association, Goodman and Kruskal's γ coefficient remains well-defined if the respective ranges I and J of X and Y are infinite and $\Pr[X = i] = \Pr[Y = j] = 0$ for any $i \in I$ and $j \in J$. The coefficient γ_{XY} is then Kendall's [8] τ measure $\tau_k = C - D$. It remains true that $-1 \leqslant \tau_k \leqslant 1$, with $\tau_k = 0$ when X and Y are independent. Estimation of τ_k is, however, best described in terms of independent pairs (X_t, Y_t), $1 \leqslant t \leqslant N$, with common distribution (X, Y). Then *Kendall's* τ statistic $\hat{\tau}_k$ is $2(N_c - N_d)/[N(N-1)]$, where there are N_c s and t with $1 \leqslant s < t \leqslant N$ such that (X_s, Y_s) and (X_t, Y_t) are concordant ($X_s < X_t$ and $Y_s < Y_t$ or $X_t < X_s$ and $Y_t < Y_s$) and there are N_d s and t with $1 \leqslant s < t \leqslant N$ such that (X_s, Y_s) and (X_t, Y_t) are discordant ($X_s < X_t$ and $Y_s > Y_t$ or $X_t < X_s$ and $Y_t > Y_s$). As N becomes large $N^{1/2}(\hat{\tau}_k - \tau_k)$ has an approximate $N(0, \sigma^2(\tau_k))$ distribution. *See* KENDALL'S TAU. Here $\sigma^2(\tau_k) = 16(F - C^2)$ and

$$\begin{aligned}
F &= \Pr[X_1 > X_2, X_1 > X_3, Y_1 > Y_2, Y_1 > Y_3] \\
&\quad + \Pr[X_1 < X_2, X_1 < X_3, Y_1 < Y_2, Y_1 < Y_3]
\end{aligned}$$

is the probability that both (X_2, Y_2) and (X_3, Y_3) are concordant with (X_1, Y_1). If X and Y are independent, then $\sigma^2(\tau_k) = 4/9$. See Hoeffding [7] and Kruskal [12] for details.

Closely related to Kendall's τ is the *Spearman rank correlation coefficient** r_s. Assume that all X_t are distinct and all Y_t are distinct. Let R_t be the number of s with $X_s \leqslant X_t$, and let S_t be the number of t with $Y_s \leqslant Y_t$. Then r_s is the sample correlation of the pairs (R_t, S_t), $1 \leqslant t \leqslant N$. The statistic r_s provides a measure

$$\rho_s = 6\{\Pr[X_1 > X_2, Y_1 > Y_3] - 1\}$$

of the probability that (X_1, Y_1) and (X_2, Y_3) are concordant. An alternative unbiased estimate of ρ_s is $\hat{\rho}_s = [(n+1)/(n-2)]r_s - [3/(n-2)]\hat{\tau}_k$, which has been termed the *unbiased grade coefficient* by Konijn [11]. One has $-1 \leqslant \rho_s \leqslant 1$, with $\rho_s = 0$ under independence of X and Y. Both $N^{1/2}(\hat{\rho}_s - \rho_s)$ and $N^{1/2}(r_s - \rho_s)$ have limiting distribution $N(0, \sigma^2(\rho_s))$ as N becomes large. The formula for $\sigma^2(\rho_s)$ has been given by Hoeffding [7]. Since it is somewhat complicated, it will be omitted here in the general case. Under independence of X and Y, $\sigma^2(\rho_s) = 1$.

For further details concerning these and related measures, see Kendall [9] and

Kruskal [12]. (*See* also KENDALL'S TAU—I, SPEARMAN RANK CORRELATION COEFFICIENT, and GOODMAN–KRUSKAL TAU AND GAMMA)

NUMERICAL EXAMPLE

To illustrate results, consider Table 1, which can be found in Goodman and Kruskal [1], among many other references. Let X refer to eye color and let Y refer to hair color.

Some resulting estimated measures of association are listed in Table 2. In the case of the measures for ordered variables, eye color is ordered from blue to brown and hair color is ordered from fair to black.

Asymptotic standard deviations are based on the assumption that the counts in Table 1 have a multinomial distribution. The asymmetry in X and Y and the large variations in the sizes of measures are to be expected. As noted as early as Goodman and Kruskal [1], instincts developed from regression analysis* are not necessarily appropriate for assessment of the size of measures of association for ordinal or nominal polytomous variables.

REFERENCES

1. Goodman, L. A. and Kruskal, W. H. (1954). *J. Amer. Statist. Ass.*, **49**, 732–764. (Goodman and Kruskal's λ, τ, and γ are defined and the basic criteria for measures of association are presented.)
2. Goodman, L. A. and Kruskal, W. H. (1959). *J. Amer. Statist. Ass.*, **54**, 123–163. (A valuable historical review and bibliography are provided.)
3. Goodman, L. A. and Kruskal, W. H. (1963). *J. Amer. Statist. Ass.*, **58**, 310–364. (Asymptotic distributions are obtained for estimates of Goodman and Kruskal's λ, τ, and γ. Further results appear in their 1972 paper.)
4. Goodman, L. A. and Kruskal, W. H. (1972). *J. Amer. Statist. Ass.*, **67**, 415–421.
5. Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classifications*. Springer-Verlag, New York. (A volume consisting of Goodman and Kruskal's four papers on measures of association, indicated above.)
6. Guttman, L. (1941). In *The Prediction of Personal Adjustment* (Bull. 48), P. Horst et al., eds. Social Science Research Council, New York, pp. 253–318.
7. Hoeffding, W. (1948). *Ann. Math. Statist.*, **19**, 293–325. (The methods presented here apply to numerous problems involving ordinal data.)
8. Kendall, M. G. (1938). *Biometrika*, **30**, 277–283.
9. Kendall, M. G. (1962). *Rank Correlation Methods*, 3rd ed. Charles Griffin, London. (A standard text on rank correlation statistics.)
10. Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*, 2nd ed. Vol. 2. Charles Griffin, London.

Table 2. Estimated Measures of Association for Table 1

Measure	Estimate	Estimated Asymptotic Standard Deviation of Estimate
$\lambda_{Y\cdot X}$	0.192	0.012
$\lambda_{X\cdot Y}$	0.224	0.013
λ_{XY}	0.208	0.010
$\tau_{Y\cdot X}$	0.081	0.005
$\tau_{X\cdot Y}$	0.089	0.005
τ_{XY}	0.085	0.005
$\eta_{Y\cdot X}$	0.075	0.004
$\eta_{X\cdot Y}$	0.085	0.005
η_{XY}	0.080	0.004
γ_{XY}	0.547	0.013
$\gamma_{X\cdot Y}$	0.346	0.009
$\gamma_{Y\cdot X}$	0.371	0.010

Table 1.

	Hair Color Group				
Eye Color Group	Fair	Red	Brown	Black	Total
Blue	1768	47	807	189	2811
Gray or green	946	53	1387	746	3132
Brown	115	16	438	288	857
Total	2829	116	2632	1223	6800

11. Konijn, H. S. (1956). *Ann. Math. Statist.*, **27**, 300–323.
12. Kruskal, W. H. (1958). *J. Amer. Statist. Ass.*, **53**, 814–861. (A very helpful review.)
13. Mosteller, F. (1968). *J. Amer. Statist. Ass.*, **63**, 1–28.
14. Pearson, K. (1904). Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation. *Drapers' Company Res. Mem., Biometric Ser. 1.* (Pearson introduces classical association measures related to the chi-square statistic* and relates them to the bivariate normal distribution.*)
15. Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, N.J.
16. Savage, L. J. (1971). *J. Amer. Statist. Ass.*, **66**, 783–801.
17. Somers, R. H. (1962). *Amer. Soc. Rev.*, **27**, 799–811.
18. Theil, H. (1970). *Amer. J. Sociol.*, **76**, 103–154.
19. Tschuprow, A. A. (1925). *Grundbegriffe und Grundprobleme der Korrelationstheorie*. Teubner, Leipzig.
20. Yule, G. U. (1900). *Philos. Trans. R. Soc. Lond. A*, **194**, 257–319. (A valuable early paper on measurement of association of dichotomous variables.)

See also Categorical Data; Correlation; Dependence, Measures and Indices of; Goodman–Kruskal Tau and Gamma; and Log-linear Models in Contingency Tables.

S. J. Haberman

ASTRONOMY, STATISTICS IN

Perhaps more than other physical sciences, astronomy is frequently statistical in nature. The objects under study are inaccessible to direct manipulation in the laboratory. The astronomer is restricted to observing a few external characteristics of objects populating the Universe, and inferring from these data their properties and underlying physics. From the seventeenth through nineteenth centuries, European astronomers were engaged in the application of Newtonian theory to the motions of bodies in the solar system. This led to discussions of the statistical treatment of scientific data, and played a critical role in the development of statistical theory. The twentieth century has seen remarkable success in the applications of electromagnetism and quantum mechanics* to heavenly bodies, leading to a deep understanding of the nature and evolution of stars, and some progress in understanding galaxies and various interstellar and intergalactic gaseous media. Statistical theory has played a less important role in these advances of modern astrophysics. However, the last few years have seen some reemergence of interest in statistical methodology to deal with some challenging data analysis problems. Some examples of these contemporary issues are presented.

EARLY HISTORY

Celestial mechanics in the eighteenth century, in which Newton's law of gravity was found to explain even the subtlest motions of heavenly bodies, required the derivation of a few interesting quantities from numerous inaccurate observations. As described in detail by Stigler [40], this required advances in the understanding of statistical inference* and error distributions. Mayer, in his 1750 study of lunar librations, suggested a procedure of reconciling a system of 27 inconsistent linear equations in three unknowns by solving the equations in groups. Laplace*, in a 1787 analysis of the influence of Jupiter's gravity on Saturn's motion, suggested a more unified approach that led to Legendre's invention of the least-squares method in an 1805 study of cometary orbits. Shortly thereafter, in an 1809 monograph on the mathematics of planetary orbits, Gauss* first presented the normal (or Gaussian) distribution of errors in overdetermined systems of equations using a form of Bayes' theorem*, though the actual derivation was flawed.

Many other individuals also contributed substantially to both astronomy and statistics [40,14]. Galileo* gave an early discussion of observational errors concerning the distance to the supernova of 1572. Halley, famous for his early contributions to celestial mechanics, laid important foundations to mathematical demography and actuarial science. Bessel, codiscoverer of stellar parallaxes and the binary companion of Sirius, introduce the notion of probable error*. Adolphe Quetelet*, founder of the Belgian

Royal Observatory, led the application of probability theory to the social sciences. Airy, a Royal astronomer, is known both for his text on statistical errors and his study of telescope optics. *See also* LAWS OF ERROR—I, II, III.

STATISTICAL ASTRONOMY

As comprehensive all-sky catalogs of star positions were compiled and astronomical photography permitted faint stars to be located, a field known as "statistical astronomy" rose to importance in the first half of this century. It is concerned with various collective properties of stars including their luminosity and mass distribution functions, their spatial distribution in the Milky Way, their distances from us and the related problem of light absorption in the interstellar medium, and their motions with respect to the Sun and to the center of the galaxy. The principal results of these studies are discussed in the monumental 1953 monograph of Trumpler and Weaver [41]. Prephotographic statistical discussions are reviewed in ref. 37 and more recent findings can be found in refs. 38 and 12.

From these studies we have learned that the Sun resides about 25 thousand light-years off-center in a disk of a differentially rotating spiral galaxy, with stars of increasing ages occupying the galaxy's spiral arms, smooth disk, and halo. By comparing the galactic mass inferred from star counts with that inferred from their rotational velocities around the galactic center, the existence of a "dark matter" component in the outer regions of the galaxy is inferred. Dynamical studies of other galaxies confirm that the mass in visible stars and gas is dwarfed by the dark matter in galaxy halos; yet astronomers do not know whether the matter is in the form of planets, elementary particles, black holes, or some more exotic form.

We give two modern examples of studies in galactic astronomy. Murray [25] derives the joint densities of the observed parallaxes, proper motions, and brightnesses for 6,125 stars, and computes the luminosity function, scale height in the galactic disk, streaming motions, and outliers with high velocities. A maximum-likelihood technique is used; the principal limitation is that parametric (e.g., Gaussian luminosity functions and observational errors, exponential scale heights for stars in the disk) forms are assumed throughout. Caldwell and Ostriker [8] seek fits of a three-component model of the mass distribution of the galaxy to 14 observationally derived quantities constraining the size, density, and motions in the galaxy. A nonlinear least-squares minimization algorithm is used to find the minimum chi-squared* solution.

Perhaps the central problem in statistical astronomy is the derivation of the *intrinsic* luminosity distribution function of a class of stars from a survey of the stars with the greatest *apparent* brightnesses (usually called a magnitude- or flux-limited survey). The observed population contains an excess of high luminosity stars, which can be seen out to large distances, and a deficit of low luminosity stars, which are bright enough to appear in the sample only if they lie close to us. Intrinsic or experimental errors scatter faint stars preferentially into flux-limited samples. These and related problems are called the "Malmquist effects," after the Swedish astronomer who derived a correction in 1920 for the bias for Gaussian luminosity distributions.

Interest in luminosity functions reemerged during the last decade with attempts to understand the phenomenon of quasar evolution [42]. The density of quasars, which are galaxies possessing extremely violent activity in their nuclei probably due to an accreting massive black hole, per unit volume was thousands of times higher when the Universe was young than it is today, indicating evolution of the shape and/or amplitude of the luminosity function. In 1971, Lynden-Bell [21] derived a remarkably simple nonparametric maximum likelihood* procedure for extrapolating from a flux-limited data set (assumed to be randomly truncated) of quasar counts to obtain a complete luminosity function. The method was largely unused by astronomers, and unknown to statisticians until its reexamination in 1985 by Woodroofe [43]. A similar method was independently developed by Nicoll and Segal [30] in support of Segal's proposed chronometric cosmology. A more common approach to the quasar evolution problem is to fit the data to parametric evolution formulas using least-squares* or maximum-likelihood criteria; see, for example, Marshall [23].

STATISTICAL ANALYSIS OF MODERN ASTRONOMICAL DATA

The modern observational astronomer is typically schooled only in elementary techniques such as the chi-squared test* and least-squares regression, using computer software such as that given by Bevington [6]. Some nonparametric methods such as Kolmogorov–Smirnov* and rank correlation* tests have come into frequent use, and computer codes distributed by Press et al. [32] are likely to bring other methods into the astronomer's repertoire. Unfortunately, very few astronomers are familiar with the major statistical software* packages such as SAS or BMDP.

Interest in more sophisticated and specialized statistical techniques of data analysis has emerged during the last decade. Murtagh and Heck [26] have written a monograph on multivariate techniques for astronomers, with a thorough bibliography of astronomical applications. The proceedings of a 1983 conference devoted specifically to the subject of statistical methods in astronomy is available [39]. An informal international Working Group for Modern Astronomical Methodology has formed, with a newsletter and announcements of workshops published in ref. 7. Articles in the statistical literature include those of Kendall [19], Scott [36], and Narlikar [28]. Following are brief descriptions of a few topics of current interest.

*Censored data**. A common experiment in astronomy entails the observations of a preselected sample of objects at a new spectral band, but some of the objects are not detected. For example, only about 10% of optically selected quasars are detected with the most sensitive radio telescopes and about 50% with the most sensitive satellite-borne X-ray telescopes, unless unacceptably long exposure times are devoted to the experiment. The data thus suffer type I left-censoring in apparent brightness, and a quasi-random censoring in intrinsic luminosities because the quasars lie at different distances. The studies seek to measure the mean radio or X-ray luminosities of the sample, differences in the luminosity functions between subsamples, correlations and linear regressions between luminosities in the various spectral bands, and so forth.

Until recently astronomers were unaware that survival analysis* statistical methods used by biostatisticians and others were available that provide many solutions to these questions. Avni, an astrophysicist, independently derived the "redistribute-to-the-right" formulation of the Kaplan–Meier product-limit estimator* [3], and a maximum-likelihood linear regression* assuming normally distributed residuals [2]. Schmitt [35] has developed a linear regression procedure, based on the two-dimensional Kaplan–Meier estimator and bootstrap error analysis, which can be applied to data censored in both the dependent and independent variables. Schmitt, Isobe, and Feigelson have brought previously known survival analysis methods into use for astronomical applications [35,10,18,17]. The principal difficulties in adapting standard survival analysis to astronomical situations are that the censoring levels are usually not known precisely, and the censoring patterns in flux-limited data are usually not completely random. These problems have yet to be addressed in detail.

Spatial analysis of galaxy clustering. Our Milky Way is but one of 10^{11} detectable galaxies, each containing up to $\sim 10^{11}$ luminous stars and a greater but uncertain amount of nonluminous matter. The galaxies appear to be rushing away from each other in a universal expansion that started 10–20 billion years ago. Correct modeling of the spatial and velocity distribution of the visible galaxies is critical for understanding the history of the Universe, and for determining whether or not the Universe will cease expanding and recollapse. (Poisson point process*). Other statistical analyses followed including the nearest-neighbor* distance distribution, the multiplicity function, and, most extensively, the two- and three-point correlation functions [31]. The power law two-point correlation function found for galaxies could be explained as the result of simple gravitational interactions of matter in an initially homogeneous Universe with an appropriate spectrum of density fluctuations.

Galaxies are not distributed randomly but are strongly clustered on many spatial scales. Neyman and Scott [29,36] were among the first to study this during the 1950s, using a hierarchy of clusters distributed as a uniform Poisson point process*.

By the middle 1980s, however, surveys of thousands of galaxy recessional velocities had been completed, revealing an unexpectedly anisotropic clustering pattern, as illustrated in Fig. 1 [9]. The galaxies seem to be concentrated along the edges of giant shells. The effect is sometimes referred to as "filaments" or "sheets" surrounding "voids," with a "spongelike" topology. The largest structures may exceed one-tenth the size of the observable Universe. In addition, large regions of the Universe may be moving in bulk with respect to other regions [22]. None of these phenomena can be easily explained by simple gravitational effects in an initially homogeneous Universe. Statistical modeling of these data has just begun. Suggested methods include minimal spanning trees* [5], random walks* [20], ridge-finding algorithms [24], measures of topological curvature [13], and a quadrupole elongation statistic [11]. The study of Lynden-Bell et al. [22] is also of methodological interest, illustrating how sophisticated maximum-likelihood models of galaxy location and motions have become.

Analysis of periodic time series. Stellar systems often exhibit periodic behavior, from vibrations or rotations of single stars to orbits of two or more bodies in mutual gravitational orbits. Time scales run from milliseconds to hundreds of years, and the data can involve any portion of the electromagnetic spectrum. For example, time series of X-ray emission from binary star systems where one companion of an accreting neutron star or

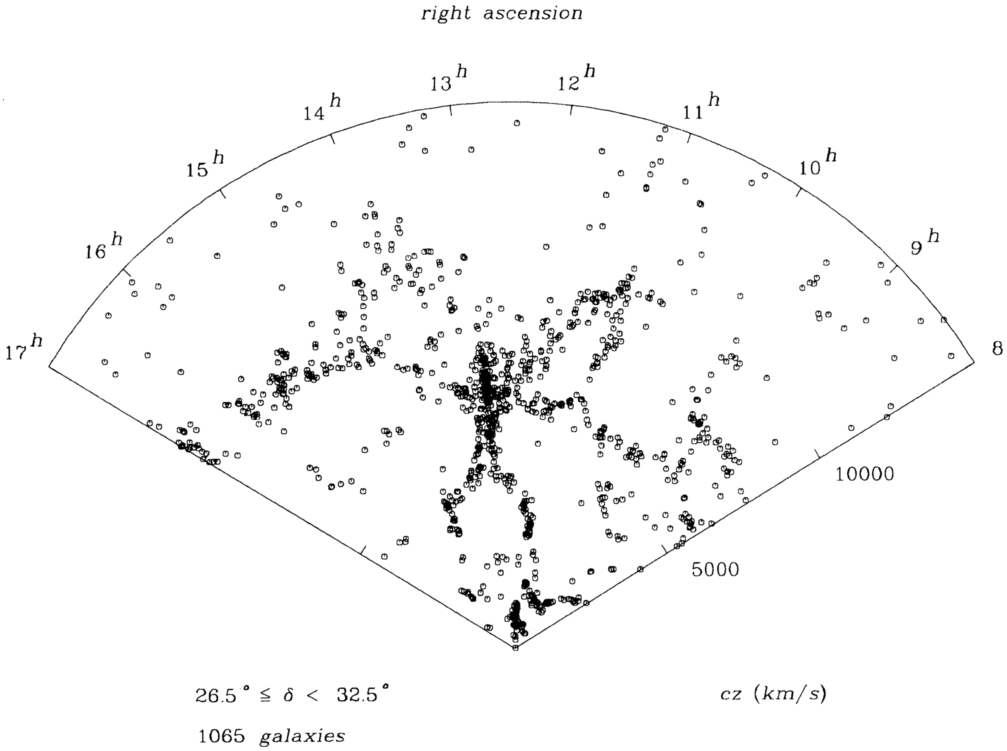


Figure 1. A "slice of the Universe" showing the spatial distribution of bright galaxies in a portion of the sky [9]. It shows the strongly anisotropic clustering pattern of "filaments" surrounding "voids." Courtesy of M. Geller, Center for Astrophysics.

black hole are quite interesting. Some systems show one or more strict periodicities, others random shot noise or white noise, others occasional sudden bursts of x-rays, and yet others combinations of 1/*f* noise and quasi-periodic oscillations. Certain of these behaviors are understood as rotational or orbital effects, whereas others are still mysterious.

A time-series* problem that has attracted recent methodological interest is the difficulty of establishing the existence of binary star orbits from optical photometric or velocity time series, which often consist of sparse, unevenly spaced, and noisy data. The issue is important because as many as half of the "stars" one sees at night may in fact be wide binaries. Classical Fourier analysis*, which assumes evenly spaced data points, is not an adequate method, and a variety of alternatives have been proposed. One is the periodogram*, which can be equivalent to least-squares fitting of sine curves to the data [34,16]. Several nonparametric methods have also been evaluated [15]. The principal difficulty with all methods is usually not finding the highest peak in the spectrum, but in evaluating its statistical significance and eliminating false alarms. Both analytical and bootstrap methods for measuring confidence limits have been suggested, but no consensus has emerged on the best treatment of the problem.

Image restoration techniques. Many classes of objects in the sky have complicated morphologies, and much effort is devoted to imaging them accurately. This entails compensation for the distortions caused by imperfect optics of the telescope or turbulence in the Earth's atmosphere, and nonlinearities in detectors such as photographic plates. Perhaps the greatest need for sophisticated image restoration* is in "aperture synthesis" interferometry, a technique developed in radio astronomy that combines the signals from separated individual radio antennas to produce a single high-resolution image [33,27]. The data consist of the two-dimensional Fourier transform of the brightness distribution in the sky, but the coverage in the Fourier plane is incomplete. Simple Fourier transformation thus gives an image contaminated by strong side lobes. Two restoration methods are commonly used: The "CLEAN" algorithm, which gives a least-squares fit to a superposition of many point sources; and the maximum entropy method*, which gives the most probable nonnegative smooth image consistent with the data. Both require prior knowledge of the coverage in the Fourier plane. The resulting CLEANed map can then be used as an improved model of the sky distribution to "self-calibrate" unpredicted instrumental or atmospheric disturbances. After many iterations, which can take hundreds of CPU-hours on large computers, images with extraordinarily high fidelity have been achieved with dynamic range (brightest spot in the image divided by the root mean square noise level) of order $10^5 : 1$. Maximum entropy image enhancement techniques are sometimes also used in optical and X-ray astronomy as well as radio interferometry.

Statistics of very few events. Important branches of astronomy have emerged in the last few decades from experimental physics that involve the detection of small numbers of discrete events. These include cosmic ray, X-ray, gamma-ray, and neutrino astronomy. With the proliferation of photon-counting detectors like charged-coupled devices, such data are becoming increasingly common in optical astronomy as well. The statistical procedures for interpreting these data are traditionally based on the Poisson distribution, though use of nonparametric and maximum likelihood techniques is also appearing.

A premier example of this problem was the detection of neutrinos from the supernova SN1987A, initiated by the gravitational collapse of a star in a satellite galaxy of our Milky Way. The detection represented the first astronomical object other than the Sun ever detected in neutrinos, and provides a unique opportunity to test a host of models of high-energy and nuclear astrophysics. The data consist of 11 events in 13 seconds in the Kamiokande-II underground detector, and 8 events in 6 seconds in the IMB detector. The timing of the events constrain the mass of the neutrino, which could be the most massive component of the Universe [4,1]. However, different treatments of the same data give considerably different neutrino mass limits.

SUMMARY

Many aspects of modern astronomy are statistical in character, and demand sophisticated statistical methodology. In light of this situation, and the rich history of the interaction between astronomy and statistics through the nineteenth century, it is surprising that the two communities have been so isolated from each other in recent decades. Astronomers dealing with censored data, for example, were unaware of the relevant progress in biostatistics* and industrial reliability applications until recently. Most are not familiar with multivariate techniques extensively used in the social sciences. Conversely, statisticians were unaware of Lynden-Bell's maximum-likelihood density estimation technique for a truncated distribution. There is little communication between astronomers analyzing the spatial distribution of galaxies (Fig. 1) and statisticians involved with point spatial processes* arising in other fields. Improved interactions between the two fields is clearly needed. It would give astronomers more effective techniques for understanding the Universe, and statisticians challenging and important problems to address.

REFERENCES

1. Arnett, W. D. and Rosner, J. L. (1987). *Phys. Rev. Lett.*, **58**, 1906.
2. Avni, Y. and Tananbaum, H. (1986). *Astrophys. J.*, **305**, 57.
3. Avni, Y., Soltan, A., Tananbaum, H., and Zamorani, G. (1980). *Astrophys. J.*, **235**, 694.
4. Bahcall, J. N. and Glashow, S. L. (1987). *Nature*, **326**, 476.
5. Barrow, J. D., Bhavsar, S. P., and Sonoda, D. H. (1985). *Monthly Notices R. Astron. Soc.*, **216**, 17.
6. Bevington, P. R. (1969). *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York.
7. *Bulletin d'Information du Centre de Données Stellaires*, Observatoire de Strasbourg, 11 Rue de l'Université, 67000 Strasbourg, France.
8. Caldwell, J. A. and Ostriker, J. P. (1981). *Astrophys. J.*, **251**, 61.
9. de Lapparent, V., Geller, M. J., and Huchra, J. P. (1986). *Astrophys. J. Lett.*, **302**, L1.
10. Feigelson, E. D. and Nelson, P. I. (1985). *Astrophys. J.*, **293**, 192.
11. Fry, J. N. (1986). *Astrophys. J.*, **306**, 366.
12. Gilmore, G. and Carswell, B., eds. (1987). *The Galaxy*. Reidel, Dordrecht, Netherlands.
13. Gott, J. R., Melott, A. L., and Dikinson, M. (1986). *Astrophys. J.*, **306**, 341.
14. Hald, A. (1986). *Int. Statist. Rev.*, **54**, 211–220.
15. Heck, A., Manfroid, J., and Mersch, G. (1985). *Astron. Astrophys. Suppl.*, **59**, 63.
16. Horne, J. H. and Baliunas, S. L. (1986). *Astrophys. J.*, **302**, 757.
17. Isobe, T. and Feigelson, E. D. (1987). *ASURV: Astronomy Survival Analysis, software package*.
18. Isobe, T., Feigelson, E. D., and Nelson, P. I. (1986). *Astrophys J.*, **306**, 490.
19. Kendall, D. G. (1985). In *Celebration of Statistics*, A. C. Atkinson and S. E. Fienberg, eds. Springer Verlag, Berlin.
20. Kuhn, J. R. and Uson, J. M. (1982). *Astrophys. J. Lett.*, **263**, L47.
21. Lynden-Bell, D. (1971). *Monthly Notices R. Astron. Soc.*, **136**, 219.
22. Lynden-Bell, D., Faber, S. M., Burstein, D., Davies, R. L., Dressler, A., Terlevich, R. J., and Wegner, G. (1988). *Astrophys. J.*, **326**, 19.
23. Marshall, H. L. (1987). *Astron. J.*, **94**, 620.
24. Moody, J. E., Turner, E. L., and Gott, J. R. (1983). *Astrophys. J.*, **273**, 16.
25. Murray, C. A. (1986). *Monthly Notices R. Astron. Soc.*, **223**, 649.
26. Murtagh, F. and Heck, A. (1987). *Multivariate Data Analysis*, *Astrophysics and Space Science Library*, **131**. Reidel, Dordrecht, Netherlands.
27. Narayan, R. and Nityananda, R. (1986). *Ann. Rev. Astron. Astrophys.*, **24**, 127.
28. Narlikar, J. V. (1982). *Sankhyā B*, **42**, 125.
29. Neyman, J. and Scott, E. L. (1952). *Astrophys J.*, **116**, 144.
30. Nicoll, J. F. and Segal, I. E. (1983). *Astron. Astrophys.*, **118**, 180.
31. Peebles, P. J. E. (1980). *The Large-Scale Structure of the Universe*. Princeton University Press, Princeton, NJ.

32. Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, London.
33. Roberts, J. A., ed (1984). *Indirect Imaging*. Cambridge University Press, London.
34. Scargle, J. D. (1982). *Astrophys. J.*, **263**, 835.
35. Schmitt, J. H. (1985). *Astrophys. J.*, **293**, 178.
36. Scott, E. L. (1976). In *On the History of Statistics and Probability*, D. B. Owen, ed. Marcel Dekker, New York.
37. Sheynin, O. B. (1984). *Archive for History of Exact Sciences*, **29**, 151–199.
38. *Stars and Stellar Systems* (1962–75). University of Chicago Press, Chicago. 9 volumes.
39. *Statistical Methods in Astronomy* (1983). SP-201, European Space Agency, ESA Scientific and Technical Publications, c/o ESTEC, Noordwijk, Netherlands.
40. Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, MA.
41. Trumpler, R. J. and Weaver, H. F. (1953). *Statistical Astronomy*. University of California Press, Berkeley, CA.
42. Weedman, D. (1986). *Quasar Astronomy*. Reidel, Dordrecht, Netherlands.
43. Woodroofe, M. (1985). *Ann. Statist.*, **13**, 163.

BIBLIOGRAPHY

Di Gesù, V., Scarsi, L., Crane, P., Friedman, J. H., and Levaldi, S., eds. (1985; 1986). *Data Analysis in Astronomy I; II*. Plenum, New York. (Proceedings of International Workshops held at Erice, Sicily, Italy in 1984 and 1985.)

Maistrov, L. E. (1974). *Probability Theory: A Historical Sketch*, translated by S. Kotz. Academic, New York.

EDITORIAL NOTE

S. R. Searle [*Commun. Statist. A*, **17**, 935–968 (1988)] notes that the earliest appearances of variance components* are in two books on astronomy, namely:

> Airy, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Macmillan, London.

and

> Chauvenet, W. (1863). *A Manual of Spherical and Practical Astronomy*. Lippincott, Philadelphia.

Searle also draws attention to the fact that the development of the least-squares* estimation of parameters in linear models was presented in books on astronomy, namely:

> Gauss, K. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conics Solem Ambientium*. Perthes and Besser, Hamburg.

and

> Legendre, A. M. (1806). *Nouvelles Methodes pour Determination des Orbites des Comètes*. Courcier, Paris.

See also Directional Data Analysis; Discriminant Analysis; Gauss, Carl Friedrich; Hierarchical Cluster Analysis; Least Squares; Linear Regression; Multivariate Analysis; Newcomb, Simon; Outlier Rejection, Chauvenet's Criterion; and Statistics: An Overview.

Eric D. Feigelson

ASYMMETRIC POPULATION

A population or a distribution that is not symmetric is *asymmetric*. The property of asymmetry is also related to skewness. It should be noted, however, that measures of skewness* usually correspond to some particular feature of symmetry. The third central moment μ_3, for example, is indeed zero for symmetric populations, but it can also be zero for populations that are asymmetric.

It is better to limit the use of the adjective "asymmetric" to distributions, and not apply it to populations.

See also Mean, Median, and Mode; Skewness: Concepts and Measures; and Skewness, Measures of.

ASYMPTOTIC EXPANSIONS—I

Many test statistics and estimators have probability distributions that may be approximated quite well by a normal distribution* in the case of a sufficiently large sample size. This is of practical use, for example, in such problems as determining critical regions* for tests of specified sizes and determining confidence regions* with specified confidence coefficients.

Let T_n denote a test statistic based on a sample $X_1, \ldots, X_n$ from a distribution F, let a_n and b_n be suitable normalizing constants, and let G_n denote the distribution of the normed statistic $(T_n - a_n)/b_n$. The "normal approximation" is expressed by

$$\lim_{n\to\infty} G_n(t) = \Phi(t) \qquad (-\infty < t < \infty), \qquad (1)$$

where

$$\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} \exp(-x^2/2)dx,$$

the standard normal distribution* function. Often (1) can be established under moderate regularity assumptions on the distribution function F and the functional form of the statistic T_n. *See* ASYMPTOTIC NORMALITY.

Of fundamental importance is the question of the error of approximation in (1) for a particular value of n. One useful type of answer is supplied by a "Berry-Esséen" rate, namely an assertion of the form

$$\sup_{-\infty<t<\infty} |G_n(t) - \Phi(t)| = O(n^{-1/2}), \qquad (2)$$

available under additional restrictions on F and the form of T_n (*see* CENTRAL LIMIT THEOREMS, CONVERGENCE RATES FOR). A more refined answer is given by an *expansion* of the error $G_n(t) - \Phi(t)$ in powers of $n^{-1/2}$. This requires additional restrictions on F and T_n. However, not only does it provide detailed information in (2), but it also supplies a way to replace $\Phi(t)$ by an improved approximation. Below we shall survey such expansions for the key special case that T_n is a *sum*, and then we shall comment briefly on other cases. For more details see the entries devoted to specific expansions.

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with distribution F, mean μ, variance σ^2, and characteristic function* ψ. Let G_n denote the distribution of the normed sum $(\sum_1^n X_i - n\mu)/(n^{1/2}\sigma)$. If Cramér's condition

$$\text{(C)} \qquad \lim_{|z|\to\infty} \sup |\psi(z)| < 1$$

is satisfied and the kth moment of F is finite, then

$$\left| G_n(t) - \Phi(t) - \sum_{j=1}^{k-3} P_{3j-1}(t) e^{-t^2/2} n^{-j/2} \right| < M n^{-(k-2)/2}, \qquad (3)$$

where M is a constant depending on k and F but not on n or t, and $P_m(t)$ is a polynomial (essentially the *Chebyshev–Hermite* polynomial*) of degree m in t. Indeed, we have the expressions

$$P_2(t)e^{-t^2/2} = -(\lambda_3/3!)\Phi^{(3)}(t),$$
$$P_3(t)e^{-t^2/2} = (\lambda_4/4!)\Phi^{(4)}(t) + (10\lambda_3^2/6!)\Phi^{(6)}(t), \ldots,$$

where λ_j denotes the jth cumulant* of F [the coefficient of $(iz)^j/j!$ in the MacLaurin series expansion of $\log \psi(z)$]. We may express λ_j (essentially) as a polynomial in the moments of F, obtaining in particular

$$\lambda_3 = \frac{E[(X-\mu)^3]}{\sigma^3} = \gamma_1$$

and

$$\lambda_4 = \frac{E[(X-\mu)^4]}{\sigma^4} - 3 = \gamma_2,$$

known as the coefficients of skewness* and of kurtosis*, respectively. Therefore, up to terms of order n^{-1}, which usually suffices in practical applications, the approximation given by (3) may be written conveniently in the form

$$G_n(t) \doteq \Phi(t) - \frac{\gamma_1}{6}(t^2-1)\phi(t)n^{-1/2} - \left[\frac{\gamma_2}{24}(t^3-3t) + \frac{\gamma_1^2}{72} \times (t^5 - 10t^3 + 15t)\right]\phi(t)n^{-1}, \qquad (4)$$

with error $O(n^{-3/2})$ uniformly in t.

The expansion (3) is called the *Edgeworth expansion* for G_n. Corresponding expansions for the density g_n follow by replacing all functions of t by their derivatives. The assumption (C) is always satisfied if the distribution F has an absolutely continuous component. Analogs of (3) hold under alternative conditions on F, e.g., the case of a lattice distribution*. Versions also have been developed allowing the X_i's to have differing distributions or to be stationary dependent. Furthermore, other metrics besides $\sup_t |G_n(t) - \Phi(t)|$ have been treated. For extensive treatments of (3) and these various ramifications, see [2,3,7,9,10,12].

An inverse problem related to (3) concerns the equation

$$G_n(t_p) = 1 - p,$$

where $0 < p < 1$. The solution t_p may be expressed asymptotically as

$$t_p \sim \mu + \sigma w, \qquad (5)$$

where w is given by the *Cornish–Fisher expansion**, which like (3) involves the quantities $\{\lambda_i\}$ and the Chebyshev–Hermite polynomials.

For detailed numerical illustration of the effectiveness of the expansions (3) and (5), see Abramowitz and Stegun [1], pp. 935–936, 955, 958–959. As noted in connection with (4), the improvement of the Edgeworth approximation over simply the normal approximation can be attributed to use of the coefficients of skewness and kurtosis; this provides a convenient intuitive basis for assessing the potential degree of improvement. Numerical illustration related to (4) is provided by Bickel and Doksum [5].

Finally, let us consider statistics other than sums. For such cases the question of asymptotic normality*, (1), has received extensive treatment. Secondarily, the associated Berry–Esséen rates, (2), have received attention. Consequently, results of types (1) and (2) are now available for several important wide classes of statistics: U-statistics*; von Mises differentiable statistical functions*; linear functions of order statistics; M-estimates*; and rank statistics*. Detailed exposition may be found in Serfling [11]. *see also* ASYMPTOTIC NORMALITY. However, except for isolated results, the question of asymptotic expansions analogous to (3) has only very recently gained intensive interest and development. For comments on Edgeworth expansions for rank statistics, such as the two-sample Wilcoxon statistic, *see* Hájek and Šidák [8], Sec. IV.4.2. Multivariate Edgeworth-type expansions are discussed by Chambers [6]. For a review of recent activity, see Bickel [4].

REFERENCES

1. Abramowitz, M. and Stegun, I. A., eds. (1965, 1970). *Handbook of Mathematical Functions*. U.S. Government Printing Office, Washington, D.C.
2. Bhattacharya, R. N. (1977). *Ann. Prob.*, **5**, 1–27.
3. Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.
4. Bickel, P. J. (1974). *Ann. Statist.*, **2**, 1–20.
5. Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*. Holden-Day, San Francisco.
6. Chambers, J. M. (1967). *Biometrika*, **54**, 367–383.
7. Cramér, H. (1970). *Random Variables and Probability Distributions*, 3rd ed. Cambridge University Press, Cambridge.
8. Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.
9. Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.
10. Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag, New York.

11. Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
12. Wallace, D. (1958). *Ann. Math. Statist.*, **29**, 635–654.

See also Approximations to Distributions; Asymptotic Normality; Central Limit Theorems, Convergence Rates for; Cornish–Fisher and Edgeworth Expansions; and Limit Theorem, Central.

R. J. Serfling

ASYMPTOTIC EXPANSIONS—II

Asymptotic expansions of functions are useful in statistics in three main ways. Firstly, conventional asymptotic expansions of special functions are useful for approximate computation of integrals arising in statistical calculations. An example given below is the use of Stirling's approximation to the gamma function. Second, asymptotic expansions of density or distribution functions of estimators or test statistics can be used to give approximate confidence limits for a parameter of interest or p-values for a hypothesis test. Use of the leading term of the expansion as an approximation leads to confidence limits and p-values based on the limiting form of the distribution of the statistic, whereas use of further terms often results in more accurate inference. Third, asymptotic expansions for distributions of estimators or test statistics may be used to investigate properties such as the efficiency of an estimator or the power of a test. The first two of these are discussed in this entry, which is a continuation of Serfling [35]. The third is discussed in the entry Asymptotics, Higher Order*.

An asymptotic expansion of a function is a reexpression of the function as a sum of terms adjusting a base function, expressed as follows:

$$f_n(x) = f_{0n}(x)[1 + b_{1n}g_1(x) + b_{2n}g_2(x) + \cdots + b_{kn}g_k(x) + O(b_{k+1,n})] \quad (n \to \infty). \tag{1}$$

The sequence $\{b_{kn}\} = \{1, b_{1n}, b_{2n}, \ldots\}$ determines the asymptotic behavior of the expansion: in particular how the reexpression approximates the original function. Usual choices of $\{b_{kn}\}$ are $\{1, n^{-1/2}, n^{-1}, \ldots\}$ or $\{1, n^{-1}, n^{-2}, \ldots\}$; in any case it is required that $b_{kn} = o(b_{k-1,n})$ as $n \to \infty$. For sequences of constants $\{a_n\}$, $\{b_n\}$, we write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$, and $a_n = O(b_n)$ if a_n/b_n remains bounded as $n \to \infty$. The notation $o_p(\cdot)$, $O_p(\cdot)$ is useful for sequences of random variables $\{Y_n\} : Y_n$ is $o_p(a_n)$ if Y_n/a_n converges in probability to 0 as $n \to \infty$, and is $O_p(a_n)$ if $|Y_n/a_n|$ is bounded in probability as $n \to \infty$.

Asymptotic expansions are used in many areas of mathematical analysis. Three helpful textbooks are Bleistein and Handelsman [9], Jeffreys [25], and DeBruijn [18]. An important feature of asymptotic expansions is that they are not in general convergent series, and taking successively more terms from the right-hand side of (1) is not guaranteed to improve the approximation to the left-hand side. In the study of asymptotic expansions in analysis, emphasis is typically on $f_n(x)$ as a function of n, with x treated as an additional parameter, and n considered as the argument of the function. In these treatments it is usual to let n be real or complex, and the notation $f(z; x)$ or $f(z)$ is more standard. The functions $g_j(\cdot)$ that we have used in (1) are then just constants (in z), and the sequence $\{b_{kn}\}$ is typically $\{z^{-k}\}$ if $z \to \infty$ or $\{z_k\}$ if $z \to 0$.

An asymptotic expansion is used to provide an approximation to the function $f_n(x)$ by taking the first few terms of the right-hand side of (1): for example, we might write

$$f_n(x) = f_{0n}(x)[1 + b_{1n}g_1(x)].$$

Although the approximation is guaranteed to be accurate only as $n \to \infty$, it is often quite accurate for rather small values of n. It will usually be of interest to investigate the accuracy of the approximation for various values of x as well, an important concern being the range of values for x over which the error in the approximation is uniform.

In statistics the function $f_n(x)$ is typically a density function, a distribution function, a moment or cumulant generating function, for a random variable computed from a sequence of random variables of length n. For example, $f_n(x)$ could be the density of the standardized mean $\overline{X}_n$, say, of n independent, identically distributed random variables $X_i : \overline{X}_n = \sum X_i/n$. In this case the function $f_{0n}(x)$ is the limiting density for the standardized version of $\overline{X}_n$, usually the normal density function $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$; *see* ASYMPTOTIC NORMALITY.

ASYMPTOTIC EXPANSIONS OF SPECIAL FUNCTIONS

A familiar asymptotic expansion is that of the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ by Stirling's formula* given, for example, in Abramowitz and Stegun [1, Sec. 6.1.37]:

$$\Gamma(z) = e^{-z}z^{z-1/2}(2\pi)^{1/2} \times \left(1 + \frac{1}{12z} + \frac{1}{288z^2} + O(z^{-3})\right) \quad (z \to \infty \quad \text{in } |\arg z| < \pi). \tag{2}$$

The leading term of the right-hand side is Stirling's approximation to the gamma function. There are similar expansions given in Abramowitz and Stegun [1, Chap. 6] for log $\Gamma(z)$ and its first two derivatives, the digamma and trigamma functions. The approximations given by the first several terms in these expansions are used, for example, in computing the maximum-likelihood estimator and its asymptotic variance for a sample from a gamma density with unknown shape parameter.

Another example is the asymptotic expansion for the tail of the standard normal cumulative distribution function:

$$1 - \Phi(z) = \int_z^\infty (2\pi)^{-1/2}\exp\left(-\frac{x^2}{2}\right)dx = \phi(z)z^{-1}\left(1 - \frac{1}{z^2} + \frac{3}{z^4} + O(z^{-6})\right). \tag{3}$$

The quantity $[1 - \Phi(z)]/\phi(z)$ is often called Mills' ratio*.

The asymptotic expansions given above are examples of expansions obtained using Laplace's method*. Laplace's method is also very useful for deriving approximations to integrals arising in Bayesian inference.

EDGEWORTH AND SADDLEPOINT EXPANSIONS

For statistics that are asymptotically normally distributed, the *Edgeworth expansion* for density or distribution functions gives a useful and readily computed approximation. Such statistics are typically either sample means or smooth functions of sample means, and the Edgeworth expansion for the distribution function of a standardized sample mean is given in Serfling [35]. We assume for simplicity that $X_1, \ldots, X_n$ are independent and identically distributed. Define $S_n = n^{1/2}(\overline{X}_n - \mu)/\sigma$, where μ and σ^2 are the mean and variance of X_i. The Edgeworth expansion for the density of S_n is

$$f_n(s) = \phi(s)\left[1 + \frac{1}{n^{1/2}}\frac{\lambda_3}{6}h_3(s) + \frac{1}{n}\left(\frac{\lambda_4}{24}h_4(s) + \frac{\lambda_3^2}{72}h_6(s)\right) + O(n^{-3/2})\right], \tag{4}$$

where λ_3 and λ_4 are the third and fourth cumulants of $(X_i - \mu)/\sigma$, and, $h_j(s) = (-1)^j \phi^{(j)}(s)/\phi(s)$ is the jth Hermite polynomial.

Note that (4) suggests the use of the first three terms as an approximation to the exact density, with the remaining terms absorbed into the expression $O(n^{-3/2})$. The full expansion for the distribution function is given in Serfling [35]. The Edgeworth expansion is derived in many textbooks; cf. the references in Serfling [35], Feller [22, Chap. 16], McCullagh [30, Chap. 6], and Barndorff-Nielsen and Cox [5, Chap. 4]. *See also* CORNISH–FISHER AND EDGEWORTH EXPANSIONS.

The Edgeworth approximation is quite accurate near the center of the density. In particular, at $\overline{x} = \mu$ the relative error in using the normal approximation is $O(n^{-1})$, and that in using the approximation suggested by (4) is $O(n^{-2})$, because the odd-order Hermite polynomials are 0 at $s = 0$. For large values of $|s|$, though, the approximation is often inaccurate for fixed values of n, as the polynomials oscillate substantially as $|s| \to \infty$. A particular difficulty is that the approximation to $f_n(s)$ may in some cases take negative values.

A different type of asymptotic expansion for the density of a sample mean is given by the *saddlepoint expansion*. Let the cumulant generating function of X_i be denoted by $K(t)$. The saddlepoint expansion for the density of $\overline{X}$ is defined by

$$f_n(\overline{x}) = \frac{1}{\sqrt{2\pi}}\left(\frac{n}{|K''(\hat{z})|}\right)^{1/2}\exp\{n[K(\hat{z}) - \hat{z}\overline{x}]\}$$
$$\times\left(1 + \frac{3\lambda_4(\hat{z}) - 5\lambda_3^2(\hat{z})}{24n} + O(n^{-2})\right), \quad (5)$$

where $\hat{z}$ is called the *saddlepoint* and is defined by $K'(\hat{z}) = \overline{x}$, and the rth cumulant function $\lambda_r(z) = K^{(r)}(z)/[K''(z)]^{r/2}$. Note that this is an asymptotic expansion in powers of n^{-1}. The next term in the expansion is a complicated expression involving cumulant functions up to order 6. The leading term of (5) is the saddlepoint approximation* to the density of $\overline{X}_n$. This expression is always positive, but will not usually integrate to exactly one, so in practice it is renormalized. The renormalization improves the order of the approximation:

$$f_n(\overline{x}) = c\left(\frac{n}{|K''(\hat{z})|}\right)^{1/2}\exp\{n[K(\hat{z}) - \hat{z}\overline{x}]\}$$
$$\times[1 + O(n^{-3/2})]. \quad (6)$$

Evaluating the saddlepoint approximation requires knowledge of the cumulant generating function $K(z)$. Approximations based on estimating $K(z)$ by estimating the first four cumulants are discussed in Easton and Ronchetti [21], Wang [41], and Cheah et al. [10].

The saddlepoint approximation to the distribution function of $\overline{X}$ can be obtained by integrating (6) or by applying the saddlepoint technique; the result, due to Lugannani and Rice [29], is

$$F_n(\overline{x}) = \Phi(r) + \phi(r)\left(\frac{1}{r} - \frac{1}{q}\right), \quad (7)$$

where

$$r = \text{sign}(q)[2n[K(\hat{z}) - \hat{z}\overline{x}]]^{1/2},$$
$$q = \hat{z}[K''(\hat{z})]^{1/2}.$$

The approximation (7) is often surprisingly accurate throughout the range of $\overline{x}$, except near $\overline{x} = \mu$ or $r = 0$, where it should be replaced by its limit as $r \to 0$:

$$F_n(\mu) = \tfrac{1}{2} - \tfrac{1}{6}\lambda_3(0)/\sqrt{2\pi n}.$$

The approximation (7) has relative error $O(n^{-1})$ for all $\overline{x}$ and $O(n^{-3/2})$ for the so-called moderate deviation region $\overline{x} - \mu = O(n^{-1/2})$. It can be expressed in an asymptotically equivalent form by defining $r^* = r + r^{-1}\log(q/r)$: the approximation

$$F_n(\overline{x}) \doteq \Phi(r^*), \quad (8)$$

originally due to Barndorff-Nielsen [4], is asymptotically equivalent to (7).

The approximations (5) and (7) were derived in Daniels [13] and Lugannani and Rice [29], respectively, using the saddlepoint technique of asymptotic analysis. Daniels [15] exemplifies the derivation of (7). Both Kolassa [27] and Field and Ronchetti [23] provide rigorous derivations of (5) using the saddlepoint method. General discussions of the saddlepoint method can be found in Bleistein and Handelsman [9] or Courant and Hilbert [11]. The approximations can also be derived from the Edgeworth expansion; cf. Barndorff-Nielsen and Cox [5, Chap. 4], where (5) is called the tilted Edgeworth expansion.

The Edgeworth and saddlepoint approximations for distribution functions discussed here apply to continuous random variables, and adjustments to the approximations are needed in the case that the variables X_i takes values on a lattice. The details are provided in Kolassa [27, Chaps. 3, 5].

For vector X_i of length m, say, multivariate versions of the Edgeworth and saddlepoint density approximations are readily available. The multivariate Edgeworth approximation requires for its expression a definition of generalized Hermite polynomials, which are conceptually straightforward but notationally complex. A brief account is given in Reid [33], adapted from McCullagh [30, Chap. 5]. The multivariate version of the saddlepoint approximation is the same as (6), with $K(z) = \log E\exp(z^T x)$, $\hat{z}^T\overline{x}$ interpreted as a scalar product of the two vectors

$\hat{z}$ and $\bar{x}$, and $|K''(\hat{z})|$ interpreted as the determinant of the $p \times p$ matrix of second derivatives of the cumulant generating function. The distribution-function approximation (7) is only available for the univariate case, but an approximation to the conditional distribution function $\Pr(\overline{X}_{(1)} \leqslant \bar{x}_{(1)}|\bar{x}_{(2)})$ is derived in Skovgaard [36] and extended in Wang [39] and Kolassa [27, Chap. 7]. The form of this approximation has proved very useful for inference about scalar parameters in the presence of nuisance parameters*.

The Edgeworth and saddlepoint approximations are both based on a limiting normal distribution for the statistic in question. Some statistics may have a limiting distribution that is not normal; in particular, sample maxima or minima usually have limiting distributions of the extreme-value form. For such statistics a series expansion of the density in which the basic function corresponds to the limiting density may be of more practical interest. In principle this is straightforward, and for example McCullagh [30, Chap. 5] derives Edgeworth-type expansions using arbitrary basis functions and the associated orthogonal polynomials. Examples of saddlepoint approximations based on nonnormal limits are discussed in Jensen [26] and Wood et al. [42].

STOCHASTIC ASYMPTOTIC EXPANSIONS

It is often very convenient in deriving asymptotic results in statistics to use *stochastic asymptotic expansions*, which are analogues of (1) for random variables. For a sequence of random variables $\{Y_n\}$ a stochastic asymptotic expansion is expressed as

$$Y_n = X_0 + X_1 b_{1_n} + \cdots + X_k b_{kn} + O_p(b_{k+1,n}), \tag{9}$$

where $\{X_0, X_1, \ldots\}$ have a distribution not depending on n. Stochastic asymptotic expansions are discussed in Cox and Reid [12] and in Barndorff-Nielsen and Cox [5, Chap. 5]. As an example, Cox and Reid [12] show that if Y_n follows a chi-squared distribution with n degrees of freedom, then

$$\frac{Y_n - n}{\sqrt{2n}} = X_0 + \frac{1}{n^{1/2}}\frac{\sqrt{2}}{3}(X_0^2 - 1) + O_p(n^{-1}),$$

where X_0 is a standard normal random variable. The relationship between stochastic asymptotic expansions and expansions for the corresponding distribution functions is discussed in Cox and Reid [12]. Expansions similar to (9) where the distributions of $X_0, X_1, \ldots$ are only asymptotically free of n are very useful in computing asymptotic properties of likelihood-based statistics.

An expansion closely related to (9) but usually derived in the context of the Edgeworth expansion for the distribution function of the sample means is the Cornish—Fisher expansion* for the quantile of the distribution function. As described in Serfling [35], an expansion for the value s_α satisfying $F_n(s_\alpha) = 1 - \alpha$ can be obtained by a reversion of the Edgeworth expansion. The result is

$$\begin{aligned} s_\alpha = z_\alpha &+ \frac{1}{6\sqrt{n}}(z_\alpha^2 - 1)\lambda_3 \\ &+ \frac{1}{24n}(z_\alpha^3 - 3z_\alpha)\lambda_4 \\ &- \frac{1}{36n}(2z_\alpha^3 - 5z_\alpha)\lambda_3^2 + O(n^{-3/2}), \end{aligned}$$

where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha$. Other asymptotic expansions for $F_n(s)$ lead to alternative expansions for s_α, and in particular the r^*-approximation given at (8) can be derived from the saddlepoint expansion for the distribution function of $\overline{X}_n$.

APPLICATIONS TO PARAMETRIC INFERENCE

There has been considerable development of statistical theory based on the use of higher-order approximations derived from asymptotic expansions. Likelihood-based inference, or inference from parametric models, has developed particularly rapidly, although the approximations are very useful in other contexts as well. Some examples of this will now be sketched.

Suppose that $X = (X_1, \ldots, X_n)$ is a sample from a parametric model that is an exponential family*, i.e., is of the form

$$f(x;\theta) = \exp[\theta^T x - b(\theta) - d(x)], \tag{10}$$

where X and θ take values in $\mathbb{R}^m$, say. The minimal sufficient statistic is $S = s(X) = \sum X_i$, and the maximum-likelihood estimator of θ is a one-to-one function of $S : c'(\hat{\theta}) = S/n$. Thus the saddlepoint approximation of S given in (7) can be used to give an approximation to the density for $\hat{\theta}$, which takes the form

$$f_n(\hat{\theta};\theta) = c|b''(\hat{\theta})|^{1/2} \times \exp[(\theta-\hat{\theta})'s - nb(\theta) + nb(\hat{\theta})].$$

If we denote the log-likelihood function for θ based on x by $\ell(\theta;x)$, and the observed Fisher information* function $-\partial^2\ell(\theta)/\partial\theta\theta^T$ by $j(\theta)$, we can reexpress this approximation as

$$f_n(\hat{\theta};\theta) = c|j(\hat{\theta})|^{1/2}\exp[\ell(\theta;x) - \ell(\hat{\theta};x)] \quad (11)$$

This approximation to the density for the maximum-likelihood estimator is usually known as *Barndorff-Nielsen's approximation*, or, following Barndorff-Nielsen, the p^* *approximation*. Although it has been used here to illustrate the saddlepoint approximation in an exponential family, the approximation (11) is valid quite generally. This was exemplified and illustrated in Barndorff-Nielsen [2, 3] and several subsequent papers. A review of the saddlepoint approximation and the literature on the p^*-formula* through 1987 is given in Reid [33]. A general proof and discussion of the interpretation of the p^* formula is given in Skovgaard [38]. Chapters 6 and 7 of Barndorff-Nielsen and Cox [6] provide an extensive discussion of the p^* formula and its applications in parametric inference. The validity of (11) in more general models requires the existence of a one-to-one transformation from the minimal sufficient statistic to $(\hat{\theta}, a)$, where a is a complementary statistic with a distribution either exactly or approximately (in a specific sense) free of θ; such statistics are called exact or approximate ancillary* statistics. The right-hand side of (11) then approximates the conditional distribution of $\hat{\theta}$, given a.

An illustration of the cumulative-distribution-function approximation (7) in the exponential family is also instructive. Suppose in (10) that $m = 1$. Then (7) provides an approximation to the distribution function for the maximum-likelihood estimate which is simply

$$\begin{aligned} F_n(\hat{\theta};\theta) &= \Phi(r)\left(\frac{1}{r} - \frac{1}{q}\right) \\ &= \Phi\left(r - r^{-1}\log\frac{r}{q}\right) = \Phi(r^*), \quad (12) \end{aligned}$$

where

$$\begin{aligned} r &= \text{sign}(\hat{\theta}-\theta)\{2[\ell(\hat{\theta}) - \ell(\theta)]\}, \\ q &= (\hat{\theta}-\theta)|j(\hat{\theta})|^{1/2} \end{aligned}$$

are the signed square root of the log-likelihood ratio statistic and the standardized maximum-likelihood estimator, respectively, and $r^* = r + r^{-1}\log(q/r)$.

As with the p^* approximation, the approximation (12) holds much more generally, with q replaced by a sometimes complicated statistic that depends on the underlying model and in particular on the exact or approximate ancillary statistic required for the validity of (11) in general models. Furthermore, (12) can be applied to marginal and conditional distributions for the maximum-likelihood estimate of a parameter of interest, after nuisance parameters have been eliminated via a marginal or conditional likelihood. A recent accessible reference is Pierce and Peters [32]. The approximation due to Skovgaard [36] is an important ingredient in this development. The r^* approximation, which for general families is due to Barndorff-Nielsen [4], is discussed in Barndorff-Nielsen and Cox [6, Chap. 6].

As an illustration of stochastic asymptotic expansions in likelihood-based inference, consider Taylor series expansion of the score equation $\ell'(\hat{\theta}) = 0$ (assuming for simplicity that this uniquely defines the maximum-likelihood estimator):

$$\begin{aligned} 0 &= \ell'(\theta) + (\hat{\theta}-\theta)\ell''(\theta) \\ &\quad + \tfrac{1}{2}(\hat{\theta}-\theta)^2\ell'''(\theta) + \cdots. \quad (13) \end{aligned}$$

Reversion of this expansion gives an expansion for the maximum-likelihood estimator that can be expressed as

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{Z_1(\theta)}{i(\theta)} + \frac{1}{\sqrt{n}}\left[\frac{Z_2(\theta)Z_1(\theta)}{[i(\theta)]^2} + \frac{Z_1^3(\theta)\rho_3(\theta)}{2[i(\theta)]^3}\right] + O_p(n^{-1}), \quad (14)$$

where $Z_1 = n^{-1/2}\ell'(\theta)$ and $Z_2 = n^{-1/2} \times [\ell''(\theta) - ni(\theta)]$, $i(\theta) = n^{-1}E[\ell'(\theta)]^2$, $\rho_3(\theta) = n^{-1}E[\ell'''(\theta)]$. The random variables Z_1 and Z_2 are $O_p(1)$ and have mean zero. In expansions of this type it is much easier to keep track of the orders of various terms using these standardized variables Z; this notation is originally due to D. R. Cox, and is extensively use as well in McCullagh [30, Chap. 7].

The expansion (14) is a type of stochastic asymptotic expansion, although strictly speaking the distributions for Z_1, Z_2 are only asymptotically free of n. The leading term of (14) gives the usual asymptotic normal approximation for the maximum-likelihood estimator, and the next-order term is useful for deriving refinements of this. For example, it is readily verified that the expected value of $\hat{\theta}$ has the expansion

$$E(\hat{\theta}) = \theta + n^{-1}\frac{i'(\theta) + \rho_3(\theta)/2}{[i(\theta)]^2} + O(n^{-2}),$$

and that $\mathrm{var}(\hat{\theta}) = [ni(\theta)]^{-1} + O(n^{-2})$.

The multivariate version of (14) is given in McCullagh [30, Chap. 7], as are extensions to the nuisance-parameter case and several illustrations of the use of these expansions. One particularly relevant application is the substitution of (14) into a Taylor series expansion of the log-likelihood ratio statistic $w(\theta) = 2[\ell(\hat{\theta}) - \ell(\theta)]$ to obtain an expansion of both the density and the expected value of $w(\theta)$. These expansions lead to the results

$$Ew(\theta) = m\left(1 + \frac{b(\theta)}{n} + O(n^{-2})\right)$$

and

$$\frac{w(\theta)}{1 + b(\theta)/n} = X_m^2[1 + O(n^{-3/2})], \quad (15)$$

where m is the dimension of θ and X_m^2 is a random variable following a χ^2 distribution on m degrees of freedom. The improvement of the approximation to the distribution of the log-likelihood ratio statistic given by (15) is called the *Bartlett correction*, after Bartlett [7], where the correction was derived for testing the equality of several normal variances (*see* BARTLETT ADJUSTMENT—I). It is a multivariate analogue of the improvement of the normal approximation to the signed square root given in (7). The expansion (15) is originally due to Lawley [28]: for details of the derivation see McCullagh [30, Chap. 7], Barndorff-Nielsen and Cox [6, Chap. 6], Bickel and Ghosh [8], and DiCiccio and Martin [19].

Approximations using the Edgeworth and saddlepoint expansions are also useful for statistics that are not derived from a likelihood-based approach to inference. Edgeworth expansions for more general statistics are discussed in Serfling [35] and in considerable generality in Pfanzagl [31]. Skovgaard [38] considers formulations for the density of minimum-contrast estimators that lead to the p^* approximation. Saddlepoint approximation to the density of M-estimators* is discussed in Daniels [14] and Field and Ronchetti [23]. Application of the saddlepoint approximation to the bootstrap is introduced in Davison and Hinkley [17], and explored further in Daniels and Young [16], DiCiccio et al. [20], Wang [40], and Ronchetti and Welsh [34].

A somewhat different application of asymptotic expansions in parametric inference is the use of the techniques outlined here to obtain asymptotic expansions for the efficiency of estimators and the power function of test statistics. One purpose of this is to provide a means for choosing among various estimators or test statistics that have the same efficiency or power to first order of asymptotic theory. Helpful surveys of these types of results are given in Skovgaard [37] and Ghosh [24]: see also the entry ASYMPTOTIC NORMALITY.

REFERENCES

1. Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.
2. Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, **67**, 293–310.

(Introduction to the p^* formula and proof of exactness for transformation models.)

3. Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum-likelihood estimator. *Biometrika*, **70**, 343–365. (Easier to read than [2]; includes derivation of p^* from the saddlepoint approximation.)
4. Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, **73**, 307–322. (First paper on the use of r^* approximation.)
5. Barndorff-Nielsen, O. E. and Cox, D. R. (1990). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London. (Presents detailed development of many asymptotic techniques that have proved useful for obtaining higher-order approximations in statistics, and includes a great number of useful examples. A very valuable reference text for this area. This article has been very strongly influenced by this text.)
6. Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London. (Follows on from ref. [5], but emphasizes the role of higher-order asymptotics in parametric inference, as outlined in the last section of this article.)
7. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. A*, **160**, 268–282.
8. Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and Bartlett correction—a Bayesian argument. *Ann. Statist.*, **18**, 1070–1090.
9. Bleistein, N. and Handelsman, R. A. (1975). *Asymptotic Expansions of Integrals*. Holt, Rinehart and Winston, New York.
10. Cheah, P. K., Fraser, D. A. S., and Reid, N. (1993). Some alternatives to Edgeworth. *Canad. J. Statist.*, **21**, 131–138.
11. Courant, R. and Hilbert, D. (1950). *Methods of Mathematical Physics*. Wiley, New York.
12. Cox, D. R. and Reid, N. (1987). Approximations to non-central distributions. *Canad. J. Statist.*, **15**, 105–114.
13. Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.*, **25**, 631–650. (The first derivation of the saddlepoint approximation for the density of the sample mean in the statistical literature. Very clearly written and important for any study of saddlepoint expansions. A helpful longer treatment is given in ref. [27].)
14. Daniels, H. E. (1983). Saddlepoint approximations for estimating equations. *Biometrika*, **70**, 89–96.
15. Daniels, H. E. (1987). Tail probability approximations. *Int. Statist. Rev.*, **55**, 37–48. (Derivation of distribution-function approximations using the saddlepoint technique. Illustration of the Lugannani—Rice approximation and two closely related approximations. Very clearly written. A helpful longer treatment is given in ref. [27].)
16. Daniels, H. E. and Young, G. A. (1991). Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika*, **78**, 169–179.
17. Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, **75**, 417–431.
18. DeBruijn, N. G. (1970). *Asymptotic Methods in Analysis*, 2nd ed. North-Holland, Amsterdam. Reprinted by Dover, 1981.
19. DiCiccio, T. J. and Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. R. Statist. Soc. B*, **55**, 305–316.
20. DiCiccio, T. J., Martin, M. A., and Young, G. A. (1992). Fast and accurate approximation to double bootstrap confidence intervals. *Biometrika*, **79**, 285–296.
21. Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L-statistics. *J. Am. Statist. Assoc.*, **81**, 420–430.
22. Feller, W. (1970). *Introduction to Probability Theory and Applications*, Vol. II, Wiley, New York.
23. Field, C. A. and Ronchetti, E. (1990). *Small Sample Asymptotics*. Institute of Mathematical Statistics, Hayward. (Emphasizes the use of exponential tilting in constructing approximations for statistics derived from sample means, with special emphasis on robust estimators.)
24. Ghosh, J. K. (1994). *Higher Order Asymptotics*. Institute of Mathematical Statistics, Hayward.
25. Jeffreys, H. (1962). *Asymptotic Approximations*. Oxford University Press.
26. Jensen, J. L. (1986). Inference for the mean of a gamma distribution with unknown shape parameter. *Scand. J. Statist.*, **13**, 135–151.
27. Kolassa, J. E. (1994). *Series Approximation Methods in Statistics*. Springer-Verlag, New York. (A very detailed derivation of saddlepoint and Edgeworth expansions that includes a lot of background material. Very helpful for

understanding how higher-order approximations are derived.)

28. Lawley, D. (1956). A general method for approximating to the distribution of the likelihood ratio criterion. *Biometrika*, **43**, 295–303.
29. Lugannani, R. and Rice, S. O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.*, **12**, 475–490. [Derivation of the approximation given in (7).]
30. McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London. (Introduces tensor calculus for use in multivariate statistical calculations. Derives multivariate Edgeworth and saddlepoint expansions, and considers application of asymptotic expansions and stochastic asymptotic expansions in parametric statistical inference.)
31. Pfanzagl, J. (1985). *Asymptotic Expansions for General Statistical Models*. Springer-Verlag, New York. (Rigorous development of expansions for efficiency of statistical estimators and power of statistical tests. A comparison of this approach and the one emphasized in this article is given in ref. [37].)
32. Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. R. Statist. Soc. B*, **54**, 701–738.
33. Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.*, **3**, 213–238. (A review of the saddlepoint approximation and its relationship to the p^* formula, and the uses of the p^* formula in inference.)
34. Ronchetti, E. and Welsh, A. (1994). Empirical saddlepoint approximations for multivariate M-estimators. *J. R. Statist. Soc. B*, **56**, 313–326.
35. Serfling, R. J. (1982). Asymptotic expansion. In *Encyclopedia of Statistical Sciences*. Wiley, New York, Vol. 1, pp. 137–138.
36. Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.*, **24**, 875–887.
37. Skovgaard, I. M. (1989). A review of higher order likelihood inference. *Bull. Int. Statist. Inst.*, **53**, 331–350.
38. Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *Ann. Statist.*, **18**, 779–789. (A very helpful derivation of the p^* formula is given in Section 3.)
39. Wang, S. (1990). Saddlepoint approximation for bivariate distributions. *J. Appl. Probab.*, **27**, 586–597.
40. Wang, S. (1990). Saddlepoint approximations in resampling. *Ann. Inst. Statist. Math.*, **42**, 115–131.
41. Wang, S. (1992). General saddlepoint approximations in the bootstrap. *Statist. Probab. Lett.*, **13**, 61–66.
42. Wood, A. T. A., Booth, J. G., and Butler, R. W. (1993). Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions. *J. Amer. Statist. Ass.*, **88**, 680–686.

See also Asymptotic Normality; Asymptotics, Higher Order; Cornish–Fisher and Edgeworth Expansions; Laplace's Method; p^*-Formula; and Saddle Point Approximations.

Nancy Reid

ASYMPTOTIC NORMALITY

The exact distribution of a statistic is usually highly complicated and difficult to work with. Hence the need to approximate the exact distribution by a distribution of a simpler form whose properties are more transparent. The limit theorems* of probability theory provide an important tool for such approximations. In particular, the classical central limit theorems* state that the sum of a large number of independent random variables is approximately normally distributed under general conditions (see the section "Central Limit Theorems for Sums of Independent Random Variables"). In fact, the normal distribution* plays a dominating role among the possible limit distributions. To quote from Gnedenko and Kolmogorov [18, Chap. 5]: "Whereas for the convergence of distribution functions of sums of independent variables to the normal law only restrictions of a very general kind, apart from that of being infinitesimal (or asymptotically constant), have to be imposed on the summands, for the convergence to another limit law some very special properties are required of the summands." Moreover, many statistics behave asymptotically like sums of independent random variables (see the fifth, sixth, and seventh sections). All of this helps to explain the importance of the normal distribution* as an asymptotic distribution.

Suppose that the statistics $T_n, n = 1, 2, \ldots$, when suitably normed, have the standard normal limit distribution; i.e., for some constants $b_n > 0$ and a_n and for every real x we have

$$\Pr[(T_n - a_n)/b_n \leqslant x] \to \Phi(x) \qquad \text{as } n \to \infty \tag{1}$$

where

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2} dy.$$

Then we say that T_n is asymptotically normal with mean a_n and variance b_n^2, or asymptotically normal (a_n, b_n^2). [Note that a_n and b_n^2 need not be the mean and the variance of T_n; indeed, (1) may hold even when T_n has no finite moments.]

It can be shown that if (1) holds for every x, the convergence is uniform in x, so that

$$\sup_{-\infty<x<\infty} |\Pr[(T_n - a_n)/b_n \leqslant x] - \Phi(x)| \to 0 \qquad \text{as } n \to \infty. \tag{2}$$

[This is due to the continuity of $\Phi(x)$.]

The knowledge that (1) or (2) holds is not enough for most statistical applications. For one thing, the statistician wants to know how large n has to be in order that the limit distribution may serve as a satisfactory approximation. Also, if the distribution of T_n depends on unknown parameters, the statistician wants to know how the values of the parameters affect the speed of convergence to the limit. Both goals are met, to some extent, by the Berry–Esseen theorem* (see below) and related results discussed in "Remainder Term in the Central Limit Theorem."

When the approximation provided by the limit distribution* is unsatisfactory, asymptotic expansions*, treated in the third section, may prove more helpful.

Conditions for the convergence of the moments of a statistic to the corresponding moments of its limit distribution are briefly discussed in the fourth section. The fifth section deals with the distributions of functions of asymptotically normal random variables. The asymptotic normality of functions of independent random variables and of sums of dependent random variables is considered in the sixth and seventh sections, respectively. The final section deals with functional central limit theorems, which are concerned with asymptotic distributions of random functions.

CENTRAL LIMIT THEOREMS* FOR SUMS OF INDEPENDENT RANDOM VARIABLES

The following classical central limit theorem for the partial sums of an infinite sequence of independent, identically distributed (i.i.d.) random variables is due to Lindeberg.

Theorem 1. Let $X_1, X_2, \ldots$ be an infinite sequence of i.i.d. random variables with finite mean a and positive and finite variance σ^2. Then, as $n \to \infty, X_1 + \cdots + X_n$ is asymptotically normal $(an, \sigma^2 n)$.

In the following theorem only finite sequences of independent (not necessarily identically distributed) random variables are involved, which makes it better adapted to most applications.

Theorem 2. For each $N = 1, 2, \ldots$ let $X_{N1}, X_{N2}, \ldots, X_{Nn}$ be $n = n(N)$ independent random variables with finite p-th moments, for some $p > 2$. Let $B_N = \sum_j \text{var}(X_{Nj})$ (the index j runs from 1 to n). If

$$B_N^{-p/2} \sum_j E|X_{Nj} - EX_{Nj}|^p \to 0 \qquad \text{as } N \to \infty, \tag{3}$$

then $\sum_j X_{Nj}$ is asymptotically normal $(\sum_j EX_{Nj}, B_N)$.

This theorem is due to Liapunov. Condition (3) may be replaced by a weaker condition, due to Lindeberg, which does not assume finite moments of order >2 (see ref. 32).

For a general central limit theorem for sums of independent random variables which assumes no finite moments and for other central limit theorems, see ref. 32, Chap. IV, Sec. 4.

Multidimensional central limit theorems give conditions for the convergence of the distribution of a sum of independent random vectors to a multivariate normal distribution*; see Cramér [10] and Uspensky [40].

If the sums of independent random variables have probability densities, the latter will converge, under certain conditions, to a normal probability density. For results of this type, known as local* central limit theorems, see ref. 32.

REMAINDER TERM IN THE CENTRAL LIMIT THEOREM

The following result, due to Esseen [17], gives an explicit upper bound for the difference between the distribution function of a sum of independent random variables and the normal distribution function.

Theorem 3. Let $X_1, \ldots, X_n$ be independent random variables,

$$EX_j = 0, \qquad E|X_j|^3 < \infty, \qquad (j = 1, \ldots, n),$$

and let

$$B_n = \sum_{j=1}^{n} EX_j^2 > 0, \qquad L_n = \sum_{j=1}^{n} E|X_j|^3 / B_n^{3/2}.$$

Then

$$\left| \Pr\left[B_n^{-1/2} \sum_{j=1}^{n} X_j \leqslant x \right] - \Phi(x) \right| \leqslant CL_n \quad \text{for all } x, \tag{4}$$

where C is a numerical constant.

The assumption $EX_j = 0$ is made merely to simplify the notation. If $EX_j = a_j$, replace X_j by $X_j - a_j$ in the statement of the theorem.

The least value of C for which (4) holds is not known. It is known [2] that (4) is true with $C = 0.7975$ and is not true with $C < 0.4097$.

Note that Theorem 3 involves only one finite sequence of independent random variables and is not a limit theorem. It easily implies Theorem 2 with $p = 3$.

Under the further assumption that $X_1, \ldots, X_n$ are identically distributed, with $EX_1 = 0$, $EX_1^2 = \sigma^2$, inequality (4) simplifies to

$$\left| \Pr\left[n^{-1/2}\sigma^{-1} \sum_{j=1}^{n} X_j \leqslant x \right] - \Phi(x) \right| \leqslant Cn^{-1/2}\sigma^{-3}E|X_1|^3. \tag{5}$$

This inequality was also derived by Berry [4] and is known as the Berry–Esseen inequality.

The upper bounds in (4) and (5) do not depend on x. S. V. Nagaev has shown that inequality (5) is still true (except perhaps for the value of the constant C) if the right side is multiplied with $(1 + |x|)^{-3}$. For this and related results, see ref. 32.

For extensions of these results to sums of independent random vectors, see ref. 5.

ASYMPTOTIC EXPANSIONS*

Let $X_1, X_2, \ldots$ be i.i.d. random variables, $EX_1 = 0$, $0 < \sigma^2 = EX_1^2 < \infty$,

$$F_n(x) = \Pr\left[\sum_{j=1}^{n} X_j \leqslant x\sigma n^{1/2} \right].$$

By Theorem 1, $F_n(x) \to \Phi(x)$ as $n \to \infty$. However, the approximation of $F_n(x)$ by $\Phi(x)$ is often too crude to be useful. There are expansions of the difference $F_n(x) - \Phi(x)$ in powers of $n^{-1/2}$ that may provide more accurate approximations.

The form of the expansion depends on whether the random variable X_1 is lattice* or nonlattice. [A random variable X is called a lattice random variable if, for some numbers $h > 0$ and a, the values of $(X - a)/h$ are integers; the largest h with this property is called the maximum span. Otherwise, X is nonlattice.]

Theorem 4. If the random variables X_1, $X_2, \ldots$ are i.i.d., nonlattice, and have a finite third moment, then

$$F_n(x) = \Phi(x) + \Phi'(x)Q_1(x)n^{-1/2} + o(n^{-1/2}) \tag{6}$$

uniformly in x. Here $\Phi'(x) = (2\pi)^{1/2} \exp \cdot (-x^2/2)$ is the standard normal density function and

$$Q_1(x) = \frac{1}{6}\frac{EX_1^3}{\sigma^3}(1 - x^2).$$

For a proof and for extensions of (6) involving higher powers of $n^{-1/2}$, see refs. 5 and 32.

Expansions of this type have been studied by Chebyshev*, Edgeworth, Cramér, Esseen, and others.

Theorem 5. If $X_1, X_2, \ldots$ are i.i.d. lattice random variables taking the values $a + kh$ ($k = 0, \pm 1, \pm 2, \ldots$), where h is the maximum span, and have a finite third moment, then

$$F_n(x) = \Phi(x) + \Phi'(x)(Q_1(x) + S_1(x))n^{-1/2} + o(n^{-1/2}) \quad (7)$$

uniformly in x. Here

$$S_1(x) = \frac{h}{\sigma} S\left(\frac{x\sigma n^{1/2} - an}{h}\right),$$
$$S(x) = [x] - x + \tfrac{1}{2},$$

and $[x]$ is the largest integer $\leqslant x$.

This theorem is due to Esseen [17]; see also ref. 18. For an extension of (7) that involves higher powers of $n^{-1/2}$, see refs. 5 and 32.

Asymptotic expansions of the distribution function and the probability density function of a sum of independent random variables that need not be identically distributed are also treated in ref. 32.

CONVERGENCE OF MOMENTS

If a statistic T_n has a normal limit distribution, its moments need not converge to the corresponding moments of the latter; in fact, T_n need not have any finite moments.

If the conditions of Theorem 2 with a fixed $p > 2$ are satisfied then for all positive integers $q \leqslant p$, the qth absolute moment of $\sum_j (X_{Nj} - EX_{Nj})/B_N^{1/2}$ converges to the corresponding moment of the standard normal distribution; see S. N. Bernstein* [3] and Hall [23]. A similar result is due to Zaremba [41]. Bounds for the remainder terms in such limit theorems for moments have been obtained by von Bahr [1] and Hall [23], among others. An interesting discussion of the convergence of moments of certain statistics can be found in Cramér [9, Chap. 27].

FUNCTIONS OF ASYMPTOTICALLY NORMAL RANDOM VARIABLES

We often encounter statistics that are functions of sample moments or of generalized sample moments of the form $M_n = n^{-1} \sum_{j=1}^n g(X_j)$. If the X_j are i.i.d., $Eg(X_1) = a$, $\operatorname{var} g(X_1) = \sigma^2$ $(0 < \sigma^2 < \infty)$, then M_n is asymptotically normal $(a, \sigma^2/n)$.

Theorem 6. Let the random variables M_n, $n \geqslant 1$, be asymptotically normal $(a, \sigma^2/n)$. If $H(x)$ is a function of the real variable x whose derivative $H'(x)$ exists and is $\neq 0$ and continuous at $x = a$, then $H(M_n)$ is asymptotically normal $(H(a), H'(a)^2\sigma^2/n)$.

This result can be extended to functions of k moment-like statistics which are asymptotically k-variate normal. We state the extension for $k = 2$.

Theorem 7. Let the random vectors (M_{1n}, M_{2n}), $n \geqslant 1$, be asymptotically bivariate normal* with mean (a_1, a_2) and covariances σ_{ij}/n, $i, j = 1, 2$. If $H(x, y)$ is a function of the real variables x and y whose partial derivatives at (a_1, a_2),

$$H_1 = \partial H(x,y)/\partial x|_{(a_1,a_2)},$$
$$H_2 = \partial H(x,y)/\partial y|_{(a_1,a_2)},$$

exist and are not both zero, and which has a total differential at (a_1, a_2), so that

$$H(x,y) = H(a_1,a_2) + H_1 x + H_2 y + x\epsilon_1(x,y) + y\epsilon_2(x,y)$$

where $\epsilon_i(x,y) \to 0$ as $(x,y) \to (a_1,a_2) (i = 1, 2)$, then $H(M_{1n}, M_{2n})$ is asymptotically normal with mean $H(a_1, a_2)$ and variance $(H_1^2\sigma_{11} + 2H_1H_2\sigma_{12} + H_2^2\sigma_{22})/n$.

Proofs of these or closely related results can be found in refs. 9 and 26.

Note that the conditions of Theorems 6 and 7 are such that $H(M_n)$ and $H(M_{1n}, M_{2n})$ can be approximated by the linear terms of their Taylor expansions*. If the linear terms vanish and they can be approximated by the quadratic terms, the asymptotic distribution will be that of a quadratic form* in normal random variables.

ASYMPTOTIC NORMALITY OF FUNCTIONS OF INDEPENDENT RANDOM VARIABLES

Let $T_n = T_n(X_1, \ldots, X_n)$ be a function of the independent random variables $X_1, \ldots, X_n$. Suppose that $ET_n^2 < \infty$.

*Hájek's projection lemma** approximates T_n by the statistic

$$\hat{T}_n = \sum_{j=1}^{n} E[T_n|X_j] - (n-1)ET_n,$$

which is a sum of independent random variables. By the corollary of that entry we have:

Theorem 8. Let the stated assumptions be satisfied for all n. Suppose that $\hat{T}_n$ is asymptotically normal $(E\hat{T}_n, \operatorname{var}\hat{T}_n)$ and that

$$(\operatorname{var}\hat{T}_n)/\operatorname{var}(T_n) \to 1 \qquad \text{as } n \to \infty.$$

Then T_n is asymptotically normal $(ET_n, \operatorname{var} T_n)$.

Hájek [21] and Dupač and Hájek [14] used the projection lemma to prove the asymptotic normality of a simple linear rank statistic*,

$$\sum_{j=1}^{n} a_n(j) b_n(R_{nj}),$$

where $a_n(j)$, $b_n(j)$ are constants, $R_{n1}, \ldots, R_{nn}$ are the respective ranks* of $X_{n1}, \ldots, X_{nn}$, and, for each $n, X_{n1}, \ldots, X_{nn}$ are mutually independent, continuously distributed random variables. (For details, see the papers cited.) On the asymptotic normality of linear ranks statistics*, see also refs. 20 and 22. Compare also the end of the following section. Related results on multivariate linear rank statistics have been obtained in Ruymgaart and van Zuijlen [37] and the papers there cited.

Another class of statistics whose asymptotic normality can be proved with the help of Hájek's lemma are the U-statistics*,

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leqslant j_1 < \cdots < j_m \leqslant n} f(X_{j1}, \ldots, X_{jm}),$$

where m is a fixed integer, $n \geqslant m$, the X_j are mutually independent random variables, and f is a real-valued function, symmetric in its m arguments; see ref. 24. A Berry–Esseen type bound for U-statistics* is derived in ref. 7.

Linear combinations of functions of order statistics* are asymptotically normal under general conditions; see Ruymgaart and van Zuijlen [36] and other work there cited.

There are statistics T_n which satisfy the conditions of Hájek's lemma and are asymptotically normally distributed, but whose asymptotic normality cannot be established by means of Theorem 8. A simple example is

$$T_n = X_1X_2 + X_2X_3 + \cdots + X_{n-1}X_n,$$

where the X_j are i.i.d. with a finite second moment. This is a special case of a sum of 1-dependent random variables; see the following section.

On the asymptotic normality of the sum of a random number of independent random variables (which is of interest in sequential analysis*), see ref. 8.

ASYMPTOTIC NORMALITY OF SUMS OF DEPENDENT RANDOM VARIABLES

Sums of independent random variables are asymptotically normal under general conditions. We may expect that the asymptotic normality will be preserved if the summands are allowed to be weakly dependent* in a suitable sense.

One way of expressing weak dependence is in terms of conditional expectations. For example, let $X_1, X_2, \ldots$ be a sequence of (possibly dependent) random variables and let $S_n = X_1 + \cdots + X_n$. Suppose that the Liapunov condition (3) with $N = n, X_{Nj} = X_j$ is satisfied and that $EX_j = 0$. In ref. 29, Sec. 31, it is shown that if, in addition, the conditional moments $E[X_j|S_{j-1}]$ and $E[X_j^2|S_{j-1}]$ differ sufficiently little (in a specified sense) from the corresponding unconditional moments, then S_n is asymptotically normal. For other, related results, see ref. 29.

Dvoretzky [15] has shown that sums of dependent random variables are asymptotically normal under conditions such as those in the central limit theorems for sums of independent random variables (e.g., Theorem 2), except that quantities such as means, and the like, are replaced by conditional means, and the like, the conditioning being relative to the preceding sum.

Another notion of weak dependence that has proved fruitful is the following. For simplicity we restrict ourselves to stationary sequences (X_n), so that, for all n, the joint distribution of $X_{h+1}, \ldots, X_{h+n}$ does not depend on h. A stationary sequence (X_n) is said to satisfy the strong mixing condition if there are numbers $d(r)$ converging to 0 as $r \to \infty$ such that

$$|\Pr[A \cap B] - \Pr[A]\Pr[B]| \leqslant d(n-m)$$

for any events A and B determined by conditions on the random variables X_k, $k \leqslant m$ and X_k, $k \geqslant n$, respectively, and for all m, n $(m < n)$.

Rosenblatt [34] has shown that the partial sums $X_1 + \cdots + X_n$ of a stationary sequence satisfying the strong mixing condition are asymptotically normal under conditions on some of their moments. For other sufficient conditions, see ref. 27.

A simple example of a sequence satisfying the strong mixing condition is an m-dependent sequence. The sequence $(X_n, n \geqslant 1)$ is said to be m-dependent if for all integers $1 \leqslant r \leqslant s < t \leqslant u$ the random vectors $(X_r, \ldots, X_s)$ and $(X_t, \ldots, X_u)$ are independent whenever $t - s > m$. A central limit theorem for sums of m-dependent random variables was proved in ref. 26. An improved version is due to Orey [30]. On Berry–Esseen type bounds for sums of m-dependent random variables, see Shergin [38].

Sums of m-dependent random variables and U-statistics have the feature in common that some subsets of the summands are mutually independent. A central limit theorem for more general sums of this type is due to Godwin and Zaremba [19].

For a central limit theorem for Markov chains* under conditions related to strong mixing, see Rosenblatt [35]. Sums of martingale differences are asymptotically normal under appropriate conditions; see, e.g., ref. 8.

Finally, we mention some of the so-called combinatorial central limit theorems, which have uses in sampling from a finite population* and in rank statistics. Let the random vector $(R_{n1}, \ldots, R_{nn})$ be uniformly distributed on the $n!$ permutations of the integers $1, \ldots, n$, and let $a_n(j)$, $b_n(j)$, $j = 1, \ldots, n$, be real numbers. Then the sums

$$\sum_{j=1}^{n} a_n(j) b_n(R_{nj})$$

are asymptotically normal under certain conditions on the $a_n(j)$, $b_n(j)$; see ref. 20. A similar result [25] holds for sums of the form

$$\sum_{j=1}^{n} a_n(j, R_{nj}).$$

FUNCTIONAL CENTRAL LIMIT THEOREMS

Functional central limit theorems form a far-reaching extension of the classical central limit theorems. We confine ourselves to a brief description of some typical results in this area.

Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean 0 and variance 1. Let $S_0 = 0$, $S_n = X_1 + \cdots + X_n$, $n \geqslant 1$, and define for $0 \leqslant t \leqslant 1$

$$Y_n(t) = n^{-1/2} S_{[nt]} + n^{-1/2}(nt - [nt]) X_{[nt]+1},$$

where $[nt]$ is the largest integer $\leqslant nt$. Thus for given values of $n, X_1, \ldots, X_n, Y_n(t)$ is a continuous, piecewise linear function of t such that $Y_n(j/n) = n^{-1/2} S_j$ for $j = 0, 1, \ldots, n$.

Now let $W(t)$, $0 \leqslant t \leqslant 1$, be the standard Brownian motion* process (Wiener process*) on [0,1]. Thus for each fixed $t \in (0, 1]$ the random variable $W(t)$ is normally distributed with mean 0 and variance t $(W(0) = 0)$, and for any finitely many points $t_1 < t_2 < \cdots < t_k$ in [0,1] the increments $W(t_2) - W(t_1)$, $W(t_3) - W(t_2), \ldots, W(t_k) - W(t_{k-1})$ are mutually independent. Each increment $W(t_j) - W(t_{j-1})$ is normally distributed with mean 0 and variance $t_j - t_{j-1}$. These facts determine the joint (normal) distribution of $W(t_1), \ldots, W(t_k)$. It is known that the random function $W(t)$, $0 \leqslant t \leqslant 1$, is continuous with probability 1.

By a theorem of Donsker [11] the random functions Y_n converge in distribution*, as $n \to \infty$, to the random function W. The exact meaning of this statement is explained, e.g., in Billingsley [6a]. *See also* CONVERGENCE OF

SEQUENCES OF RANDOM VARIABLES. It has the important implication that for a large class of functionals $h(f)$ of a continuous function $f(t)$, $0 \leqslant t \leqslant 1$, the distributions of the random variables $h(Y_n)$ converge to that of $h(W)$. A trivial example is $h(f) = f(1)$. The implication that the distributions of $Y_n(1) = n^{-1/2}S_n$ converge to that of $W(1)$ is essentially equivalent to the central limit theorem, Theorem 1. A more interesting functional to which Donsker's theorem applies is $h(f) = \max_{0 \leqslant t \leqslant 1} f(t)$. Since $\max_{0 \leqslant t \leqslant 1} Y_n(t) = n^{-1/2} \max(0, S_1, \ldots, S_n)$, Donsker's theorem implies that

$$\begin{aligned}\lim_{n\to\infty} &\Pr[n^{-1/2}\max(0, S_1, \ldots, S_n) \leqslant x] \\ &= \Pr[\max_{0 \leqslant t \leqslant 1} W(t) \leqslant x]. \qquad (8)\end{aligned}$$

A proof of Donsker's theorem can be found in Billingsley [6a], where also other applications of the theorem are discussed and other similar theorems are proved.

Donsker's theorem and theorems of a similar type are called functional central limit theorems.

For the limit in (8) we have

$$\Pr[\max_{0 \leqslant t \leqslant 1} W(t) \leqslant x] = \max(2\Phi(x) - 1, 0). \quad (9)$$

This can be proved from the properties of the Wiener process, or by applying the so-called invariance principle (not to be confused with the invariance principle* in statistical inference*). Donsker's theorem, just as the central limit theorem with $a = 0$ and $\sigma^2 = 1$, assumes only that the X_n are i.i.d. with mean 0 and variance 1. Thus in either theorem the limit is invariant in this class of distributions of X_n. Once it is known that the limit (8) exists, it can be evaluated directly by choosing the distribution of X_n in a convenient way; for details, see Billingsley [6a]. The idea of the invariance principle was first conceived by Erdös and Kac [16].

We conclude with another functional central limit theorem. Let $X_1, X_2, \ldots$ be i.i.d. random variables with common distribution function $F(t)$. Let $F_n(t)$ be the empirical distribution function* corresponding to the sample $X_1, \ldots, X_n$. Define the random function $Z_n(t)$, t real, by

$$Z_n(t) = n^{1/2}(F_n(t) - F(t)).$$

First suppose that the X_n are uniformly distributed* with $F(t) = t$, $0 \leqslant t \leqslant 1$. In this case $F_n(t) - F(t) = 0$ outside of [0, 1], and we may restrict t to the interval [0, 1].

Let $W^0(t)$, $0 \leqslant t \leqslant 1$, be the Brownian* bridge process. This is the Gaussian process* on [0,1] whose distribution is specified by the requirements

$$EW^0(t) = 0,$$
$$EW^0(s)W^0(t) = \min(s, t) - st.$$

In the present case (F uniform) the random functions Z_n converge in distribution to the random function W^0, in a similar sense as the convergence of Y_n to W; see Donsker [12] or Billingsley [6a]. (An important difference is that the functions Z_n are not continuous as the Y_n are.) One implication is that

$$\begin{aligned}\lim_{n\to\infty} &\Pr[\sup_t n^{1/2}|F_n(t) - F(t)| \leqslant x] \\ &= \Pr[\sup_t |W^0(t)| \leqslant x]. \qquad (10)\end{aligned}$$

Earlier, Kolmogorov [28] proved, by a different method, that the limit in (10) equals

$$1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2x^2} \qquad (11)$$

for $x > 0$. Thus the probability on the right of (10) is equal to (11). The present approach to deriving results such as (10) was heuristically described by Doob [13] and made rigorous by Donsker [12].

The case where $F(t)$ is an arbitrary continuous distribution function can be reduced to the uniform case (by noting that $X_n' = F(X_n)$ is uniformly distributed on [0, 1]), and (10) remains valid. For the case where $F(t)$ is any distribution function on [0, 1], see Billingsley [6a].

The general foundations underlying the functional central limit theorems were laid by Prohorov [33] and Skorohod [39] and are expounded in the books of Billingsley [6a,6b]. *See also* Parthasarathy [31] and Loève [29, Chap. 13].

REFERENCES

1. Bahr, B. von (1965). *Ann. Math. Statist.*, **36**, 808–818.
2. Beek, P. van (1972). *Zeit. Wahrscheinlichkeitsth.*, **23**, 187–196.
3. Bernstein, S. N. (1939). *Dokl. Akad. Nauk SSSR* (Compt. rend.), **24**, 3–8.
4. Berry, A. C. (1941). *Trans. Amer. Math. Soc.*, **49**, 122–136.
5. Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York. (Thorough treatment of normal approximations and asymptotic expansions of distributions of sums of independent random variables and random vectors, with emphasis on error bounds.)

6a. Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

6b. Billingsley, P. (1971). *Weak Convergence of Measures*. SIAM, Philadelphia.

7. Callaert, H. and Janssen, P. (1978). *Ann. Statist.*, **6**, 417–421.
8. Chow, Y. -S. and Teicher, H. (1978). *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, New York. (Includes careful proofs of representative results on asymptotic normality.)
9. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N.J. (A classic text, still unsurpassed.)
10. Cramér, H. (1970). *Random Variables and Probability Distributions*, 3rd ed. Cambridge Tracts in Mathematics and Mathematical Physics No. 36. Cambridge University Press, Cambridge (first ed., 1937). (Concise monograph in the classical vein.)
11. Donsker, M. (1951). An Invariance Principle for Certain Probability Limit Theorems. *Mem. Amer. Math. Soc. No. 6*.
12. Donsker, M. (1952). *Ann. Math. Statist.*, **23**, 277–281.
13. Doob, J. L. (1949). *Ann. Math. Statist.*, **20**, 393–403.
14. Dupač, V. and Hájek, J. (1969). *Ann. Math. Statist.*, **40**, 1992–2017.
15. Dvoretzky, A. (1972). *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, Vol. 2. University of California Press, Berkeley, Calif., pp. 513–535.
16. Erdös, P. and Kac, M. (1946). *Bull. Amer. Math. Soc.*, **52**, 292–302.
17. Esseen, C. -G. (1945). *Acta Math.*, **77**, 1–125. (A fundamental paper on asymptotic normality.)
18. Gnedenko, B. V. and Kolmogorov, A. N. (1968). *Limit Distributions for Sums of Independent Random Variables*, rev. ed. (Translated from the Russian by K. L. Chung.) Addison-Wesley, Reading, Mass. (A classic monograph on the subject of the title.)
19. Godwin, H. J. and Zaremba, S. K. (1961). *Ann. Math. Statist.*, **32**, 677–686.
20. Hájek, J. (1961). *Ann. Math. Statist.*, **32**, 501–523.
21. Hájek, J. (1968). *Ann. Math. Statist.*, **39**, 325–346.
22. Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York. (Includes results on asymptotic normality of rank statistics.)
23. Hall, P. (1978). *J. Aust. Math. Soc. A*, **25**, 250–256.
24. Hoeffding, W. (1948). *Ann. Math. Statist.*, **19**, 293–325. (On U-statistics.)
25. Hoeffding, W. (1951). *Ann. Math. Statist.*, **22**, 558–566.
26. Hoeffding, W. and Robbins, R. (1948). *Duke Math. J.*, **15**, 773–780. (On sums of m-dependent random variables.)
27. Ibragimov, I. A. and Linnik, Yu. V. (1971). *Independent and Stationary Sequences of Random Variables*. (Translated from the Russian and edited by J. F. C. Kingman.) Wolters-Noordhoff, Groningen. (Includes central limit theorems for sums of stationary random variables.)
28. Kolmogorov, A. N. (1933). *G. Ist. Att.*, **4**, 83–91.
29. Loève, M. (1977). *Probability Theory*, 4th ed., Vols. 1 and 2. Springer-Verlag, New York. (Includes treatment of sums of independent and dependent random variables.)
30. Orey, S. (1958). *Duke Math. J.*, **25**, 543–546.
31. Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.
32. Petrov, V. V. (1975). *Sums of Independent Random Variables*. (Translated from the Russian by A. A. Brown.) Springer-Verlag, New York. (Contains a wealth of information on the subject of the title, with proofs of the more important results.)
33. Prohorov, Yu. V. (1956). *Theory Prob. Appl.*, **1**, 157–214.
34. Rosenblatt, M. (1956). *Proc. Natl. Acad. Sci. USA*, **42**, 43–47.
35. Rosenblatt, M. (1971). *Markov Processes: Structure and Asymptotic Behavior*. Springer-Verlag, New York. (Contains a chapter

on the central limit theorem for Markov processes.)

36. Ruymgaart, F. H. and van Zuijlen, M. C. A. (1977). *Ned. Akad. Wetensch. Proc. A*, **80**, 432–447.

37. Ruymgaart, F. H. and van Zuijlen, M. C. A. (1978). *Ann. Statist.*, **6**, 588–602.

38. Shergin, V. V. (1979). *Teor. Veroyatn. Ee Primen.*, **24**, 781–794. (In Russian. English translation to appear in *Theory Prob. Appl.*, **24**, 782–796.)

39. Skorohod, A. V. (1956). *Theory Prob. Appl.*, **1**, 261–290.

40. Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York. (This early text includes much material not found in other books.)

41. Zaremba, S. K. (1958). *Math. Zeit.*, **69**, 295–298.

See also Central Limit Theorems, Convergence Rates for; Convergence of Sequences of Random Variables; Ergodic Theorems; Limit Theorems; and Limit Theorem, Central.

W. Hoeffding

ASYMPTOTIC NORMALITY OF EXPERIMENTS

Asymptotic normality* is a feature of many probabilistic or statistical studies. It is often connected with the central limit theorem* for sums of independent variables, for martingales* [13], or for empirical processes* [9,29]. In contrast to these essentially probabilistic results, we consider a more statistical aspect of the situation.

Our framework is that of Wald's theory of *statistical decision functions** [37]. As in that book and in Blackwell [3], we abstract the idea of a statistical experiment by a mathematical structure consisting of a set Θ, a σ-field $\mathscr{A}$ carried by a space χ, and a family $\mathscr{E} = \{P_\theta : \theta \in \Theta\}$ of probability measures on $\mathscr{A}$ (*see* Measure Theory in Probability and Statistics). (This is for one-stage experiments or for sequential ones where the stopping rule* is prescribed in advance; otherwise, additional structure is needed.) We don't care how the family $\mathscr{E}$ was created, whether by observing independent random variables or stochastic processes*, but only study how the likelihood ratios* depend on the parameter $\theta \in \Theta$.

The set Θ is usually called the *set of states of nature* or the *parameter set*. It need not have any special structure, though in many cases it is a Euclidean space.

To further specify a statistical problem, Wald defines a set D of possible decisions and a loss function W on $\Theta \times D$ (*see* Decision Theory). A decision procedure ρ is a rule that attaches to each observable point $x \in \chi$ a probability measure ρ_x on D. Having observed x, one selects a $d \in D$ according to ρ_x. Then the statistician suffers a loss $W(\theta, d)$ if θ is the true state of nature. The risk $R(\theta, \rho)$ of the procedure ρ is the expected value under P_θ of the sustained loss.

We can now define Gaussian shift experiments, first in the familiar case where Θ is a Euclidean space, and then, to cope with signal detection with noise* or nonparametric* situations, in a more general setup that covers infinite-dimensional parameter sets.

Also, we will introduce a concept of distance between two experiments $\mathscr{E}$ and $\mathscr{F}$ having the same parameter set Θ. The distance Δ is defined through comparison of the risk functions available on $\mathscr{E}$ and $\mathscr{F}$ respectively. Combining this with the Gaussian shift experiments and introducing a variable n that tends to infinity, one obtains definitions of asymptotically Gaussian (or normal) sequences of experiments.

We then look at the classical local asymptotic normality (LAN) conditions for this framework, and discuss a concept of weak convergence* of experiments by using the distance Δ taken on finite subsets of Θ. Two of the most valuable results of that theory, the Hájek—Le Cam asymptotic minimax and convolution theorems, employ weak convergence as a basic assumption.

Finally, we revisit the distance Δ on experiments, showing that for a number of cases it allows automatic transfer, in asymptotic situations, of results known in the Gaussian case. This is illustrated by recent results of Nussbaum [27] on density estimation*. He shows that most nonparametric density estimation problems are asymptotically equivalent to estimation of a signal in a Gaussian-white-noise problem. The use of the distance Δ

provides the statistician with a firm grip on possible rates of convergence of estimates.

GAUSSIAN SHIFT EXPERIMENTS

Every statistician is familiar with some form of what we call Gaussian (or, for emphasis, Gaussian shift) experiments. A common example is one where the observable variable is a random vector X taking values in some Euclidean space with norm $\| \cdot \|$. One has a function $m(\theta)$ from the parameter space Θ to the same Euclidean space, and the density of X (with respect to the Lebesgue measure) is proportional to $\exp[-\frac{1}{2} \parallel x - m(\theta) \parallel^2]$.

The square norm $\parallel x \parallel^2$ is often written $\parallel x \parallel^2 = x'Mx$, where M is a positive definite matrix. These norms are characterized by the median* equality

$$\parallel (x+y)/2- \parallel^2 + \parallel (x-y)/2 \parallel^2$$
$$= \tfrac{1}{2}[\parallel x \parallel^2 + \parallel y \parallel^2].$$

When a vector space is complete with respect to such a norm, it is a Hilbert space. Hilbert spaces are like Euclidean spaces except that they may be infinite-dimensional. They provide the basic framework for nonparametric estimation problems.

On Euclidean spaces there are more general experiments called heteroskedastic* Gaussian experiments. Their densities are proportional to

$$[\det M(\theta)]^2$$
$$\times \exp\{-\tfrac{1}{2}[X - m(\theta)]'M(\theta)[X - m(\theta)]\},$$

where M depends on θ. We shall not consider them, but concentrate instead on shift (or homoskedastic) experiments.

The name "Gaussian" refers to a theorem of C. F. Gauss [10]. He proved, in the one-dimensional case, that Gaussian experiments are the only shift families (*see* LOCATION-SCALE PARAMETER) where the average of the observations is the maximum-likelihood estimate* (*see* GAUSS, CARL FRIEDRICH).

There are Gaussian experiments where the observable entity is not a finite-dimensional vector X but an infinite-dimensional one or a stochastic process. For instance, the theory of signal detection (in noise) introduces situations where the observable entity is a stochastic process $\{X(t) : t \in T\}$ consisting of a "signal" $\theta(t)$ and a "noise" $W(t)$ so that $X(t) = \theta(t) + W(t)$. Here θ is an unknown nonrandom function, and W is a Gaussian process* in the sense that for every finite set $\{t_1, t_2, \ldots, t_k\}$ the vector $\{X(t_1), X(t_2), \ldots, X(t_k)\}$ has a normal distribution* (cf. Ibragimov and Has'minskii [14, p. 321]). The study of nonparametric or semiparametric* inference quickly introduces infinite-dimensional Gaussian experiments (cf. Bickel et al. [1]).

Let us take a closer look at the experiment $\mathscr{G} = \{G_\theta : \theta \in \Theta\}$ whose densities are proportional to $\exp[-\frac{1}{2} \parallel X - m(\theta) \parallel^2]$ on a Euclidean space. Consider the log likelihood $\Lambda(u; s) = \ln dG_u/dG_s$. If u runs through Θ and X has the distribution G_s, this gives a stochastic process. Assume, for convenience, that $m(s) = 0$. Then

$$\Lambda(u; s) = \langle m(u), X\rangle - \tfrac{1}{2} \parallel m(u) \parallel^2,$$

where $\langle x, y\rangle$ is the inner product associated with $\| \cdot \|$; i.e.,

$$\langle x, y\rangle = \parallel (x+y)/2 \parallel^2 - \parallel (x-y)/2 \parallel^2 .$$

Observe the following properties:

1. The process $u \to \Lambda(u; s)$ is a Gaussian process.
2. It stays Gaussian if, instead of G_s inducing the distribution of X, one uses G_θ, $\theta \in \Theta$. The covariance* kernel $\mathrm{Cov}[\Lambda(u; s), \Lambda(v; s)]$ is independent of the particular G_θ used.
3. Changing s to θ to induce the distributions of $\Lambda(u; s)$ adds $\mathrm{Cov}[\Lambda(u; s), \Lambda(\theta; s)]$ to the expectation of $\Lambda(u; s)$.

The first property suggests the following definition:

Definition 1. Let $\mathscr{G} = \{G_\theta : \theta \in \Theta\}$ be an experiment with parameter set Θ. It is called *Gaussian* if:

(i) the G_θ are mutually absolutely continuous* and

(ii) there is an $s \in \Theta$ such that $\Lambda(\cdot; s)$ is a Gaussian process under G_s.

It is easy to show that when (i) holds, if (ii) holds for some $s_0 \in \Theta$, then (ii) holds for all $s \in \Theta$. Furthermore, properties 2 and 3 above are automatic. The covariance kernel of the process is independent of s. Shifting the distributions to those induced by G_θ adds $\mathrm{Cov}[\Lambda(u; s), \Lambda(\theta; s)]$ to the expectation of $\Lambda(\theta; s)$ [23, p. 23].

Our Euclidean example had expectations $\mathbf{m}(\theta)$ in the same Euclidean space as $\mathbf{X}$. For the general case one needs Hilbert spaces. What takes the place of $\mathbf{X}$ will be a linear process.

Consider the space $\mathscr{M}_0$ of finite signed measures μ with finite support* on Θ and such that $\mu(\Theta) = 0$. The integrals $\int \Lambda(u; s)\mu(du)$ are finite linear combinations, with Gaussian distributions. We shall write them as $\langle \mu, Z \rangle$, where $\langle \cdot, \cdot \rangle$ is an inner product attached to the (Hilbert) square norm $\| \mu \|^2$, which is the variance of $\int \Lambda(u; s)\mu(du)$. To embed $\mathscr{M}_0$ into a true Hilbert space $\mathscr{H}$, equate to zero all μ such that $\| \mu \|^2 = 0$ and complete the resulting linear space.

Our likelihood ratio process becomes

$$\Lambda(u; s) = \langle (\delta_u - \delta_s), Z \rangle - \tfrac{1}{2} \| \delta_u - \delta_0 \|^2,$$

where δ_u is the probability measure that gives mass one to u. This can be extended to all of $\mathscr{H}$, giving new Gaussian measures $G(\mu)$ with log likelihood

$$\ln \frac{dG(\mu)}{dG(0)} = \langle \mu, Z \rangle - \tfrac{1}{2} \| \mu \|^2 .$$

If $\theta' \neq \theta''$ implies $G_{\theta'} \neq G_{\theta''}$, then the map $\theta \rightsquigarrow \delta_\theta - \delta_s$ is an embedding of Θ onto a certain subset Θ^* of $\mathscr{H}$.

The map $\mu \rightsquigarrow \langle \mu, Z \rangle$ is often called the canonical Gaussian process of $\mathscr{H}$. It is characterized (up to equivalence) by the property that it is linear, that the variable $\langle \mu, Z \rangle$ is an ordinary random variable with a normal distribution, and that under G_s, $\langle \mu, Z \rangle$ has expectation zero and variance $\| \mu \|^2$. (As before, variances and covariances do not depend on which $G(\upsilon)$, $\upsilon \in \mathscr{H}$, is used to induce the distributions.)

Assuming that $\theta' \neq \theta''$ implies $G_{\theta'} \neq G_{\theta''}$, the statistical properties of the original experiment $\mathscr{H}$ depend only on the geometric properties of the image Θ^* of Θ in $\mathscr{H}$. More specifically, let $\mathscr{G}_1 = \{G_u : u \in \Theta_1\}$ and $\mathscr{G}_2 = \{G_\upsilon : \upsilon \in \Theta_2\}$ be two Gaussian experiments with parameter spaces Θ_1 and Θ_2 respectively. Introduce the square Hellinger distance*

$$H^2(u, \upsilon) = \tfrac{1}{2} \int (\sqrt{dG_u} - \sqrt{dG_\upsilon})^2.$$

A simple computation shows that the variance of $\Lambda(u; s) - \Lambda(\upsilon; s)$ is given by $-8 \ln[1 - H^2(u, \upsilon)]$. Thus an isometry (distance-preserving transformation) of the Θ_i^* corresponds to an isometry of the initial spaces Θ_i for H.

More concretely, take a point $s_1 \in \Theta_1$, and let s_2 be the point of Θ_2 that corresponds to s_1 in the isometry for H. Carry out the procedure described above for the log likelihoods $\Lambda_i(u; s_i)$. This will give spaces $\mathscr{H}_i$ and processes Z_i. The isometry between the Θ_i^* then extends to a linear isometry of the spaces $\mathscr{H}_i$ (cf. Le Cam [22, p. 239]). The process Z_1 is carried by the isometry to a process Z_1' that has the same distributions as Z_2. Thus the experiments $\mathscr{G}_1$ and $\mathscr{G}_2$ differ only by a relabeling of their parameter sets. Such relabelings are used very frequently. For instance $\mathscr{G}_1$ may be parametrized by a set of square-integrable functions f on the line, but one can then reparametrize it by the Fourier transforms $\tilde{f}$ of f. Whether one represents a square-integrable f by its Fourier coefficients or by its coefficients in some other orthonormal basis does not change anything essential.

For a simple example, consider the signal-plus-noise problem $X(t) = \theta(t) + \tilde{W}(t)$, where $t \in [0, 1]$ and $\tilde{W}$ is the standard Wiener process (*see* BROWNIAN MOTION). Assume that $\theta(t) = \int_0^t f(c)dc$ for a square-integrable function f. Then the associated Hilbert space $\mathscr{H}$ is the space L_2 of equivalence classes of square-integrable functions with the squared norm $\| f \|^2 = \int_0^1 f^2(c)dc$. Let Z denote standard white noise (equal to the "derivative" of $\tilde{W}$), and then the inner product $< f, Z >$ is $\int f(t)Z(dt)$.

Actually the experiment obtained in that manner is the archetype of all separable

Gaussian experiments. [Here "separable" means that the set $\{G_\theta : \theta \in \Theta\}$ contains a countable dense set for the Hellinger distance*, or equivalently, for the L_1-norm $\| G_u - G_v \| = \sup_{|\varphi| \leqslant 1} \int \varphi(dG_u - dG_v)$.] Any separable Gaussian experiment can be represented by taking for Θ^* a suitable subset of the space L_2, with the prescription that the observable process is $dX(t) = f_\theta(t)dt + Z(dt)$.

A DISTANCE BETWEEN EXPERIMENTS

Let $\mathscr{E} = \{P_\theta : \theta \in \Theta\}, \mathscr{F} = \{Q_\theta : \theta \in \Theta\}$ be two experiments with the same parameter set Θ. They can have observables in different spaces, say $(\chi, \mathscr{A})$ for $\mathscr{E}$ and $(\mathscr{Y}, \mathscr{B})$ for $\mathscr{F}$. The two experiments can be considered close to each other in a statistical sense if for any arbitrary pairs (D,W) of decision spaces D with loss function W, a risk function available on one of the two experiments can be closely matched by a risk function of the other experiment.

The idea of such possible matchings goes back to Wald [36]. He considered i.i.d. observations and matched any measurable set in the space of n observations to one in the space of maximum-likelihood estimates. His matching is such that the probabilities of the two sets differ little, as $n \to \infty$, uniformly in θ. Wald's demands on this set-to-set relation are too strict; they can be satisfied only under special circumstances.

Consider arbitrary decision spaces D and arbitrary loss functions W on $\Theta \times D$, but *restrict attention to functions W such that* $0 \leqslant W(\Theta, d) \leqslant 1$. Let $\mathscr{R}(\mathscr{E}, W)$ be the set of functions from Θ to $[0, \infty]$ that are possible risk functions for $\mathscr{E}$ and W, or at least as large as such possible risk functions (the use of functions larger than actually possible risk functions is a technical convenience). Instead of working with $\mathscr{R}(\mathscr{E}, W)$ we shall work with its closure $\overline{\mathscr{R}}(\mathscr{E}, W)$ for pointwise convergence on Θ.

In most cases $\mathscr{R}$ itself is already closed and thus equal to $\overline{\mathscr{R}}$. This happens for instance if the family $\{P_\theta : \theta \in \theta\}$ is dominated and if the W are lower semicontinuous in d on a decision space D that is locally compact. We shall call $\overline{\mathscr{R}}(\mathscr{E}, W)$ the *augmented* space of risk functions for $\mathscr{E}$ and W.

Definition 2. For $\epsilon \in [0,1]$, the deficiency $\delta(\mathscr{E}, \mathscr{F})$ of $\mathscr{E}$ with respect to $\mathscr{F}$ does not exceed ϵ if for every D, every W such that $0 \leqslant W \leqslant 1$, and every $g \in \mathscr{R}(\mathscr{F}, W)$ there is an $f \in \overline{\mathscr{R}}(\mathscr{E}, W)$ such that $f(\theta) \leqslant g(\theta) + \epsilon$ for all $\theta \in \Theta$. The *deficiency* $\delta(\mathscr{E}, \mathscr{F})$ is the minimum of the numbers ϵ with that property.

Definition 3. The *distance* $\Delta(\mathscr{E}, \mathscr{F})$ between $\mathscr{E}$ and $\mathscr{F}$ is the number

$$\Delta(\mathscr{E}, \mathscr{F}) = \max\{\delta, (\mathscr{E}, \mathscr{F}), \delta(\mathscr{F}, \mathscr{E})\}.$$

Actually, Δ is not a distance but a pseudometric. It satisfies the triangle inequality (for given Θ), but there are always pairs $(\mathscr{E}, \mathscr{F})$ such that $\mathscr{E}$ and $\mathscr{F}$ are different but $\Delta(\mathscr{E}, \mathscr{F}) = 0$. This happens for instance if $\mathscr{E}$ is an experiment with observations $X_1, X_2, \ldots, X_n$ and $\mathscr{F}$ uses only a sufficient statistic*$T(X_1, X_2, \ldots, X_n)$ instead of the whole set $\{X_1, X_2, \ldots, X_n\}$.

To reiterate, if $\Delta(\mathscr{E}, \mathscr{F}) \leqslant \epsilon$, this means that for all D and all loss functions W subject to $0 \leqslant W \leqslant 1$, every possible function in the (augmented) set of risk functions for one experiment can be matched within ϵ by a function in the (augmented) set of risk functions for the other experiment.

The introduction of "deficiencies" and distances occurs in Le Cam [19]. Blackwell [3] and Stein [31] had previously considered the case where $\delta(\mathscr{E}, \mathscr{F}) = 0$, in which case one says that $\mathscr{E}$ is "better" or "stronger" than $\mathscr{F}$. The deficiency can be computed in terms of Bayes' risks. It is also obtainable by a *randomization* criterion. Let T be a linear transformation that transforms positive measures into positive measures of the same mass. Then

$$\delta(\mathscr{E}, \mathscr{F}) = \inf_T \sup_\theta \tfrac{1}{2} \| Q_\theta - TP_\theta \|,$$

where $\| m \|$ is the L_1-norm $\| m \| = \sup_\varphi \{\int \varphi dm : |\varphi| \leqslant 1\}$. Those linear transformations are limits of transformations obtainable through randomizations by Markov kernels.

A *Markov kernel* from $(\chi, \mathscr{A})$ to $(\mathscr{Y}, \mathscr{B})$ is a function $x \rightsquigarrow \chi$ and a probability measure K_x on $\mathscr{B}$ such that the functions $x \rightsquigarrow K_x(B)$ are $\mathscr{A}$-measurable or at least equivalent to $\mathscr{A}$-measurable functions. The corresponding transformation T is given by $(TP)(B) = \int K_x(B)P(dx)$.

The distance Δ applies to experiments $\mathscr{E}$ and $\mathscr{F}$ with the same parameter set Θ, but one can use it to define other distances: If S is a subset of φ, let $\mathscr{E}_s = \{P_\theta : \theta \in S\}$ be the experiment $\mathscr{E}$ with parameter set restricted to S. One can compute a distance $\Delta(\mathscr{E}_s, \mathscr{F}_s)$.

For asymptotic purposes one treats two sequences $\{\mathscr{E}_n\}$ and $\{\mathscr{F}_n\}$, with parameter set Θ_n. These sequences are asymptotically equivalent if $\Delta(\mathscr{E}_n, \mathscr{F}_n) \to 0$ as $n \to \infty$. For instance one can say that the $\{\mathscr{E}_n\}$ are *asymptotically Gaussian* if there are Gaussian experiments $\mathscr{G}_n = \{G_{\theta,n} : \theta \in \Theta_n\}$ such that $\Delta(\mathscr{E}_n, \mathscr{G}_n) \to 0$ as $n \to \infty$. Le Cam and Yang [23] say that the $\mathscr{E}_n$ are *weakly asymptotically Gaussian* if there are Gaussian $\mathscr{G}_n$ such that $\Delta(\mathscr{E}_{n,S_n}, \mathscr{G}_{n,Sn}) \to 0$ as long as the cardinality of $S_n \subset \Theta_n$ remains bounded above independently of n.

There is also a topology of weak convergence of experiments.

Definition 4. Let $\{\mathscr{E}_n\}$ be a sequence of experiments with common parameter set Θ. Let $\mathscr{F} = \{Q_\theta : \theta \in \Theta\}$ be another experiment with the same parameter set. One says that $\mathscr{E}_n \to \mathscr{F}$ weakly if for every *finite* subset $S \subset \Theta$ the distance $\Delta(\mathscr{E}_{n,S}, \mathscr{F}_S)$ tends to zero.

Weak convergence of experiments is equivalent to convergence in the ordinary sense of distributions of likelihood ratios [24, pp. 10–15]. One takes a finite set $\{\theta_0, \theta_1, \ldots, \theta_m\}$, and looks at the vector $\{dP_{\theta_{i,n}}/dP_{\theta_0,h} : i = 0, 1, 2, \ldots, m\}$ with the distributions induced by $P_{\theta_{0,n}}$. These should converge in the ordinary sense to the distributions of $\{dQ_{\theta_i}/dQ_{\theta_0} : i = 0, 1, \ldots, m\}$ under Q_{θ_0}. (Note that θ_0 is an *arbitrary* point in an *arbitrary* finite set. So the distributions under $P_{\theta_{i,n}}$ also converge. At times it is enough to consider only one particular θ_0, for instance if the $\{P_{\theta_{i,n}}\}$ and $\{P_{\theta_0,n}\}$ are contiguous sequences.)

Besides Δ, several other distances have been used. Torgersen [33,34] has made a deep study of a distance Δ_k defined exactly like Δ, but restricting the decision space D to have at most k elements (k decision problems) or even two elements (test problems). As k increases, so do the Δ_k; and Δ itself is $\sup_k \Delta_k$.

Convergence for the distance Δ is also linked to convergence in distribution of stochastic processes as defined by Prohorov [30]. Suppose $\mathscr{E} = \{P_\theta : \theta \in \Theta\}$ and $\mathscr{F} = \{Q_\theta : \theta \in \Theta\}$ can be "coupled" to be on the same space and that they are dominated by a suitable probability measure μ. Let X be the process $X(\tau) = dP_\tau/d_\mu$ and let $Y(\tau) = dQ_\tau/d_\mu$ for $\tau \in \Theta$. Suppose that the two experiments are not only coupled on the same space but that they are coupled so that

$$\tfrac{1}{2} \sup_\tau E_\mu |X(\tau) - Y(\tau)| \leqslant \epsilon.$$

Then $\Delta(\mathscr{E}, \mathscr{F}) \leqslant \epsilon$. (An example of coupling occurred in our discussion of isometries for the parameter set of Gaussian experiments.)

By comparison, using the Skorohod embedding theorem (*see* SKOROHOD EMBEDDINGS) or Strassen's theorems [32], convergence of stochastic processes (for the uniform norm) would correspond to convergence to zero of $E_\mu \sup_\tau \{|X(\tau) - Y(\tau)| \wedge 1\}$, a much more delicate affair, since $E_\mu \sup_\tau$ is larger than $\sup_\tau E_\mu$.

LOCAL ASYMPTOTIC NORMALITY: THE LAN CONDITIONS

There are many situations in which one fixes a parameter value θ_0 and studies the behavior of experiments $\mathscr{E}_n = \{P_{\theta,n} : \theta \in \Theta\}$ in small shrinking neighborhoods of θ_0. For i.i.d. data, a famous example is Cramér's work on the roots of maximum-likelihood equations [4, p. 500]. Another example, for Markov processes*, is given by Billingsley [2].

Cramér imposed differentiability conditions on $f(x, \theta)$. An alternative approach, taken by Hájek and Le Cam, assumes that the densities $f(x, \theta)$ are defined with respect to some measure μ and requires the existence of a derivative in μ-quadratic mean of $\sqrt{f(x, \theta)}$. That is, one requires the existence of random vectors $Y(\theta_0)$ such that

$$\int \left(\frac{1}{|t|} \left| \sqrt{f(x, \theta_0 + t)} - \sqrt{f(x, \theta_0)} - \tfrac{1}{2} t' Y(\theta_0) \right| \right)^2 d\mu$$

tends to zero as $|t| \to 0$. The function $Y(\theta_0)$ is then square-integrable, and $\int |t' Y(\theta_0)| d\mu$ is equal to $t' J(\theta_0) t$, where $J(\theta_0)$ is the Fisher information* matrix $\int Y(\theta_0) Y'(\theta_0) d\mu$.

One relies upon local expansions of the log likelihood to study the behavior of these experiments. More precisely, assume that Θ is a subset of a vector space V with norm $|\cdot|$. One selects a sequence $\{\delta_n\}$ of linear maps from V to V and looks at the experiments

$$\mathscr{F}_{\theta_0,n} = \{P_{\theta_0+\delta_n t}, n : t \in V, \theta_0 + \delta_n t \in \Theta\}.$$

It is usually assumed that $\delta_n t \to 0$ as $n \to \infty$ for fixed $t \in V$. In the following, we always require that $\theta_0 + \delta_n t_n \in \Theta$.

For the Cramér conditions, or Hájek and Le Cam's condition of differentiability in quadratic mean, the space V is Euclidean and the maps δ_n are simply multiplications by $1/\sqrt{n}$, so that $\delta_n t = t/\sqrt{n}$. In contrast, the LAN* conditions of Le Cam [18] do not refer to i.i.d. or Markov processes. They ignore how the $P_{\theta,n}$ were arrived at and focus on the logarithm of likelihood ratios

$$\wedge_n(t) = \ln \frac{dP_{\theta_0} + \delta_n t, n}{dP_{\theta_0,n}}.$$

To state the conditions we shall use a particular Hilbertian or Euclidean norm $\|\cdot\|$ with its associated inner product $<\cdot,\cdot>$. We shall also need a contiguity* condition: Two sequences $\{P_n\}$ and $\{Q_n\}$ of probability measures on σ-fields $\mathscr{A}_n$ are called *contiguous* if sequences $\{X_n\}$ that tend to zero in probability for one of the sequences of measures also tend to zero for the other sequence.

The LAN conditions are as follows, with V an arbitrary vector space with norm $|\cdot|$:

LAN (1). There are random variables X_n and Hilbertian norms $\|\cdot\|$ such that whenever the nonrandom sequence $\{t_n\}$ tends to a limit t in the sense that $|t_n - t| \to 0$, the difference

$$\bigwedge_n (t_n) - \langle t_n, X_n\rangle + \tfrac{1}{2} \parallel t_n \parallel^2$$

tends to zero in $P_{\theta_0,n}$ probability.

LAN (2). If $|t_n - t| \to 0$, then the sequences $\{P_{\theta_0} + \delta_n t_n, n\}$ and $\{P_{\theta_0}, n\}$ are contiguous.

One doesn't want the sets $\{t : t \in V, \theta_0 + \delta_n t \in \Theta\}$ to be excessively small. To prevent this, assume:

LAN (3). If V is finite-dimensional, the limit points $t = \lim_n t_n$ with $\theta_0 + \delta_n t_n \in \Theta$ are dense in an open subset of V. If V is infinite-dimensional, the same holds for every finite-dimensional subspace of V.

In these conditions the random X_n, the maps δ_n, and the norms $|\cdot|$ and $\|\cdot\|$ can depend on the particular θ_0 selected for attention. For instance, in the Hájek-Le Cam approach one can take $\parallel \mathbf{t} \parallel^2 = \mathbf{t}'\mathbf{J}(\theta_0)\mathbf{t}$ or one can take $\parallel \mathbf{t} \parallel^2 = \mathbf{t}'\mathbf{t}$ but make $\delta_n = \delta_n(\theta_0)$ so that $[\delta_n(\theta)_0\delta'_n(\theta_0)]^{-1} = n\mathbf{J}(\theta_0)$.

Le Cam [18] contains a fourth, nonlocal requirement for estimates that converge at a suitable speed. If for each $\theta \in \Theta$ one has selected a sequence $\delta_n(\theta)$ of linear maps, the condition is as follows:

LAN (4). There exist estimates $\tilde{\theta}_n$ such that for each $\theta \in \Theta$ the norms $|\delta_n^{-1}(\theta)(\tilde{\theta}_n - \theta)|$ stay bounded in $P_{\theta,n}$ probability.

This permits the construction of asymptotically sufficient estimates (see Le Cam and Yang [24, §5.3]).

Together, these four conditions force Θ to be finite-dimensional. But although they were originally designed for Euclidean subsets Θ, the first three LAN conditions can be used for the infinite-dimensional spaces V needed in the study of signal-plus-noise processes or nonparametric investigations.

Let K be a compact subset of $(V, |\cdot|)$. Let $\mathscr{F}_{\theta_0,n}(K) = \{P_{\theta_0+\delta_n t,n} : t \in K, \theta_0 + \delta_n t \in \Theta\}$. Then the conditions LAN (1), (2), (3) imply the following:

(A) There are Gaussian experiments

$$\mathscr{G}_{\theta_0,n}(K) = \{G_{t,n} : t \in K, \theta_0 + \delta_n t \in \Theta\}$$

such that $\Delta[\mathscr{F}_{\theta_0,n}(K), \mathscr{G}_{\theta_0,n}(K)] \to 0$ as $n \to \infty$ for the distance Δ of Definition 3.

(B) For the norm $\|\cdot\|$ of LAN (1) and suitable random variables W_n one has

$$\ln \frac{dG_{t,n}}{dG_{0,n}} = \langle t, W_n\rangle - \tfrac{1}{2} \parallel t \parallel^2$$

with $\langle t, W_n\rangle$ distributed as $N(0, \parallel t \parallel^2)$. Thus the $\mathscr{G}_{\theta_0,n}(K)$ reflect the linear structure of V.

Conversely, if (A) and (B) hold, so do LAN (1) and LAN (2). But LAN (3) and LAN (4) cannot be consequences of (A) and (B).

For an example in the nonparametric i.i.d. case we work with densities f defined with respect to Lebesgue measure. For any integer $n \geqslant 1$ the $\sqrt{f}$ can be written in the form

$$\sqrt{f} = 1 - c\left(\frac{\| \upsilon_n \|}{2\sqrt{n}}\right) + \frac{\upsilon_n}{2\sqrt{n}},$$

where υ_n is in the space $L_{2,0}$ of functions υ such that $\int \upsilon d\lambda = 0$ and $\| \upsilon \|^2 = \int \upsilon^2 d\lambda < \infty$. Here c is a function from [0, 1] to [0, 1] defined by $[1 - c(z)]^2 = 1 - z^2$; it is used to ensure $\int f d\lambda = 1$. We restrict attention to the subset $L_{2,0}(n)$ of $L_{2,0}$ where

$$1 - c(\| \upsilon_n \| /2\sqrt{n}) + \upsilon_n/2\sqrt{n} \geqslant 0$$

and $\| \upsilon_n \|^2 \leqslant 4n$.

Consider subsets Θ_n of $L_{2,0}(n)$, and take n independent observations from the density

$$\left[1 - c\left(\frac{\| \upsilon \|}{2\sqrt{n}}\right) + \frac{\upsilon}{2\sqrt{n}}\right]^2.$$

It can be shown that such experiments satisfy LAN (1), (2) and (A), (B) for the squared norm $\| \upsilon \|^2 = \int \upsilon^2 d\lambda$. The notation has already rescaled the densities, so that $| \cdot |$ and δ_n do not appear, or one can take $| \cdot | = \| \cdot \|$ and δ_n to be the identity. This is all that is needed in most of the arguments and examples of Bickel et al. [1].

THE ASYMPTOTIC MINIMAX AND CONVOLUTION THEOREMS

Hájek [11,12] weakened LAN (1) and LAN (2) by letting $t_n = t$, independent of n. This necessitated the following strengthening of LAN (3).

1. There is a dense subset D of the metric space $(V, | \cdot |)$ such that for $t \in D$ one has $\theta_0 + \delta_n t \in \Theta$ for all n larger than some $n(t)$.

These weakened LAN (1), (2) conditions, together with (H), will be called the Hájek conditions. They do not imply statement (A) of the previous section, but they do imply the following:

Proposition. Under Hájek's conditions the experiments

$$\mathscr{F}_{\theta_0,n}(D) = \{P_{\theta_0+\delta_n+t,n} : t \in D, \theta_0 + \delta_n t \in \Theta\}$$

converge weakly to a Gaussian experiment $\mathscr{G} = \{G_t : t \in D\}$ with log likelihood

$$L(t) = \langle t, Z\rangle - \tfrac{1}{2} \| t \|^2 .$$

The weak convergence is that in Definition 4.

Such a weak convergence already has some interesting statistical consequences. One of them is as follows:

Asymptotic Minimax Theorem. Let $\mathscr{F}_n = \{Q_{t,n} : t \in D\}$ be experiments that converge weakly to a limit $\mathscr{G} = \{G_t : t \in D\}$, not necessarily Gaussian. Fix a loss function W that is bounded below for each fixed t. Let $\mathscr{R}(\mathscr{F}_n, W)$ be the (augmented) set of risk functions introduced in Definition 2, and let r be a function that *does not* belong to $\overline{\mathscr{R}}(\mathscr{G}, W)$. Then there is a finite set $F \subset D$, an $\alpha > 0$, and $N < \infty$ such that if $n \geqslant N$, the function $r + \alpha$ does not belong to $\overline{\mathscr{R}}(\mathscr{F}_{n,F}, W)$, where $\mathscr{F}_{n,F} = \{Q_{t,n} : t \in F\}$.

This result leads directly to Hájek's version of the asymptotic minimax theorem [12]. A stronger form of it is proved in Le Cam [20; 22, pp. 109–110]; for the i.i.d case, under Cramér-type conditions, see [17]. The theorem in [20] relies entirely on weak convergence of experiments and does not use the fact that the limit is Gaussian.

By contrast, Hájek's version of the convolution theorem relies on the fact that the limit a Gaussian shift experiment:

Convolution Theorem. Let V be a Euclidean space, and let Hájek's conditions be satisfied. Let $\mathscr{F}_n = \{Q_{t,n} : t \in g\}$ converge weakly to a Gaussian experiment with log-likelihood ratios

$$\ln \frac{dG_t}{dG_0} = \langle t, Z\rangle - \tfrac{1}{2} \| t \|^2 .$$

Consider statistics T_n defined on $\mathscr{F}_n$ and such that $\mathscr{L}[T_n - t|Q_{t,n}]$ has a limit M independent of t for all $t \in S$. Then there is a probability measure π such that M is the convolution $\pi * G$, where $G = \mathscr{L}(Z)$.

This was proved by Hájek [11]; *see also* HÁJEK–INAGAKI CONVOLUTION THEOREM. It also applies to linear functions AT_n and AZ; the limiting distribution of $\mathscr{L}(AT_n)$ is a convolution of $\mathscr{L}(AZ)$. The theorem has been extended to infinite-dimensional (nonparametric) cases where V is a Hilbert space, with the norm $\| \cdot \|$ that occurs in LAN (1) (see Moussatat [26], Millar [25], and van der Vaart [35]).

The convolution theorem has been taken by many as a definition of optimality of estimates. If the T_n are such that $M = \pi * G$ with π concentrated at a point, they are called *optimal*, as other possible limiting distributions are more dispersed.

The Hájek-Le Cam asymptotic minimax and convolution theorems are widely used in asymptotic statistics. But they don't tell the whole story, as is shown in the following section. One of the reasons they cannot tell the whole story is that these theorems can yield only *lower* bounds for the risk of estimates. They cannot yield upper bounds. This is because they are really "local" theorems, and cannot imply such things as the global condition LAN (4). But they do imply, for example, that in the nonparametric i.i.d. case there are no estimates $\tilde{f}_n$ that can identify the square root of densities in such a way that $(n/2)\int(\sqrt{\tilde{f}_n} - \sqrt{f})^2 d\lambda$ stays bounded in probability for all densities f. We shall elaborate on this in the next section.

SOME GLOBAL ASYMPTOTIC NORMALITY SITUATIONS

The Hájek-Le Cam theorems yield lower bounds for the risk of estimates. Convergence in the strong sense of the distance Δ can yield both lower and upper bounds. In principle, this is possible only for bounded loss functions W; however, for unbounded loss functions (such as ordinary quadratic loss), if $\Delta(\mathscr{E}_n, \mathscr{F}_n) \to 0$ then one can truncate W to W_n such that $0 \leqslant W_n \leqslant b_n$ with $b_n\Delta(\mathscr{E}_n, \mathscr{F}_n)$ tending to zero. This is usually sufficient for practical purposes.

As an example of the need for strong convergence, consider the i.i.d. case where $P_{\theta,n}$ is the joint distribution of n variables $X_1, X_2, \ldots, X_n$. Suppose each X_j is Cauchy* with center θ. Let $\mathscr{E}_n = \{P_{\theta,n} : \theta \in \mathbb{R}\}$ and $\mathscr{G}_n = \{G_{\theta,n} : \theta \in \mathbb{R}\}$, where $G_{\theta,n}$ is the distribution of Y, a $N(\theta, 2/n)$ random variable.

The $\mathscr{E}_n$ converge weakly to a limit where for $\theta \neq \theta'$ the corresponding measures are disjoint. This gives little information. For a fixed θ_0, say $\theta_0 = 0$, the rescaled experiments $\mathscr{F}_n = \{P_{\delta_n t,n} : t \in \mathbb{R}\}$ with $\delta_n t = t/\sqrt{n}$ satisfy the conditions of LAN (1), (2), and (3). By itself, this does not imply LAN (4), but it is sufficient to yield lower bounds on the asymptotic risk of estimates.

In this case one can prove that $\Delta(\mathscr{E}_n, \mathscr{G}_n) \to 0$ as $n \to \infty$. In fact one can prove that $\Delta(\mathscr{E}_n, \mathscr{G}_n) \leqslant C/\sqrt{n}$ for a suitable constant C. This says that the risk of estimates for $\mathscr{E}_n$ will behave like estimates for the Gaussian $\mathscr{G}_n$, except for a term of the type

$$C \parallel W_n \parallel /\sqrt{n},$$

where $\parallel W_n \parallel = \sup_{\theta,d} |W_n(\theta, d)|$. A similar assertion can be made for the general i.i.d. location family case, provided the density $f(x - \theta)$ has finite Fisher information.

As a more interesting problem, Nussbaum [27] considers a nonparametric situation where the variables are i.i.d. with densities f on [0,1]. Let Θ be the space of densities such that (1) for a fixed constant K and some fixed $\alpha > \frac{1}{2}$,

$$|f(x) - f(y)| \leqslant K|x - y|^{\alpha},$$

and (2) there exists $\epsilon_0 > 0$ such that $f(x) \geqslant \epsilon_0$ for all $x \in [0, 1)$. Let $\mathscr{E}_n = \{P_{f,n} : f \in \theta\}$ be the experiment where $P_{f,n}$ is the joint distribution of n independent observations from $f \in \Theta$. Consider the Gaussian experiment $\mathscr{G}_n = \{G_{f,n} : f \in \Theta\}$ where the observable element is a process $Y(\tau)$ for $\tau \in [0, 1]$ such that

$$dY(\tau) = \sqrt{f(\tau)}d\tau + \frac{1}{2\sqrt{n}}Z(d\tau),$$

where Z is the standard Gaussian white noise* on [0, 1]. Nussbaum's result is as follows:

Theorem. As $n \to \infty$ the distance $\Delta(\mathscr{E}_n, \mathscr{G}_n)$ tends to zero.

Note that the shift parameter of $\mathscr{G}_n$ is the square root of the density. As in the nonparametric example of the previous section, this ensures homoskedasticity of $\mathscr{G}_n$.

People know a lot about estimation in the Gaussian case (see Donoho et al. [8], Donoho and Liu [7], Donoho and Johnstone [6], and the references therein). Using Nussbaum's theorem, one can readily transfer all of these Gaussian results to the case of density estimation in Θ, at least for bounded loss functions (and for unbounded ones, by truncation). It follows for instance that the speed of convergence of estimates will be the same in $\mathscr{E}_n$ as in $\mathscr{G}_n$.

Nussbaum gives many applications; we only cite one. Let $W_2^m(K)$ be the Sobolev ball of functions g where for the standard Fourier orthonormal basis on [0, 1] the Fourier coefficients g_j satisfy $\sum(2nj)^{2m}g_j^2 \leqslant K$. Consider a truncated Hellinger loss

$$L(\hat{f}, f, n, c) = \min\{c, n^{1-r}\int(\hat{f}^{1/2} - f^{1/2})^2 d\lambda\},$$

where $r = (2m+1)^{-1}$.

Let $\mathscr{F}(m, k, \epsilon)$ be the set of densities f such that $f \geqslant \epsilon > 0$ and $f^{1/2} \in W_2^m(K)$. Then for each $\epsilon > 0$ and for $m \geqslant 4$ there is a sequence c_n tending to infinity such that the minimax risk on $\mathscr{F}(m, k, \epsilon)$ for $L(\hat{f}, f, n, c_n)$ tends to $2^{2(r-1)}K^r\gamma(m)$, where $\gamma(m)$ is the Pinsker constant [28] relative to the Gaussian case. This result had been guessed but not proved before Nussbaum's theorem.

RELATED RESULTS

There are situations where approximability by Gaussian shift experiments is not possible, but where the Gaussian results are still useful. For inference on stochastic processes or time series, the LAN conditions may apply, but the locally asymptotically mixed normal (LAMN) conditions apply more generally (cf. [5,15,16]). Here the log-likelihood ratios have an expansion of the type

$$\wedge_n(\mathbf{t}) = \langle \mathbf{t}, S_n \rangle - \tfrac{1}{2}\mathbf{t}'\mathbf{\Gamma}_n\mathbf{t} + \epsilon_n,$$

where ϵ_n tends to zero in probability but where the matrices, or linear maps, $\mathbf{\Gamma}_n$ stay random. They are not approximable by nonrandom maps. Nonetheless, a large number of Gaussian results can be applied *conditionally* given the $\mathbf{\Gamma}_n$.

Several bounds on risks can be obtained through the study of Bayes' procedures when the posterior distributions are approximately Gaussian (see Le Cam and Yang [24, §5.4]). These results stay valid in the LAMN case. For a discussion of such Gaussian approximations to posterior distributions, see Jeganathan [15,16] and Le Cam [22, Chap. 12].

REFERENCES

1. Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins Series in the Mathematic Sciences.
2. Billingsley, P. (1961). *Statistical Inference for Markov Processes*. University of Chicago Press, Chicago.
3. Blackwell, D. (1951). Comparison of experiments. *Proc. 2nd Berkeley Symp. Math. Statist. Probab*. J. Neyman, ed. University of California Press, Berkeley, pp. 93–102.
4. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
5. Davies, R. (1985). Asymptotic inference when the amount of information is random. *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, vol. II, L. Le Cam and R. Olshen, eds. Wiley, New York, pp. 241–264.
6. Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann Statist.*, **26**, 879–921.
7. Donoho, D. L. and Liu, R. (1991). Geometrizing rates of convergence, II and III. *Ann. Statist.*, **19**, 633–701.
8. Donoho, D. L., Liu, R., and MacGibbon, B. (1990). Minimax risk over hyperrectangles and implications. *Ann. Statist.* **18**, 1416–1437.
9. Dudley, R. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, **6**, 899–929.
10. Gauss, C. F. (1809). *The Heavenly Bodies Moving about the Sun in Conic Sections (reprint)*. Dover, New York, 1963.
11. Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. u. Verw. Geb.* **14**, 323–330.
12. Háek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, vol. 1, L. Le Cam, J. Neyman, and E. Scott, eds. University of California Press, pp. 175–194.
13. Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.
14. Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*. Springer-Verlag, New York.

15. Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log likelihood is mixed normal. *Sankhyā Ser. A*, **44, (part 2)**, 173–212.
16. Jeganathan, P. (1983). Some asymptotic properties of risk functions when the limit of the experiments is mixed normal. *Sankhyā Ser. A*, **45**, 66–87.
17. Le Cam, L. (1953). On some asymptotic properties of maximum-likelihood and related Bayes' estimates. *Univ. Calif. Publ. Statist.*, **1**, 277–330.
18. Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.*, **3**, 37–98.
19. Le Cam, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.*, **35**, 1419–1455.
20. Le Cam, L. (1979). On a theorem of J. Hájek. In *Contribution to Statistics: J. Hájek Memorial Volume*, J. Jurečková, ed. Akademia Praha, Czechoslovakia, pp. 119–137.
21. Le Cam, L. (1985). Sur l'approximation de familles de mesures par des familles gaussiennes. *Ann. Inst. Henri Poincaré, Probab. Statist.*, **21**, 225–287.
22. Le Cam. L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
23. Le Cam, L. and Yang, G. L. (1988). On the preservation of local asymptotic normality under information loss. *Ann. Statist.*, **16**, 483–520.
24. Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics*, Springer-Verlag, New York.
25. Millar, P. W. (1985). Nonparametric applications of an infinite dimensional convolution theorem. *Z. Wahrsch. u. Verw. Geb.*, **68**, 545–556.
26. Moussatat, W. (1976). *On the asymptotic theory of statistical experiments and some of its applications*. Ph.D. thesis, University of California, Berkeley.
27. Nussbaum, M. (1996). Asymptotic equivalence of density estimation and white noise. *Ann. Statist.*, **24**, 2399–2430.
28. Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Prob. Inf. Transmission*, **1**, 120–123.
29. Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.
30. Prohorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability. *Theory Probab. Appl.*, **1**, 157–214.
31. Stein, C. (1951). *Notes on the Comparison of Experiments*, University of Chicago Press.
32. Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.*, **36**, 423–439.
33. Torgersen, E. N. (1970). Comparison of experiments when the parameter space is finite. *Z. Wahrsch. u. Verw. Geb.*, **16**, 219–249.
34. Torgersen, E. N. (1991). *Comparison of Statistical Experiments*. Cambridge University Press.
35. van der Vaart, A. (1988). *Statistical Estimation in Large Parameter Spaces*. Tract 44, Centrum voor Wiskunde en Informatica, Amsterdam.
36. Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426–482.
37. Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.

See also ABSOLUTE CONTINUITY; ASYMPTOTIC EXPANSIONS—II; CONTIGUITY; DECISION THEORY; GAUSSIAN PROCESSES; HÁJEK–INAGAKI CONVOLUTION THEOREM; MEASURE THEORY IN PROBABILITY AND STATISTICS; and MINIMAX DECISION RULES.

L. LE CAM

ASYMPTOTICS, HIGHER ORDER

Statisticians often seek to approximate quantities, such as the density of a test statistic evaluated at a fixed ordinate, that depend on a known parameter, such as the sample size, in a very complicated way. The resulting approximation should be tractable analytically or numerically. Asymptotic expansions* are approximations that can be expressed as a sum of a small number of terms, each of which is the product of a factor that is a simple function of the parameter, and a coefficient that does not depend on the parameter. Asymptotic expansions are distinguished from other potential approximations in that their accuracy is assessed by examining the limiting behavior of errors as the parameter approaches some limiting value. This limiting value is usually infinity in the very common case in which the known parameter is a measure of sample size.

Specifically, we will consider asymptotic approximations to a quantity f_n, depending on n, of the form

$$f_n = \sum_{j=1}^{r} g_j b_{jn} + R_{r,n}, \tag{1}$$

with attention paid to the size of the remainder term $R_{r,n}$ as $n \to \infty$. Usually the coefficients b_{jn} are quantities that decrease as n moves toward its limiting value, and decrease more quickly for larger j. Typical choices for b_{jn} are inverse powers of n or its square root.

Frequently the quantity to be approximated is a function of another variable. For instance, when f_n is a function of x, (1) becomes

$$f_n(x) = \sum_{j=1}^{r} g_j(x) b_{jn} + R_{r,n}(x); \tag{2}$$

the coefficients g_j and the error term are allowed to depend on x, and the factors b_{jn} and the error term are allowed to depend on n.

Methods with $r = 1$, using $g_1 b_{1n}$ to approximate f_n, are generally described by the term *first-order asymptotic*, and when additional terms are included one calls r defined in (1) the *order* of asymptotic approximation applied. Higher-order asymptotic methods, then, are defined here to be applications of (1) or (2) with $r > 1$; this includes second-order asymptotic methods.

We will consider innovative approaches to using terms in the series beyond the initial term $g_1 b_{1n}$ for assessing powers of tests and efficiencies of estimators, both asymptotically as n becomes large, and for finite n. Details may be omitted in discussing the applications of higher-order asymptotics, in order to avoid duplication with other *ESS* entries.

EFFICIENCY IN GENERAL

Much of the material in this section also appears in the entry Efficiency, Second-Order. Common definitions of the relative efficiency of two statistical procedures involve comparisons of sample sizes necessary to provide equivalent precision. Specifically, suppose that k_n observations are required to give the second procedure the same precision as is realized with n observations from the first procedure; then the relative efficiency of these two procedures is k_n/n, and the *asymptotic relative efficiency* is ARE $= \lim_{n\to\infty} k_n/n$. Placing this in the context of (1),

$$\frac{k_n}{n} = \text{ARE} + R_{1,n},$$
$$\text{where } \lim_{n\to\infty} R_{1,n} = 0. \tag{3}$$

Here equality of precision may refer to equality of mean square errors* of two estimates, or powers of two tests for a similar alternative; various precise definitions of equality of precision are discussed in the next section. The first procedure is preferable if ARE > 1, and the second procedure is preferable if ARE < 1.

Fisher [7] considered discrimination between estimation procedures when the asymptotic relative efficiency is unity, and explored situations in which a second-order version of (3) exists. Suppose that ARE $= 1$ and

$$\frac{k_n}{n} = 1 + \frac{d}{n} + R_{2,n},$$
$$\text{where } \lim_{n\to\infty} \frac{R_{2,n}}{n} = 0. \tag{4}$$

The first procedure is preferable if $d > 0$, and the second procedure is preferable if $d < 0$. The relation (4) implies that

$$d = \lim_{n\to\infty} (k_n - n).$$

Hodges and Lehmann [9] define the *deficiency* of the two procedures to be $k_n - n$; the asymptotic deficiency of the second procedure relative to the first is then d, when this limit exists. A simple example is presented in the next section.

Hodges and Lehmann [9] give an example involving estimation in which the limit d is infinite, and so an expansion as above need not always exist. Pfanzagl [14] and Ghosh [8] present regularity conditions under which such an expansion will exist when comparing the higher-order efficiencies of first-order efficient tests; see the review article by Skovgaard [19].

COMPARISONS OF ESTIMATORS

Consider assessing an asymptotically normal and asymptotically unbiased estimator of a parameter. One might take as a definition of the efficiency of this estimator the ratio of the Cramér—Rao* lower bound for its variance to the actual achieved variance. One might also define the estimator's asymptotic efficiency as the limit of this ratio as some parameter, usually reflecting the sample size, increases [2, pp. 137 f.; 4, pp. 304 ff.]. Estimates with efficiency closer to one are preferable to those with lower efficiency. Estimators whose asymptotic efficiency is unity are *first-order efficient*.

When one considers collections of estimators one often is interested in their behavior relative to one another. The relative efficiency for two estimators is the inverse of the ratio of their variances, and the asymptotic relative efficiency is the limit of this inverse ratio. When the two estimators have variances approximately proportional to n in large samples, this definition of asymptotic relative efficiency coincides with the definition in terms of relative sample sizes needed to give equivalent precision. This section will place the differentiation between estimators in the context of (1).

As a simple example, consider estimating a mean of a population with finite variance, using a sample of n independent observations. If one procedure estimates the mean as the sample mean, and the second procedure estimates the mean as the sample mean with the first observation ignored, then $k_n = n + 1$, the relative efficiency is $(n + 1)/n$, and the asymptotic relative efficiency is 1. The deficiency is then 1 for all values of n, and so the asymptotic deficiency is also 1.

Fisher [7] argues heuristically that maximum-likelihood estimation* produces estimators that are first-order efficient, with variances, to first order, given by the Fisher information* in the whole sample, and that the loss in efficiency incurred by other estimators might be measured by the correlation of these other estimators with the maximum-likelihood estimator, or alternately by the differences between the whole sample information and the information in the sampling distribution of the estimator. Other authors have made these claims rigorous, and some of these results will be reviewed below. Wong [23] presents a more thorough rigorous review. The correlation between estimators may be used to build a definition for second-order efficiency which is equivalent to the Fisher information difference, under certain regularity conditions [17,18].

Higher-order asymptotic expansions for the mean square error for the maximum-likelihood estimator can be generated. Since expansions for the mean square error are related to expansions for the information content of the maximum-likelihood estimator, and the information expansion is simpler, the information expansion will be considered first. Efron [6] uses methods from differential geometry to define the statistical curvature* γ_θ of an inference problem, and relates it to the loss of efficiency when inference procedures designed for local alternatives are applied globally.

Consider a family of distributions on a sample space χ parametrized by θ taking values in $\Theta \subset R$, and suppose that $X \in \chi^n$ is a vector of n independent and identically distributed variables X_j. Let $l = l(\theta; X)$ be the log likelihood for X. Let $\ddot{l}(\theta, X)$ be the second derivative of the log likelihood with respect to θ. Let $i_n(\theta) = -E[\ddot{l}(\theta, X); \theta]$ be the Fisher information in the sample X, let $i_n^{\hat{\theta}}(\theta)$ be the Fisher information for the sampling distribution of $\hat{\theta}$, the maximum-likelihood estimator for θ, and let $i_1(\theta)$ be the Fisher information for X_1. If γ_θ^1 is the curvature defined for the distribution of a random variable X_1, and γ_θ^n is the curvature calculated for the distribution of X, then $\gamma_\theta^1 = \gamma_\theta^n/\sqrt{n}$. One may show that $\lim_{n\to\infty}[i_n(\theta) - i_n^{\hat{\theta}}(\theta)] = i_1(\theta)(\gamma_\theta^1)^2$, and hence $i_n^{\hat{\theta}}(\theta)/n = i_1(\theta) - i_1(\theta)\gamma_\theta^1/n + R_{2,n}$, where $\lim_{n\to\infty} nR_{2,n} = 0$, giving an asymptotic expansion* for the average information contained in a maximum-likelihood estimator. Efron [6] also produced an asymptotic expansion for the variance of the maximum-likelihood estimator at a parameter value θ_0, which contains the statistical curvature and additional terms involving the curvature of the bias of the result of one scoring iteration, and the bias of the maximum-likelihood estimator at θ_0. These terms are all of size $O(n^{-2})$, which means that after dividing by n^{-2} they remain bounded, and the error is of size $o(n^{-2})$, which means that after dividing by n^{-2} the error converges to zero as $n \to \infty$. For more details *see* EFFICIENCY, SECOND-ORDER.

COMPARISONS OF TESTS

Asymptotic comparisons of powers* of families of tests having exactly or approximately the same significance level have been examined by many authors. Generally their investigations have considered the problem of testing a null hypothesis that a statistical parameter θ takes a null value, of the form $H_0 : \theta = \theta_0$, using two competing tests $T_{1,n}$ and $T_{2,n}$, indexed by a parameter n, generally indicating the sample size. Their critical values $t_{1,n}$ and $t_{2,n}$ satisfy

$$P[T_{i,n} \geqslant t_{i,n}; H_0] = \alpha \quad \text{for} \quad i = 1, 2. \qquad (5)$$

Measures of global asymptotic efficiency of two competing tests such as Bahadur efficiency* are generally functions of the parameter determining the distribution under an alternative, and since these functions generally do not coincide for competing tests, higher-order terms in sample size are generally not needed for choosing among tests of identical first-order efficiency.

Single-number measures of efficiency often times compare powers of tests whose sizes, exactly or approximately, are fixed and identical. For consistent tests and a fixed alternative hypothesis, distribution functions for the test statistics (or asymptotic approximations to these distribution functions) indicate an identical first-order asymptotic power of unity. Distinguishing between such tests, then, requires a local measure of relative efficiency such as Pitman efficiency*, which is the ratio of sample sizes necessary to give the same power against a local alternative. That is, alternatives of the form $H_A : \theta = \theta_0 + \epsilon/c_n$, where $c_n \to \infty$, are considered, and the limit $\lim_n(k_n/n)$ is desired, where

$$P[T_{1,n} \geqslant t_{1,n}; H_A] = P[T_{2,k_n} \geqslant t_{2,n}; H_A]. \qquad (6)$$

Often competing tests can be found whose measures of asymptotic relative efficiency against local alternatives is unity. One frequently wishes to discriminate between two such tests. Hodges and Lehmann [9] apply their concept of deficiency to the problem of comparing tests in cases in which sizes can be calculated exactly.

Often exact expressions for the probabilities in (5) and (6) are unavailable. In such cases the critical value, as well as the power usually must be approximated. Pfanzagl [14] notes that asymptotic comparisons of power are only interesting when significance levels of the tests agree to the same asymptotic order, and achieves this equality of size through a process of Studentization* in the presence of nuisance parameters*. Such equality of size might be obtained using a Cornish—Fisher* expansion to calculate the critical value for the test (*see* ASYMPTOTIC EXPANSIONS—II). Albers et al. [1] apply Edgeworth series to calculate the powers of nonparametric tests. The primary difficulty in such cases arises from the discrete nature of the distributions to be approximated. Pfaff and Pfanzagl [15] present applications of Edgeworth and saddlepoint* expansions to the problem of approximating power functions for test statistics with continuous distributions; they find that the Edgeworth series is more useful for analytic comparisons of power, while saddlepoint methods give more accurate numerical results. Applications to tests for sphericity*, and for the significance of a common cumulative logit in a model for ordered categorical data*, are presented by Chan and Srivastava [3] and Kolassa [10], respectively.

Asymptotic expansions of cumulative distribution functions can also be used to calculate asymptotic relative efficiencies and deficiencies. Those applications discussed here are under local alternatives. First-order approximations are generally sufficient to calculate asymptotic relative efficiencies; deficiency calculations generally require that second-order terms be included as well. Peers [13] uses approximations of the form (1) to approximate the cumulative distribution function of the likelihood-ratio statistics, Wald statistics*, and score statistics* as a mixture of noncentral χ^2 distributions. Taniguchi [21] calculates these terms to examine cases when the deficiency is zero and a third-order counterpart of (4) is required; comparisons may then be based on the final coefficient.

Many authors use higher-order asymptotic methods to aid in the construction of hypothesis tests; the most classical uses involve deriving the Bartlett correction to likelihood-ratio tests; *see also* BARTLETT ADJUSTMENT—I.

AN EXAMPLE

As an example, consider calculations of power in tests of equality of distribution for two populations classified into ordered categories. The Mann—Whitney—Wilcoxon statistic* is the score statistic for testing this null hypothesis against the alternative that the data are generated by the constant-cumulative-odds model of McCullagh [12]. Whitehead [22] proposes a method for approximating the power of the resulting test. Critical values are derived using a normal approximation with an approximate variance. Powers are calculated by approximating the alternative distribution as normal with the mean approximated by the parameter value times an approximation to the expected information. These first-order asymptotic methods will be compared with higher-order Edgeworth series methods.

One-sided tests of size $\alpha = 0.025$ for 2×4 contingency tables* with row totals both equal to 30 were examined. The column probabilities are those investigated by Whitehead [22], (0.289, 0.486, 0.153, 0.072). For a variety of model parameters the first-order, higher-order Edgeworth, and Monte Carlo* approximations to these powers are presented. The Edgeworth series used was the distribution function corresponding to the density (4) from the entry on asymptotic approximations*:

$$F_n(s) = \Phi(s) - \phi(s)\left[\frac{\lambda_3}{6}h_2(s) + \frac{\lambda_4}{24}h_3(s) + \frac{\lambda_3^2}{72}h_5(s)\right],$$

where the cumulants λ_k contain inverse powers of the row totals and are calculated by Kolassa [10], and the test statistic has been standardized to have zero mean and unit variance. All Monte Carlo approximations involve 1,000,000 samples, and so have standard errors no larger than 0.0005. Figure 1 contains the results. The Edgeworth approximations to the power are visually indistinguishable from the Monte Carlo approximations; the first-order approximations are not so close. Kolassa [10] shows that actual and nominal test sizes correspond more closely than the corresponding power values. Because of discreteness in the distribution of the test statistics, the first- and higher-order size calculations coincide in many cases; where they disagree, the nominal levels generating the Cornish—Fisher expansion appear to be substantially more accurate than the first-order approximations.

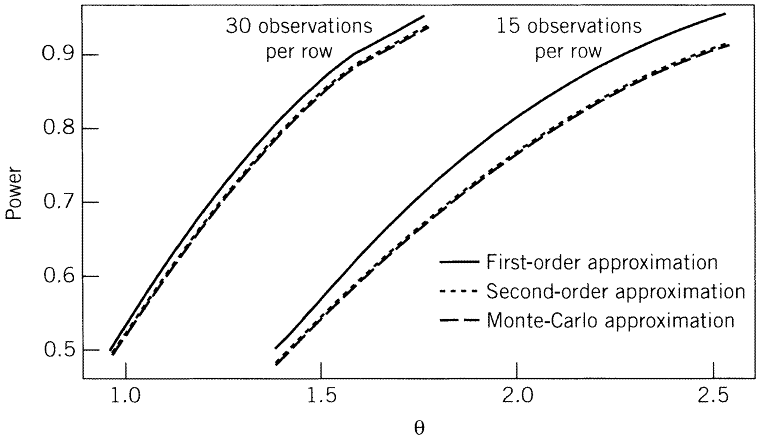


Figure 1. Comparisons of first- and second-order power approximations for ordered categorical data. (The second-order and Monte Carlo approximations are indistinguishable.)

OPEN QUESTIONS

Three topics among the open questions and ongoing difficulties in applying higher-order asymptotic methods will be reviewed here.

These are extending asymptotic methods to the more difficult situations involving statistics generated from nonindependent observations, simplifying regularity conditions, and simplifying the form of the resulting approximation.

Some results in the first area, particularly with references to ARMA processes, are presented by Swe and Taniguchi [20] and in the references they cite. The second area is perhaps not so important. Asymptotic methods, and saddlepoint methods in particular, are often remarkably accurate for a wide range of applications, and intuition often serves better than a careful evaluation of regularity conditions to indicate those contexts in which asymptotic methods are likely to be useful. Nevertheless, for certain applications careful examination of these conditions is sometimes required, and these conditions can often be quite complicated and not particularly intuitive [21].

Simplifying the form of asymptotic expansions is important for increasing their usefulness. Higher-order terms often are quite complicated [11,21]. Simplification might arise from choosing among various different asymptotic approximations accurate to the same order; see Daniels [5] and Pierce and Peters [16].

REFERENCES

1. Albers, W., Bickel, P. J., and van Zwet, W. R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem. *Ann. Statist.*, **4**, 108–156. (This paper presents Edgeworth series results for sums of discrete random variables, useful for implementing nonparametric tests.)
2. Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland, Calif. (A standard introductory text in theoretical statistics.)
3. Chan, Y. M. and Srivastava, M. S. (1988). Comparison of powers for the sphericity tests using both the asymptotic distribution and the bootstrap method, *Comm. Statist. Theory Methods*, **17**, 671–690.
4. Cox, D. R. and Hinkley, D. V. (1982). *Theoretical Statistics*. Chapman and Hall, London. (A standard text in theoretical statistics.)
5. Daniels, H. E. (1987). Tail probability approximations. *Internat. Statist. Rev.*, **55**, 37–46.
6. Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second-order efficiency). *Ann. Statist.*, **3**, 1189–1242. (This paper presents a geometric approach to asymptotics for models that can be embedded in curved exponential families.)
7. Fisher, R. A. (1925). Theory of statistical estimation, *Proc. Cambridge Phil. Soc.*, **22**, 700–725.
8. Ghosh, J. K. (1991). Higher order asymptotics for the likelihood ratio, Rao's and Wald's tests. *Statist. Probab. Lett.*, **12**, 505–509.
9. Hodges, J. L., Jr., and Lehmann, E. L. (1970). Deficiency. *Ann. Math. Statist.*, **41**, 783–801. (This paper is an important work on higher-order asymptotic efficiency.)
10. Kolassa, J. E. (1995). A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Statist. Med.*, **14**, 1577–1581.
11. Kolassa, J. E. (1996). Higher-order approximations to conditional distribution functions. *Ann. Statist.*, **24**, 353–364.
12. McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B*, **42**, 109–142.
13. Peers, H. W. (1971). Likelihood ratio and associated test criteria. *Biometrika*, **58**, 577–587. (This paper calculates approximations to statistical power at local alternatives in order to approximate efficiencies.)
14. Pfanzagl, J. (1980). Asymptotic expansions in parametric statistical theory. In *Developments in Statistics*, vol. 3, P. R. Krishnaiah, ed. Academic Press, New York, pp. 1–98. (This paper is an extended review article on the uses of asymptotics in statistical inference.)
15. Pfaff, T. and Pfanzagl, J. (1985). On the accuracy of asymptotic expansions for power functions. *J. Statist. Comput. Simul.*, **22**, 1–25.
16. Pierce, D. A. and Peters, D. (1992). Practical use of higher-order asymptotics for multiparameter exponential families. *J. Roy. Statist. Soc. Ser. B*, **54**, 701–737. (This paper reviews alternative forms of saddlepoint distribution function expansions.)

17. Rao, C. R. (1962). Efficient estimates and optimum inference procedures. *J. Roy, Statist. Soc. Ser. B*, **24**, 47–72.
18. Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā*, **25**, 189–206.
19. Skovgaard, I. M. (1989). A review of higher-order likelihood methods. In *Bull. Int. Statist. Inst. Proc. Forty-seventh Session*. vol. III, International Statistical Institute, Paris, pp. 331–351. (This paper is a recent review article on higher-order asymptotic methods.)
20. Swe, M. and Taniguchi, M. (1991). Higher-order asymptotic properties of a weighted estimator for Gaussian ARMA processes. *J. Time Ser. Anal.*, **12**, 83–93.
21. Taniguchi, M. (1991). Third-order asymptotic properties of a class of test statistics under a local alternative. *J. Multivariate Anal.*, **37**, 223–238.
22. Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statist. Med.*, **12**, 2257–2271.
23. Wong, W. H. (1992). On asymptotic efficiency in estimation theory. *Statist. Sinica*, **2**, 47–68.

See also Asymptotic Expansions—II; Bahadur Efficiency; Efficiency, Second-Order; Fisher Information; Likelihood Ratio Tests; Log-linear Models in Contingency Tables; Mann–Whitney–Wilcoxon Statistic; Maximum Likelihood Estimation; Pitman Efficiency; Score Statistics; Statistical Curvature; and Wald's *W*-Statistics.

John E. Kolassa

ATMOSPHERIC STATISTICS

The term atmospheric statistics covers a large body of work. In broad terms, this work can be divided into two categories: *statistical atmospheric statistics* and *atmospheric atmospheric statistics*.

Although the distinction is not always clearcut, in statistical atmospheric statistics, the application of statistics to problems in atmospheric science is, to a large extent, incidental. The hallmark of statistical atmospheric statistics is that, with little modification, the same analysis could be applied to data from an entirely different field. For example, the possibility of weather modification* attracted a lot of interest in the atmospheric science community during the 1970s, and a number of cloud-seeding experiments were planned and performed. As in other fields, statistical issues arose over the design of these experiments and over the analysis and interpretation of their results. This work is reviewed in [5] and [12]; *see also* Weather Modification—I and Weather Modification—II.

A second example of statistical atmospheric statistics is the development of stochastic rainfall models [28,29]; *see also* Rainfall, Landforms, and Streamflow. Here, the problem is to construct a model of variations in rainfall intensity over time and space that can be used in the design of reservoirs and storm-sewer or flood control systems (*see* Dam Theory). Other examples of statistical atmospheric statistics are described in [21] and Meteorology, Statistics in.

In recent years, atmospheric issues with global environmental implications have received considerable attention. Some statistical work on two of these issues—stratospheric ozone depletion and global warming—is reviewed below.

STATISTICS IN WEATHER PREDICTION

In contrast to statistical atmospheric statistics, atmospheric statistics is characterized by a close connection to atmospheric science. Compared to other scientific fields, atmospheric science is organized to an unusual degree around a single practical problem: the prediction of the future state of the atmosphere. Bjerknes [1] referred to this as the ultimate problem in meteorology. Since the time of Bjerknes, atmospheric prediction has been viewed as an initial-value problem. That is, to predict the future state of the atmosphere it is necessary, first, to describe the current state of the atmosphere, and second, to predict the future state by applying the laws of atmospheric dynamics to the current state. The historical development and modern view of this two-step process are described in ref. 8.

The first step in this two-step process is the estimation of the current state of the atmosphere, as represented by one or more atmospheric variables, at a dense grid of locations (called *analysis points*) from relatively sparse observations made at irregularly spaced locations (called *observation points*). This is called *objective analysis*; the most prominent role for statistics in atmospheric prediction lies in it. Early methods for objective analysis were based on fitting polynomial or other parametric functions to observations distributed in space by weighted least squares* [14,23]. The weights in this fitting are used to allow for unequal variances in the observations, which may be from quite different types of instruments (surface stations, weather balloons, satellites, etc.). Both local fitting and global fitting have been used.

As noted, the state of the atmosphere is described by more than one variable (temperature, pressure, humidity, etc.). These variables are coupled through governing equations. Methods have been developed to incorporate the constraints imposed by these governing equations in *function fitting*. A popular approach was devised by Flattery [11] for the objective analysis of the geopotential and the two horizontal components of wind. Under this approach, the three variables are expressed through an expansion in Hough functions. Hough functions are the eigenfunctions of a linearized version of the Laplace tidal equations, a relatively simple model of atmospheric dynamics on the sphere [6]. Because the model is linearized, the constraint—in this case, the geostrophic relation under which the Coriolis forces are balanced by the atmospheric pressure gradient—is only approximately met.

During the early 1970s, the use of function fitting for objective analysis declined, due in part to numerical problems, and interest shifted to a method called *statistical interpolation*, which was introduced into atmospheric science by Gandin [13]. It is essentially equivalent to linear interpolation methods developed by Kolmogorov [18] and Wiener [40] and popularized by Matheron [20] under the name *kriging*.*

To begin with, consider the univariate case. Let $Y(x)$ be the value of the variable of interest at location x, let x_0 be an analysis point, and let $x_i, i = 1, 2, \ldots, n$, be nearby observation points. Under statistical interpolation, the estimate of $Y(x_0)$ is given by

$$Y^*(x_0) = \mu(x_0) + \sum_{i=1}^{n} w_i[Y(x_i) - \mu(x_i)],$$

where $\mu(x)$ is the mean of $Y(x)$. This can be rewritten in obvious matrix notation as

$$Y^*(x_0) = \mu(x_0) + \mathbf{w}'(\mathbf{Y} - \boldsymbol{\mu}).$$

The vector w of interpolation weights is chosen to minimize the variance of the interpolation error $Y(x_0) - Y^*(x_0)$. This vector is given by $\mathbf{w} = \mathbf{C}^{-1}\mathbf{c}$, where

$$\mathbf{C} = [\mathrm{cov}(Y(x_i) - \mu(x_i), Y(x_j) - \mu(x_j))],$$
$$i,j = 1, 2, \ldots, n,$$
$$\mathbf{c} = [\mathrm{cov}(Y(x_0) - \mu(x_0), Y(x_i) - \mu(x_i))],$$
$$i = 1, 2, \ldots, n.$$

The presence of measurement error* can also be incorporated into this approach.

To implement this approach, it is necessary to specify the mean field $\mu(x)$ and the spatial covariance function of the deviation field. In atmospheric science, $\mu(x)$ is called the *background*. In early work, the background was taken to be the average of historical measurements, and was called the *climatological background* or simply *climatology*. Currently, an iterative approach is used under which predictions from a numerical model based on the previous objective analysis are used as background for the current objective analysis [8]. The shift from climatology to forecast for background is a reflection of improved forecast accuracy. Numerical models generate predictions only at the analysis points and not at the observation points. Background at the observation points is interpolated from background at the analysis points. This is called *forward interpolation*.

In the simplest case, the spatial covariance of the deviation field is assumed to be separable into horizontal and vertical components; both of these components are assumed to be stationary, and the horizontal component is assumed to be isotropic. Under these assumptions, the covariance depends only on the distance between points. The deviations at the observation locations can be used to estimate spatial covariance at a number of distances and a parametric covariance function is fitted to these estimates. In some cases, the form of these parametric models is based on physical models [24]. Extensions of these methods have been made to accommodate anisotropic and nonstationary covariances. Statistical interpolation can be extended in a straightforward way to the multivariate case. Physical constraints (like geostrophic balance) can be imposed on the interpolated field, either exactly or approximately, by imposing constraints on the covariance structure of the multivariate field [19].

Research on objective analysis remains active. One area of current interest concerns the incorporation of asynoptic satellite data into objective analysis. This is called *continuous data assimilation*—the original idea was due to Charney et al. [7]. It involves what might be called *sequential objective analysis*, in which data are incorporated into an essentially continuous forward model integration. Another focus of current research involves the estimation of forecast errors in data-poor regions, such as over the oceans. Methods based on the Kalman filter* have shown promise here [9,38]. An area of research that may be of interest to statisticians concerns the use of splines* in objective analysis [19].

STATISTICS IN STRATOSPHERIC OZONE DEPLETION

Turning now to statistical atmospheric statistics, two problems of current interest in atmospheric science are the depletion of stratospheric ozone by chlorine-containing compounds such as chlorofluorocarbons (CFCs) and atmospheric warming due to increased atmospheric concentrations of carbon dioxide and other radiatively active gases. Statistical work on these issues, both of which have substantial environmental and economic implications, is reviewed in [32].

The potential depletion of stratospheric ozone, which shields the Earth's surface from ultraviolet radiation, is a serious concern. Some scientific aspects of ozone depletion are reviewed in [37]. While there is little doubt that significant depletion has occurred over Antarctica, this is related to special atmospheric conditions (low temperatures and small atmospheric mixing), and questions remain about the rate of depletion at low latitudes. Total column ozone has been measured at 36 ground stations around the world since the 1960s. These stations are commonly grouped into seven regions. Because the measurements were made by a Dobson spectrophotometer, they are referred to as the *Dobson data*.

Statistical analyses of the Dobson data include refs. 4, 25, 26, and 36. The goal of these analyses was to estimate a global trend in stratospheric ozone concentration. One complicating factor in these analyses is that ozone concentration may be affected by such factors as solar variability, atmospheric nuclear testing, and volcanic eruptions, so that selection of variables becomes an issue. A second complicating factor concerns the form of the ozone depletion curve. Reinsel and Tiao [26] used a two-phase linear model, with slope 0 prior to 1970 and unknown slope ω thereafter. In contrast, Bloomfield et al. [4] used a depletion curve of the form ωm_t, where m_t was the depletion curve predicted by a photochemical model. For both models, $\omega = 0$ corresponds to the case of no depletion. For the specification adopted in [4], $\omega = 1$ corresponds to the case in which the predictions of the photochemical model are correct.

Reinsel and Tiao estimated depletion curves in each of the seven regions separately and combined these estimates through a random-effects model to estimate a global depletion rate. Bloomfield et al. used a more explicitly spatial model to estimate a global depletion rate. In neither case was significant global depletion found. However, using indirect measurements, Reinsel and Tiao identified significant ozone depletion of up to 4% per decade in the upper atmosphere where photochemical theory suggests that depletion should be greatest.

Since the late 1970s, the Total Ozone Mapping Spectrometer (TOMS) aboard the Nimbus 7 satellite has provided daily total column ozone data on a global 1° latitude by 1.25° longitude grid [17]. These data have recently been analyzed by Niu and Tiao [22]. The spatial coverage of these observations is much denser than of the ground stations, so more careful modeling of spatial correlation is needed. For a fixed latitude, let $Y_j(t)$ be the average ozone observation in month t at longitude j. Niu and Tiao adopted the additive model

$$Y_j(t) = s_j(t) + r_j(t) + \xi_j(t),$$

where $s_j(t)$ is a seasonal component and $r_j(t)$ is a linear trend component. The noise term $\xi_j(t)$ is assumed to follow a space—time autoregressive (STAR) model of the form

$$\begin{aligned}\xi_j(t) = \sum \alpha_k \xi_{j-k}(t) + \theta_k \xi_{j+k}(t) \\ + \sum \phi_l \xi_j(t-l) + \epsilon_j(t),\end{aligned}$$

where the $\epsilon_j(t)$ are independent, normal errors with mean 0 and variance that depends on t only through the month. Under this model, the error in period t at longitude j is related both to previous errors at the same longitude and to errors in the same period at different longitudes. The first summation in this expression runs from $k = 1$ to $k = q$, and the second from $l = 1$ to $l = p$. The parameters q and p define the spatial and temporal order of the STAR model, respectively. An interesting feature of this model is that longitude is a circular variable.

Niu and Tiao fitted this model to seven latitude bands in each hemisphere for the period November 1978 to May 1990. For the most part, a symmetric STAR (2, 1) error model was selected. This model includes two spatial lags and one temporal lag. The model is symmetric in the sense that $\alpha_k = \theta_k$. The results indicate significant reductions of up to several percent per decade at high latitudes. One possible extension of this analysis is to allow for latitudinal dependence in the error term $\xi_j(t)$.

STATISTICS IN GLOBAL WARMING

Concern over global warming stems from the observation that the atmospheric concentration of carbon dioxide and other radiatively active gases has increased significantly since the Industrial Revolution, due primarily to the consumption of fossil fuels and other human activities. These gases are responsible for the greenhouse effect that keeps the atmosphere warm enough to support life. An increase in their concentration effectively increases radiative forcing of the atmosphere. The direct effect of this increase in radiative forcing is atmospheric warming. However, the rate and magnitude of this warming is determined by certain feedbacks in the climate system, and there is considerable uncertainty about the ultimate atmospheric response. Scientific issues relating to global warming are reviewed in the reports of the Intergovernmental Panel on Climate Change [15,16].

Estimates of mean global surface temperature based on direct, instrumental measurements are available for the past 120 years. These estimates show an irregular warming trend amounting to around 0.5°C over the length of the record. Some statistical work in the area of global warming has involved efforts to assess the significance of this apparent trend. Let T_t be the observed increase in mean global surface temperature in year t over some baseline year. The basic model is

$$T_t = \beta_0 + \beta_1 t + \epsilon_t,$$

where ϵ_t represents variations in temperature around the linear trend. Interest centers on estimating β_1 and on testing the null hypothesis that $\beta_1 = 0$. A central problem in this work concerns the specification of the properties of the variability around the trend. In particular, failure to allow for positive serial dependence* in this variability will exaggerate the significance of the estimate of β_1. This issue was considered in generality in ref. 3. Woodward and Gray [42] focused on the case where ϵ_t follows an autoregressive moving-average (ARMA)* process, while Smith [30] considered a model with long-range dependence*. The results indicate that the estimate of β_1 (which is on the order of 0.004°C per year) is at least marginally significant. The one exception considered by Woodward and Gray is the case in which the process generating ϵ_t is a correlated random walk* (i.e., the ARMA model has a unit root*). However, this model is implausible on scientific grounds. In related work, Richards [27] used econometric methods to test a number of hypotheses about changes in global temperature.

This work on global warming uses only the data and exploits little if any scientific understanding of the climate system. Other statistical work in this area has focused on estimation and inference about parameters that are physically more meaningful than the slope of a linear trend. The sensitivity of the global surface temperature to changes in radiative forcing is summarized in a parameter called *temperature sensitivity*, commonly denoted by ΔT; it is defined as the equilibrium response of mean global surface temperature to an increase in radiative forcing of 4.4 W m^{-2} (corresponding to a doubling of the atmospheric concentration of carbon dioxide). A central problem in climatology is the estimation of ΔT.

It is possible to estimate the secular trend in the atmospheric concentrations of the principal greenhouse gases since 1860. These estimated trends can be converted into an estimated trend in overall radiative forcing over the same period. Radiative forcing has increased by around 2.2 W m^{-2} since 1860. The expected response of mean global surface temperature to this pattern of increase can be estimated using a simple climate model that includes ΔT as a parameter. As noted, annual estimates of the mean global surface temperature over this period show an overall warming of around 0.5°C.

The temperature sensitivity can be estimated by matching the model response to the observed warming. The basic model is

$$T_t = m_t(\Delta T) + \epsilon_t,$$

where $m_t(\Delta T)$ is the model response in year t to observed changes in radiative forcing for temperature sensitivity ΔT, and, as before, ϵ_t represents other temperature variations. It is again important in fitting this model—and crucial in constructing a confidence interval for ΔT—to recognize that the process ϵ_t is serially dependent. This process includes temperature responses to fluctuations in radiative forcing that are not included in the estimated secular trend. The effect of these fluctuations (which are due to events like volcanic eruptions and to other variations in solar activity, cloudiness, etc.) on temperature can persist for several years. This general approach was first applied in ref. 41. In its most careful application, Bloomfield [2] adopted a fractionally integrated white-noise* model for ϵ_t. The corresponding point estimate of ΔT was around 1.4°C with an approximate 0.95 confidence interval of 0.7°C to 2.2°C.

The general approach to estimating ΔT outlined above depends on the accuracy of the estimated secular trend* in radiative forcing. For example, the radiative-forcing effects of sulfate aerosols were not included in Bloomfield's analysis. Sulfate aerosols, which are produced by volcanic eruptions and by burning coal, tend to have a cooling effect on the surface. This effect is due to the reflectivity of the aerosols and also to their role in providing condensation nuclei for clouds. Unfortunately, it is difficult to incorporate these effects in the kind of analysis undertaken by Bloomfield. Historical data on sulfate aerosols are incomplete. Unlike greenhouse gases, the cooling effect of sulfate aerosols tends to be localized, so that historical data on their regional distribution are needed. Finally, the radiative effects of sulfate aerosols are complicated, depending, for example, on whether the aerosol is sooty or not.

Motivated in part by the difficulties of incorporating sulfate aerosols into Bloomfield's approach, Solow and Patwardhan [35] developed an alternative approach to estimating ΔT that does not require knowledge of the secular trend in radiative forcing. This approach is based on the result that the statistical characteristics of the temperature response to short-term fluctuations in radiative forcing also depend on ΔT. Specifically, if ΔT is high, this response is large and persistent, while if ΔT is low, it is relatively small and transient. Briefly, Solow and Patwardhan estimated ΔT (via likelihood in the frequency domain) by fitting the spectrum of observed temperature variations around a smooth trend to a spectrum generated by model simulations with a fixed value of ΔT. In generating these model simulations, it was necessary to adopt a statistical model of short-term variations in radiative forcing. Following the literature, Solow and Patwardhan assumed that these variations followed a white-noise process with variance 1.0 W m^{-2}. This model is consistent with a short record of satellite measurements of the global radiation budget. The point estimate of ΔT found in this way was 1.4°C with an approximate 0.95 confidence interval of 0.9°C to 2.3°C. These results are virtually identical to those found by Bloomfield, suggesting that the cooling effect of sulfate aerosols has been minimal.

FUTURE DIRECTIONS

Unlike weather prediction, atmospheric environmental science has no real central problem. For this reason, directions in statistical work in this area are difficult to predict. Taking a somewhat personal view, one general area of future research concerns the use of observations to validate physical models that produce values of a suite of variables in two or three dimensions through time [31]. There is a conceptual difficulty here, because no physical model is truly correct and it is therefore only a matter of acquiring enough data to discover this. Apart from that, the problem is complicated by the need to describe the pattern of covariance of multivariate spatial—time series. On the subject-matter side, there is growing interest in understanding and predicting climate variability on the annual to decadal time scale. The most important component of climate variability on this time scale is the El Niño Southern Oscillation [10]. The best known manifestation of ENSO are so-called El Niño events, associated with regional changes in precipitation and other climate variables. An historical record of the timing and magnitude of these events was analyzed in refs. 33 and 34. However, this analysis just scratches the surface, and the area seems ripe for further statistical work.

REFERENCES

1. Bjerknes, V. (1911). *Dynamic Meteorology and Hydrography. Part II. Kinematics*. Carnegie Institute, New York.
2. Bloomfield, P. (1992). Trends in global temperature. *Climate Change*, **21**, 1–16.
3. Bloomfield, P. and Nychka, D. (1992). Climate spectra and detecting climate change. *Climatic Change*, **21**, 275–287.
4. Bloomfield, P., Oehlert, G., Thompson, M. L., and Zeger, S. (1984). A frequency domain analysis of trends in Dobson total ozone records. *J. Geophys. Res.*, **88**, 8512–8522.
5. Braham, J. J. (1979). Field experimentation in weather modification (with discussion). *J. Amer. Statist. Ass.*, **74**, 57–104.
6. Chapman, S. and Lindzen, R. (1970). *Atmospheric Tides*. Reidel, Hingham, Mass.
7. Charney, J., Halem, M., and Jastrow, R. (1969). Use of incomplete historical data to infer the present state of the atmosphere. *J. Atmos. Sci.*, **26**, 1160–1163.
8. Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, UK.
9. Daley, R. (1992). Estimating model-error covariances for application to atmospheric data assimilation. *Monthly Weather Rev.*, **120**, 1735–1750.
10. Diaz, H. F. and Markgraf, V. (1992). *El Niño*. Cambridge University Press, Cambridge, UK.
11. Flattery, T. (1971). *Spectral models for global analysis and forecasting*. Proc. Sixth Air Weather Service Exchange Conf. Tech. Rep. 242, Air Weather Service, pp. 42–54.
12. Gabriel, K. R. (1979). Some issues in weather experimentation. *Commun. Statist. A*, **8**, 975–1015.
13. Gandin, L. (1965). *Objective Analysis of Meteorological Fields*. Israel Program for Scientific Translation, Jerusalem.
14. Gilchrist, B. and Cressman, G. (1954). An experiment in objective analysis. *Tellus*, **6**, 309–318.
15. IPCC (1990). *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press, Cambridge, UK.
16. IPCC (1996). *Climate Change 1995*. Cambridge University Press, Cambridge, UK.
17. JGR (1984). Nimbus 7 scientific results. *J. Geophys. Res.*, **89**, 4967–5382.
18. Kolmogorov, A. (1941). Interpolated and extrapolated stationary random sequences. *Izv. Akad. Nauk SSSR Ser. Mat.*, **5**, 85–95.
19. Lorenc, A. (1981). A global three-dimensional multivariate statistical interpolation scheme. *Monthly Weather Rev.*, **109**, 701–721.
20. Matheron, G. (1971). *The Theory of Regionalized Variables and Its Application*. Ecole des Mines, Paris.
21. Neyman, J., Scott, E. L., and Wells, M. A. (1969). Statistics in meteorology. *Rev. Int. Statist. Inst.*, **37**, 119–148.

22. Niu, X. and Tiao, G. C. (1995). Modelling satellite ozone data. *J. Amer. Statist. Ass.*, **90**, 969–983.
23. Panofsky, H. (1949). Objective weather-map analysis. *J. Appl. Meteorol.*, **6**, 386–392.
24. Philips, N. (1986). The spatial statistics of random geostrophic modes and first-guess errors. *Tellus*, **A38**, 314–322.
25. Reinsel, G., Tiao, G. C., Wang, M. N., Lewis, R., and Nychka, D. (1981). Statistical analysis of stratospheric ozone data for detection of trend. *Atmos. Environment*, **15**, 1569–1577.
26. Reinsel, G. and Tiao, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone. *J. Amer. Statist. Ass.*, **82**, 20–30.
27. Richards, G. R. (1993). Change in global temperature: a statistical analysis. *J. Climate*, **6**, 546–558.
28. Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. (1987). Some models for rainfall based on stochastic point processes. *Proc. R. Soc. London A*, **410**, 269–288.
29. Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. (1989). A point-process model for rainfall: further developments. *Proc. R. Soc. London A*, **417**, 283–298.
30. Smith, R. L. (1993). Longrange dependence and global warming. In *Statistics for the Environment*, V. Barnett and K. Turkman, eds. Wiley, Chichester.
31. Solow, A. R. (1991). On the statistical comparison of climate model output and climate data. In *Greenhouse-Gas-Induced Climatic Change*, M. Schlesinger, ed. Elsevier, Amsterdam.
32. Solow, A. R. (1994). Statistical methods in atmospheric science. In *Handbook of Statistics 12: Environmental Statistics*, G. P. Patil and C. R. Rao, eds. North-Holland, Amsterdam, pp. 717–734.
33. Solow, A. R. (1995). An exploratory analysis of a record of El Niño events, 1800–1987. *J. Amer. Statist. Ass.*, **90**, 72–79.
34. Solow, A. R. (1995). Testing for change in the frequency of El Niño events. *J. Climate*, **18**, 2563–2566.
35. Solow, A. R. and Patwardhan, A. (1994). Some model-based inference about global warming. *Environmetrics*, **5**, 273–279.
36. St. John, D., Bailey, W. H., Fellner, W. H., Minor, J. M., and Snee, R. D. (1982). Time series analysis of stratospheric ozone. *Commun. Statist. Theory and Methods*, **11**, 1293–1333.
37. Stolarski, R. S. (1982). Fluorocarbons and stratospheric ozone: a review of current knowledge. *Amer. Statist.*, **36**, 303–311.
38. Todling, R. and Cohn, S. E. (1994). Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Monthly Weather Rev.*, **122**, 2530–2545.
39. Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Rev.*, **108**, 1122–1143.
40. Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York.
41. Wigley, T. M. L. and Raper, S. C. B. (1990). Natural variability of the climate system and detection of the greenhouse effect. *Nature*, **344**, 324–327.
42. Woodward, W. A. and Gray, H. L. (1993). Global warming and the problem of testing for trend in time series data. *J. Climate*, **6**, 953–962.

See also DAM THEORY; KRIGING; METEOROLOGY, STATISTICS IN; WEATHER FORECASTING, BRIER SCORE; WEATHER FORECASTING, EPSTEIN SCORING RULE IN; WEATHER MODIFICATION—I; and WEATHER MODIFICATION—II.

ANDREW SOLOW

ATTRIBUTE

In statistical usage this term often has the connotation "nonquantitative." It is applied to characteristics that are not easily expressed in numerical terms: e.g., temperament, taste, and species. The term *qualitative character* is used synonymously.

The *theory of attributes* (see Chapters 4 to 6 of Yule and Kendall [2]) is mainly concerned with analysis of contingency tables and categorical data*.

Attribute sampling, i.e., sampling "from a population whose members exhibit either an attribute, A, or its complement, *not* $- A$", includes repeated trials from a Bernoulli distribution [1, (Sec. 9.28)], leading to consideration of probabilities in binomial distributions (*see* BINOMIAL AND MULTINOMIAL PARAMETERS, INFERENCE ON and [1 (Secs. 9.28–9.33)]). However, attribute sampling may lead to consideration of probabilities in hypergeometric distributions* (when finite populations are involved), in negative binomial distributions* (for certain sequential sampling models) [1, (Sec. 9.37)] and in multinomial distributions*.

REFERENCES

1. Stuart, A. and Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics*, Vol. 1: Distribution Theory (5th ed.). Oxford University Press, New York.
2. Yule, G. U. and Kendall, M. G. (1950). *Introduction to the Theory of Statistics*, 14th ed. Hafner, New York, Charles Griffin, London. (The early editions of this work, by G. U. Yule alone, contain a considerably longer discussion of attributes.)

AUDITING, STATISTICS IN

Auditing is a specialized area of accounting involving examination by independent auditors of the operation of an organization's system of internal control and of the financial statements in order to express an opinion on the financial statements prepared by the management of the organization. Internal auditors, who are members of the organization, also review the operation of the system of internal control to see if it is operating effectively.

Uses of statistical methods in auditing are a relatively recent phenomenon and are still evolving. Three major current uses are sampling of transactions to study the operation of the internal control system, sampling of accounts to study the correctness of recorded account balances, and regression analysis for analytical review.

STUDY OF INTERNAL CONTROL SYSTEM

Transactions, such as payments of bills received, are sampled to make inferences about the effectiveness of the internal control system. Usually with this type of sampling, auditors are concerned with a qualitative characteristic, namely, whether the internal control was operating improperly for the transaction. Inferences are made about the process proportion of transactions that are handled improperly by the internal control system. Typically, auditors desire assurance that the process proportion is reasonably small. Random samples of transactions are utilized and evaluated by standard statistical procedures based on the binomial* distribution or the Poisson* approximation. No special statistical problems are encountered in this application. Consequently, little research is currently being done in this area. A good summary is contained in Roberts [8].

STUDY OF ACCOUNT BALANCES

Accounts, such as accounts receivable and inventory, often consist of thousands of individual line items that may have a total value of millions of dollars. For such large accounts, it is not economical for the auditor to audit every line item. Consequently, auditors frequently select a sample of line items and audit these, on the basis of which inferences are made about the total error amount in the population. Thus, the characteristic of interest here is quantitative. Denoting the total amount recorded by the company for the account by Y and the total amount that the auditor would establish as correct if each line item in the account were audited by X, the total error amount in the account is $E = Y - X$.

Since the auditor usually knows the amounts recorded by the company for each line item in the population (called the *book amounts*), this information can be utilized in estimating the total error amount E. One estimator that incorporates this supplementary information is the difference estimator, which for simple random sampling of line items is

$$\hat{E} = \frac{N}{n} \sum_{i=1}^{n} d_i = N\overline{d},$$

where

$$d_i = y_i - x_i$$

is the difference between the book and audit amounts for the ith sample line item, $\overline{d}$ is the mean difference per sample line item, and N is the number of line items in the population.

Other estimators that incorporate the information on book amounts are the ratio and regression estimators. Each of these can be used with simple or stratified random sampling of line items. Still another means of incorporating the information about book amounts is through sampling with probability proportional to book amount and then utilizing an unbiased estimator.

Each of these procedures has some serious limitations in auditing applications. When the sample contains no errors, the estimated standard error of the estimator equals zero, as can be seen from the estimated variance of the difference estimator with simple random sampling of line items:

$$s^2(\hat{E}) = N^2 \frac{N-n}{Nn(n-1)} \sum_{i=1}^{n} (d_i - \overline{d})^2.$$

An estimated standard error of zero suggests perfect precision, which would be an unwarranted conclusion.

A second limitation of these supplementary information procedures is that a number of simulation* studies have found that large-sample confidence limits based on the normal distribution are frequently not applicable for sample sizes used in auditing. Neter and Loebbecke [5] studied four actual accounting populations and from these constructed a number of study populations with varying error rates. They utilized sample sizes of 100 and 200 with unstratified, stratified, and PPS sampling*. They found for the supplementary information procedures that the coverages for the large-sample upper confidence bounds for the total error amount (i.e., the proportion of times the bound is correct in repeated sampling) were frequently substantially below the nominal confidence level.

The search for alternative inference procedures that do not depend on large-sample theory has frequently involved monetary unit sampling, which involves the selection of individual monetary units from the population. Since the auditor cannot audit a single monetary unit but only the line item to which the unit pertains, any error found is then prorated to the monetary units belonging to the line item. The prorated errors are called *taints* in auditing, and are denoted by $t_k = d_k/y_k$, where $d_k \neq 0$. An important case in auditing is when the errors in the population are all overstatement errors and the taints are restricted to positive values not exceeding 1.0. For this case, a conservative confidence bound for the total error amount can be constructed by assuming that all overstatement taints are at the maximum value of 1.0. The problem then becomes one of obtaining an upper confidence bound for the population proportion of monetary units that contain an error. Multiplying this bound by Y, the total number of monetary book units in the population, yields a conservative upper confidence bound for the total error amount E.

A difficulty with this conservative bound is that it is usually not tight enough. Stringer [9] developed a heuristic for reducing this bound when all observed taints are not at the maximum value of 1.0:

$$Yp_u(1-\alpha; n, m) - Y\sum_{k=1}^{m}[p_u(1-\alpha; n, k) \\ -p_u(1-\alpha; n, k-1)](1-t_k),$$

where m is the observed number of errors in the sample of n monetary units, $t_1 \geqslant t_2 \geqslant \cdots \geqslant t_m$ are the observed taints, and $p_u(1-\alpha; n, m)$ is the $1-\alpha$ upper bound for a binomial proportion when m errors are observed in a sample of n.

Simulation studies (e.g., Reneau [7]) have shown that this Stringer bound has coverages always exceeding the nominal level and often close to 100%; also that the Stringer bound is not very tight and may involve substantial risks of making incorrect decisions when the bound is used for testing purposes. Leslie et al. [3] developed another heuristic bound, called a *cell bound*, that tends to be tighter than the Stringer bound and has good coverage characteristics for many accounting populations. The cell bound is based on cell sampling, where the frame of monetary units is divided into strata or cells of equal numbers of monetary units and one monetary unit is selected at random from each cell, the cell selections being made independently.

Fienberg et al. [2] developed a bound based on the multinomial distribution* by viewing monetary unit sampling in discretized form. The multinomial classes represent the different possible taints rounded to a specified degree. The procedure involves obtaining a joint confidence region for the multinomial parameters and then maximizing a linear function of the parameters, representing the total error amount in the population, over the confidence region. Because of the complexities of computation, Fienberg et al. utilized only a partial ordering of the sample outcomes for developing the joint confidence region. However, simulation studies (e.g., Plante et al. [6]) have shown that coverages for the multinomial bound are near to or above the nominal level for a variety of populations, particularly if cell sampling is employed.

Several approaches based on Bayesian* methodology have been proposed when monetary unit sampling is employed. Cox and Snell [1] proposed a Bayesian bound for which they assume that the population error rate for monetary units and the population mean taint for monetary units in error are independent parameters. These and other assumptions lead to a posterior distribution* of the total error amount that is a simple multiple of the F distribution*. Hence Bayesian bounds are very easy to obtain by this method. Neter and Godfrey [4] studied the behavior of the Cox and Snell bound in repeated sampling from a given population for sample size 100, and found that conservative prior parameter values exist so that the Cox and Snell bound has coverages near or above the Bayesian probability level for a wide range of populations. However, research in progress suggests that robustness may be a function of sample size, and not include all possible populations. Another recent proposal by Tsui et al. [11] is to combine the multinomial sampling model with a Dirichlet* prior distribution* to obtain Bayesian bounds for the total error amount.

ANALYTICAL REVIEW

Analytical review procedures are utilized by auditors to make internal and external consistency comparisons, such as comparing the current financial information with comparable information for preceding periods, or comparing operating data for the firm with data for the industry. Ratios have frequently been used in making these comparisons, such as comparing the ratio of cost of sales to sales for the current period with corresponding ratios for prior periods. Use of ratios for making comparisons assumes that the relationship between the two variables is linear through the origin. Often, this relationship may not be of this form. Also, use of ratio analysis does not permit the study of the simultaneous relationship of one variable with several others, such as when it is desired to examine the relationship of the revenue of a public utility to kilowatt hours, rate per kilowatt hour, and seasonal effects.

Regression analysis is now being used by some auditors for certain analytical review procedures. The data often are time series*, as when current periods are to be compared to prior periods. Sometimes, the data are cross-sectional, as when data for several hundred retail outlets of a firm are studied to identify ones that are outliers*. No special problems in applying regression analysis for analytical review have been encountered. The regression models employed have tended to be relatively simple ones. In using regression models, auditors are particularly concerned about identifying outliers that are worthy of further investigation. Much of the research to data has been concerned with developing rules that relate the identification of outliers with the amount of subsequent audit work that should be performed. Stringer and Stewart [10] have provided a comprehensive summary of the use of regression methods in auditing for analytical review.

REFERENCES

1. Cox, D. R. and Snell, E. J. (1979). *Biometrika*, **66**, 125–132.
2. Fienberg, S., Neter, J., and Leitch, R. A. (1977). *J. Amer. Statist. Ass.*, **72**, 295–302.
3. Leslie, D. A., Teitlebaum, A. D., and Anderson, R. J. (1979). *Dollar-Unit Sampling*. Copp, Clark, Pitman, Toronto, Canada. (This book provides a comprehensive description of the Stringer and cell bounds and their uses in auditing.)
4. Neter, J. and Godfrey, J. (1985). *J. R. Statist. Soc. C, (Appl. Statist.)*, **34**, 157–168.

5. Neter, J. and Loebbecke, J. K. (1975). *Behavior of Major Statistical Estimators in Sampling Accounting Populations*. American Institute of Certified Public Accountants, New York.
6. Plante, R., Neter, J., and Leitch, R. A. (1985). *Auditing*, **5**, 40–56.
7. Reneau, J. H. (1978). *Accounting Rev.*, **53**, 669–680.
8. Roberts, D. M. (1978). *Statistical Auditing*. American Institute of Certified Public Accountants, New York.
9. Stringer, K. W. (1963). *Proc. Bus. Econ. Statist. Sec., 1963*. American Statistical Association, pp. 405–411.
10. Stringer, K. W. and Stewart, T. R. (1985). *Statistical Techniques for Analytical Review in Auditing*. Deloitte, Haskins, and Sells, New York.
11. Tsui, K. W., Matsumura, E. M., and Tsui, K. L. (1985). *Accounting Rev*. **60**, 76–96.

See also CONTROL CHARTS; FINANCE, STATISTICS IN; and QUALITY CONTROL, STATISTICAL.

JOHN NETER

AUSTRALIAN AND NEW ZEALAND JOURNAL OF STATISTICS

[This entry has been updated by the Editors.]

The *Australian Journal of Statistics (AJS)* was founded by the (then) Statistical Society of New South Wales in 1959. Until 1997 it appeared three times a year, the three issues constituting a volume. The founding editor of the *AJS* was H. O. Lancaster, who served from 1959–1971. He was followed by C. R. Heathcote (1971–1973), C. C. Heyde (1973–1978), E. J. Williams (1978–1983), J. S. Maritz (1983–1989), C. A. McGilchrist (1989–1991), I. R. James (1991–1997) and S. J. Sheather (1997).

In 1998 *AJS* merged with *The New Zealand Statistician* as Volume 40 of *The Australian and New Zealand Journal of Statistics* (*see* also NEW ZEALAND STATISTICAL ASSOCIATION); the website is www.statsoc.org.au/Publications/ANZJS.htm, and the publisher is Blackwell.

AUSTRALIAN JOURNAL OF STATISTICS: HISTORY

The editorial policy of *AJS* aimed to achieve a balance between theoretical and applied articles in the following areas: (1) mathematical statistics, econometrics, and probability theory; (2) new applications of established statistical methods; (3) applications of newly developed methods; (4) case histories of interesting practical applications; (5) studies of concepts (particularly in economic and social fields) defined in terms suitable for statistical measurement; (6) sources and applications of Australian statistical data; and (7) matters of general interest, such as surveys of the applications of statistics in broad fields. No ranking is implied in this list. *AJS* also published critical book reviews and short book notices. A news and notes section regularly appeared until 1977, but this function was taken over by the *Statistical Society of Australia Newsletter*, which first appeared in May 1977.

An international perspective and coverage was intended for *AJS* and contributions from outside Australia were always welcomed; *see* STATISTICAL SOCIETY OF AUSTRALIA for further discussion. All papers were refereed.

At the time of establishment of *AJS*, the Statistical Society of New South Wales, based in Sydney and founded in 1947, was the only society of its kind in Australia. It assumed the responsibility for starting the journal, which, as its name implies, was intended to serve the statistical profession in Australia. The then president of the Statistical Society of New South Wales, P. C. Wickens, wrote in a foreword to the first issue of the journal: “It is hoped . . . that it will not be long before statistical societies will be firmly established in other States, and when this occurs it will undoubtedly be necessary to reconsider the management of the Journal.”

The hopes expressed in this statement were not long in coming to fruition. The Statistical Society of Canberra was formed in 1961. In October 1962 the Statistical Society of Australia was formed by the amalgamation of the New South Wales and Canberra Societies, which then became branches of the main society. At this stage the journal became the responsibility of the new society, but its editor and editorial policy remained unchanged. Further branches of the society were later formed in Victoria (1964), Western Australia (1964), and South Australia (1967). In 1976, responsibility for *AJS* was assumed by the Australian Statistical Publishing Association Inc., whose membership is coterminous with membership of the Central Council of the Statistical Society of Australia.

THE MERGER

Following the merger of *AJS* and *The New Zealand Statistician* in 1998, the combined *Australian and New Zealand Journal of Statistics (ANZJS)* has been published in four issues per year, under two Theory and Methods Editors (one of whom also serves as Managing Editor), an Applications Editors from New Zealand, a Book Review Editor, a Technical Editor, and an international Editorial Board (10 or so for Applications and 35 or so for Theory and Methods).

ANZJS publishes articles in four categories:

Applications: Papers demonstrate the application of statistical methods to problems faced by users of statistics. A particular focus is the application of newly developed statistical methodology to real data and the demonstration of better use of established statistical methodology in an area of application.
Theory & Methods: Papers make a substantive and original contribution to the theory and methodology of statistics, econometrics or probability. A special focus is given to papers motivated by, and illustrated with, real data.
Reviews: Papers give an overview of a current area of statistical research which consolidate and reconcile existing knowledge and make judgments about the most promising directions for future work.
Historical and General Interest: Papers discuss the history of statistics in Australia and New Zealand, the role of statistical organizations in private and government institutions and the analysis of datasets of general interest.

See also NEW ZEALAND STATISTICAL ASSOCIATION and STATISTICAL SOCIETY OF AUSTRALIA.

C. C. HEYDE
The Editors

AUSTRALIAN JOURNAL OF STATISTICS. See *AUSTRALIAN AND NEW ZEALAND JOURNAL OF STATISTICS*

AUTOCORRELATION FUNCTION. See AUTOREGRESSIVE–INTEGRATED MOVING AVERAGE (ARIMA) MODELS

AUTOMATIC INTERACTION DETECTION (AID) TECHNIQUES

Automatic interaction detection (AID) is a technique for analyzing a large quantity of data (whose number of cases is typically hundreds or thousands) by subdividing it into disjoint exhaustive subsets so as best to “explain” a dependent variable on the basis of a given set of categorized predictors. Although first noted under a different name in 1959 [1], the current computer versions were introduced by a series of papers dating from 1963, starting with one by J. N. Morgan and J. A. Sonquist [8]. Since its inception AID has grown in popularity and found application (and misapplication) in many applied fields. It can be used as an end analysis in itself, or as a prior screening device to sift the variables and draw attention to certain interactions for specific inclusion in subsequent analysis by other methods.

All AID techniques (normally implemented as computer programs) operate in a stepwise manner, first subdividing the total data base according to one of the predictors* (chosen by the algorithm to maximize some given criterion), then reexamining separately and subdividing each of the groups formed by the initial subdivision, and continuing in a like manner on each new subgroup formed until some stopping criterion is reached. Some versions have a “look-ahead” feature which allows possible subdivisions of subdivisions to be examined before the primary subdivision is effected—a procedure that may be theoretically desirable but is currently not cost-effective or usually worthwhile for the majority of databases met in practice. This results in a tree-like structure called a dendrogram* in AID literature.

The "automatic" of AID refers to the computer program making all the decisions as to which predictor to use when, and how. This attribute can be countermanded by the user in some versions of AID. The "interaction" in AID refers to its design property that each subgroup formed in the analysis is treated *individually*. Thus the technique does not constrain itself to producing "additive" or "symmetric" models, although these may result anyway.

DEPENDENT VARIABLE

The various types of AID are distinguished by the nature of the dependent variable. Standard AID [10] operates on one that is an ordinal scalar, while the multivariate extension to an ordinal vector is handled by MAID [3]. A nominal scalar-dependent variable can be analyzed by "chi-squared* AID" CHAID [6] or "theta AID" THAID [7]. The theta statistic of the latter technique is the total proportion of observations that belong to a modal category in an AID subdivision. Thus, theta is bounded below by d^{-1} for a d-category dependent variable, and bounded above by unity.

The criterion for a "good" subdivision that "explains" the data depends on the nature of the dependent variable. Table 1 lists the possible types of dependent variable together with the appropriate classical-type criterion for each one. The criteria are all interrelated in that the suitable special case of each type (e.g., ordinal vector of dimension 1, a dichotomous nominal variable) reduces to one of the others in the list. Other criteria have been proposed (e.g., THAID above), but as they are dissimilar to the classical criteria (even asymptotically), they are in need of theoretical backing.

Table 1. Classical-Type Criteria for Various Dependent Variables

Type of Dependent Variable	Criterion
Ordinal scalar	t-test*, ANOVA*
Ordinal scalar and covariate	F-test*, ANOCOVA*
Nominal	Chi-square*
Ordinal vector	Hotelling's T^{2*}, MANOVA*
Ordinal vector and covariates	MANOCOVAR*

Superficially, AID bears resemblance to stepwise regression* both in intent and procedure. A comparison of AID with this and other statistical techniques is given in ref. 10.

PREDICTORS

The categorized predictors are used to determine the possible subdivisions of the data at any one stage, and hence result in the subdivisions being meaningful in that a particular subgroup can be labeled according to which categories of which predictors define the subgroup. Anything from 2 to 100 predictors are typically used in a single analysis (depending also on the size of the data base), each predictor having from 2 up to about 10 categories (depending on type), although predictors with more categories have appeared in the literature.

The categories of each predictor are grouped by AID to define the subdivisions (in many versions this reduction is *always* to two subdivisions—i.e., a binary split), the allowable grouping determined by the type of predictor. These types, which carry names peculiar to AID, include monotonic (ordered categories in which only adjacent categories may be grouped), floating (monotonic plus an additional category that may be grouped with any of the others—usually used for missing information or "don't-knows"), free (nominal categories with no restriction on grouping), interleaved (a number of monotonic categories used for combining different ordered scales; special cases include the aforementioned types), and cyclic (monotonic with "wraparound," useful for U^*- or inverted U-distributions*). Modern theory has concentrated on the behavior of AID on monotonic and free predictors, where not unsurprisingly the Bonferroni inequality* applied to a dichotomous predictor produces reasonable conservative significance tests of subdivision differences for the multicategory case. The theory of piecewise-regression and clustering are applicable to monotonic and free predictors, respectively, but unfortunately their asymptotic theories, which dwell on the increase in the number of categories (in our terminology), are not relevant, as the practical application of AID involves a small number of categories.

CRITICISM

Many criticisms have been leveled at AID-type techniques, the two major ones being that (1) AID produces too many subdivisions or idiosyncratic results (based on experiments with random data in which *no* subdivisions should be produced), and (2) the interpretation of AID results are often fallacious. The first criticism is based on earlier versions of AID that contained no valid hypothesis test as to whether the subdivisions produced were statistically significant. Further, the older versions of AID made no distinction between the types of the predictors involved when choosing one on which to base the subdivisions—thus introducing a bias in favor of using multicategory free predictors. Appropriate significance testing for standard AID has been developed in ref. 5. The latest versions of AID introduce testing as part and parcel of the technique, preferring to subdivide on the most "significant" rather than most "explanatory" predictor and thus remove both the aforementioned bias and the generation of nonsignificant splits. Ideally, a stringent significance level should be used to take into account the number of predictors present in the analysis, using e.g., Boole's (Waring) inequality*.

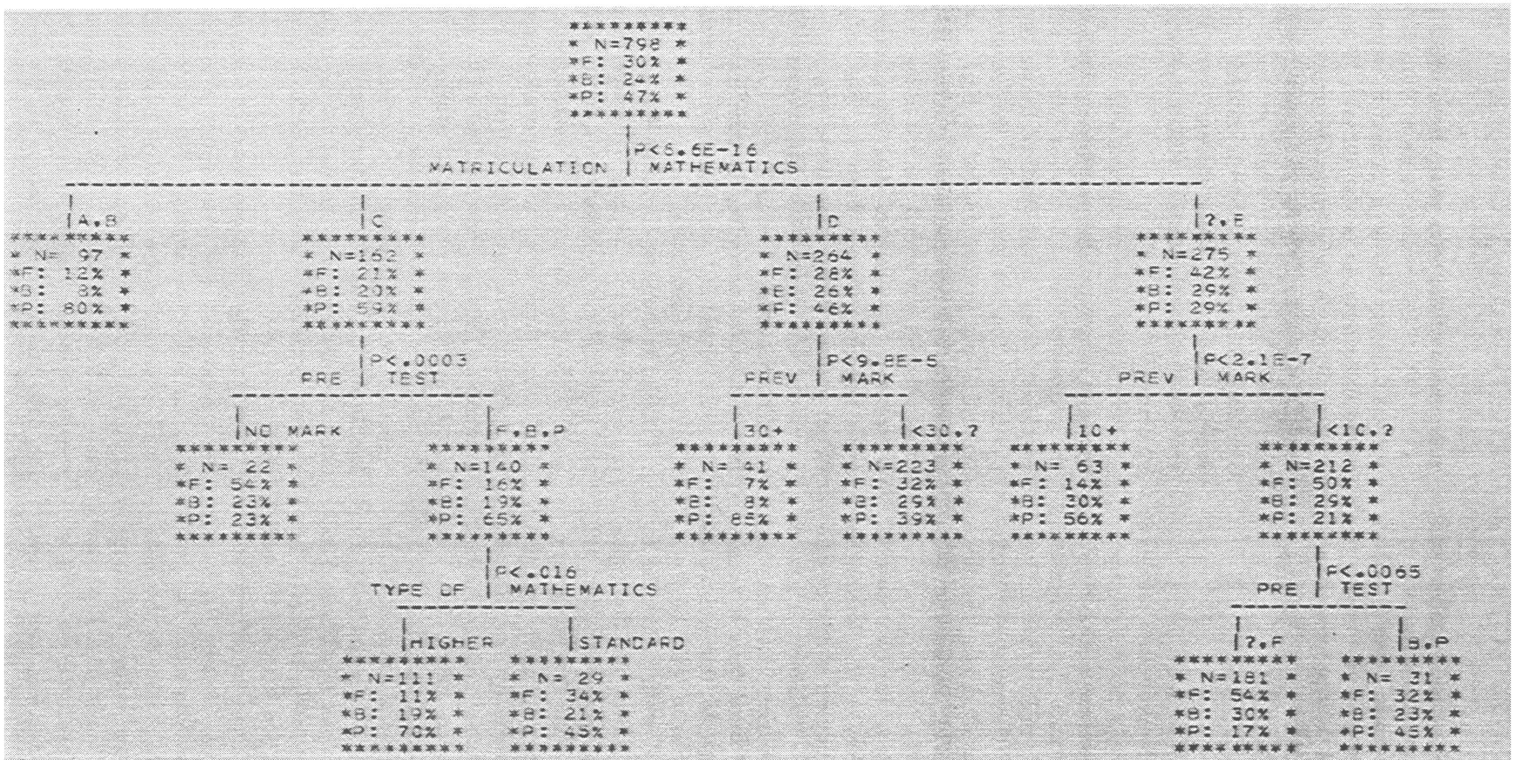
The second criticism is not so much an indictment of AID but rather of the ignorance displayed by some of its users. Added to this is the inadequate manner in which many current users present their AID results. Examples are completely omitting mention of sample sizes in the various subdivisions, the statistical significance* of these subdivisions, or appropriate auxiliary information that would allow a reader to repair this lack, at least in his or her own mind. The dendrogram resulting from an AID analysis is so appealing that a tendency has arisen for users to ignore the possible existence of competing predictors that would be revealed by an examination of the AID statistics produced for each predictor both before and after a subdivision. Certain apparent "interactions" quoted in the literature (see, e.g., the criticism [2] of the study [4] could well be spurious for this reason, and in certain circumstances, small changes in the data could cause different predictors to be selected by AID, with consequent different conclusions reached by a naive user. This fault is probably magnified by the "automatic" nature of AID mentioned above).

Finally, among other possible limitations, the appropriateness of an AID analysis can be no better than the appropriateness of the splitting criterion as given in Table 1, and all the assumptions inherent in the theory behind these criteria and their properties naturally carry through to AID.

EXAMPLE

Figure 1 depicts the dendrogram* resulting from a CHAID analysis of 798 first-year commerce students enrolled at the University of the Witwatersrand. The dependent variable is the student's midyear mark (June 1979) in the course Mathematics and Statistics IB (M & S1B), classified into the three groups: passed (>49%), borderline (40–49%), and failed (<40%).

Table 2 gives details of the first stage of the analysis and indicates the predictive power and optimal grouping of the categories of each predictor. Predictors 2 and 5 (whether a local matriculant, and sex) are immediately discarded since there is no significant difference between the groups, as indicated by a significance level of $p = 1$. Predictors 3, 6, 7, and 8, while possessing an optimal grouping as shown on the right-hand side of the table, are nevertheless not considered significant. Note that predictor 6, which has nine categories optimally grouped into two groups, is ostensibly "significant" since the 2×3 contingency table so formed has $p = 0.0014$; however, taking into account the optimization that went into forming this table, a conservative estimate of the "significance" using a Bonferroni inequality* is $p = 0.35$—clearly



F = FAIL
B = BORDERLINE
P = PASS

THE SIGNIFICANCE LEVELS QUOTED ARE CONSERVATIVE IN THAT THEY TAKE INTO ACCOUNT THE TYPE OF THE PREDICTOR USED AND THE NUMBER OF ITS CATEGORIES. BUT DOES NOT REFLECT THE POSSIBLE NUMBER OF PREDICTORS AVAILABLE AT THAT STAGE.

Figure 1. Dendrogram of CHAID analysis of first-year commerce students.

Table 2. First Stage of CHAID Analysis

Predictor	Number of Categories	Type	Ostensible Significance Level	Conservative Significance Level	Optimum Number of Groups	Optimum Grouping[a]
1. Matriculation mathematics (A, high mark; E, low)	6	Floating	3.0×10^{-17}	6.6×10^{-16}	4	AB C D ?E
2. Local matriculant? (Y, yes; N, no)	2	Monotonic	1.0	1.0	1	YN
3. Matriculation English (A, high mark; E, low)	6	Floating	0.018	0.40	3	?A B CDE
4. Year of matriculation (4, up to 1974; 5, 1975; 8, 1978)	6	Floating	4.1×10^{-5}	9.1×10^{-4}	3	8 7 ?654
5. Sex (M, male: F, female)	3	Floating	1.0	1.0	1	?MF
6. Course and rules registered under (G, general; BCom. full time; etc.	9	Free	0.0014	0.35	2	G/7C9 LOPT
7. Year of study (1, first; . . .; 5, fifth or more)	6	Floating	0.048	0.43	2	1 23?45
8. Type of matriculation mathematics (H, higher; S, standard)	3	Floating	0.036	0.11	2	?H S
9. Previous M & S1B mark (?, not a repeat student; 1, first; etc.)	9	Floating	1.1×10^{-9}	1.4×10^{-8}	2	23456 ?78
10. Pre-test (F, failed; B, borderline; P, passed)	4	Floating	8.7×10^{-11}	4.3×10^{-10}	2	?F BP
11. First midterm (April) test (F, failed; B, borderline; P, passed)	4	Floating	6.1×10^{-50}	3.0×10^{-49}	3	?F B P

[a] The symbol "?" refers to missing information and is the "floating" category.

not significant. Predictor 11 (midterm test mark) is clearly the best with conservative $p \leqslant 3.0 \times 10^{-49}$, but was included in the analysis for information only, and precluded from forming the basis of a subdivision of the data since the purpose of the analysis is to predict on the basis of information available before the commencement of the academic year. Predictor 1 is the best usable predictor ($p \leqslant 6.6 \times 10^{-16}$) and divides the data into four groups.

Figure 1 displays in detail the four-way subdivision of the total group from which it is clear that the pass rate declines from 80% for the students with high (A and B) matriculation mathematics marks, through 59 and 46% for those with intermediate marks (C and D) down to 29% for the lower marks (E). Those students for whom no mark is available (14 such students coded "?", mainly foreign or older students) are interestingly enough grouped with the poorest students.

The analysis then continued with each of the four subgroups. Information on each of the predictors similar to that in Table 2 was produced, from which further details are available. The students with mathematics symbol C are further divided in Fig. 1 according to their pre-test, where it is seen that those who did not attend the test (it was not compulsory, and implies that the students were skipping classes even at this early stage) perform worse than those who did—no matter what their mark! The groups with mathematics marks D and E were each further subdivided according to their mark in the same course last year (clearly only available for repeat students).

Finally, in the lowest level in Fig. 1, there are two further subdivisions. Although they are technically significant ($p \leqslant 0.016$ and $p \leqslant 0.0065$) they should be considered marginal, since these levels make no allowance for the number of predictors (effectively 10) examined. (They do, however, take into account the type and number of categories in the predictor used to define the subdivision.) Nevertheless, considering the pretest, it is comforting to note the similar poor performance of students who did not attend it, on the two occasions where this predictor was used.

This is merely a brief summary of some of the information and conclusions available from an AID-type analysis. The secondary details concerning the predictive power and optimum grouping of the categories within each predictor for each of the subdivisions provide valuable insight to the structure of the data and interrelationships of the predictors.

STATE OF THE ART AND FUTURE DEVELOPMENTS

The underlying theory behind valid hypothesis testing in AID is still embryonic. At present only standard AID has provision for a covariate; the other versions have yet to be so extended. Computer installations with powerful interactive terminals or personalized computer systems do not need the "automatic" decision making of AID. Instead, they could offer the researcher the opportunity to introduce additional background information and take various decisions dynamically along the lines of ref. 9. Such a feature is still to be implemented in AID.

REFERENCES

1. Belson, W. (1959). *Appl. Statist.*, **8**, 65–75.
2. Doyle, P. (1973). *Operat. Res. Quart.*, **24**, 465–467.
3. Gillo, M. W. and Shelly, M. W. (1974). *J. Amer. Statist. Ass.*, **69**, 646–653. (MAID.)
4. Heald, G. I. (1972). *Operat. Res. Quart.*, **23**, 445–457.
5. Kass, G. V. (1975). *Appl. Statist.*, **24**, 178–189. (Theory and additional references to AID.)
6. Kass, G. V. (1980). *Appl. Statist.*, **29**, 119–127. (CHAID.)
7. Morgan, J. N. and Messenger, R. C. (1973). *THAID—a sequential analysis program for the analysis of nominal scale dependent variables*. ISR, University of Michigan, Ann Arbor, Mich. (THAID.)
8. Morgan, J. N. and Sonquist, J. A. (1963). *J. Amer. Statist. Ass.*, **58**, 415–434.
9. Press, L. I., Rogers, M. S. and Shure, G. H. (1969). *Behav. Sci.*, **14**, 364–370.
10. Sonquist, J. A., Baker, E. L. and Morgan, J. N. (1971). *Search for Structure (Alias-AID-III)*. ISR, University of Michigan, Ann

Arbor, Mich. (Mainly a user manual, but contains background material, examples, and further references.)

See also CLASSIFICATION—I; COMPUTERS AND STATISTICS; PREDICTIVE ANALYSIS; and STEPWISE REGRESSION.

G. V. KASS

AUTOREGRESSIVE ERROR, HILDRETH–LU SCANNING METHOD

Consider the linear regression model

$$y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_{p-1} X_{t,p-1} + \epsilon_t, \qquad t = 1, \ldots, N, \tag{1}$$

where y_t is the tth observation on the dependent variable, X_{tj} is the tth observation on the jth nonstochastic independent variable, and ϵ_t is the tth observation on the error term. This can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where y is $(N \times 1)$ in dimension, $\mathbf{X}$ is $(N \times p)$, $\boldsymbol{\beta}$ is $(p \times 1)$, and $\boldsymbol{\epsilon}$ is $(N \times 1)$. The usual assumptions on the error vector $\boldsymbol{\epsilon}$ are that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$. In this case, the ordinary least-squares* estimator of β, denoted by $\hat{\boldsymbol{\beta}}_0$, is given by $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (*see* GENERAL LINEAR MODEL).

The assumption that the error terms are uncorrelated often breaks down in time-series* studies and sometimes in cross-sectional studies, in which case we state that the error terms are autocorrelated or serially correlated. We denote this by writing $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Sigma}$. Mixed autoregressive-moving average* processes are used to describe this serial correlation (*see* AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS). Specifically,

$$\begin{aligned}\epsilon_t &= \phi_1\epsilon_{t-1} + \cdots + \phi_p\epsilon_{t-p} + a_t \\ &\quad -\theta_1 a_{t-1} - \cdots - \theta_q a_{t-q},\end{aligned} \tag{2}$$

where $E(a_t) = 0, V(a_t) = \sigma_a^2$, and the a_t's are uncorrelated. To ensure stationarity, we require that the roots of $1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0$ lie outside the unit circle. The case that has been considered most frequently in the econometrics* literature is when the error terms are AR (1). That is,

$$\epsilon_t = \phi_1\epsilon_{t-1} + a_t. \tag{3}$$

In (3), it has become customary to replace ϕ_1 by ρ, where we require $|\rho| < 1$ for a stationary process*. We will focus our attention on the autocorrelation structure specified in (3).

For an AR (1) process, it is shown in Box and Jenkins [2] that $\sigma_\epsilon^2 = \sigma_a^2/(1-\rho^2)$ and $E(\epsilon_t\epsilon_{t-k}) = \rho^k$. Thus the covariance matrix $\boldsymbol{\Sigma}$ associated with $(\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$ is

$$\begin{aligned}\boldsymbol{\Sigma} = \sigma_\epsilon^2\mathbf{A} = \sigma_\epsilon^2 &\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{bmatrix} \\ &= \sigma_a^2\mathbf{B}.\end{aligned} \tag{4}$$

When ρ *is known*, the generalized least-squares estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_G$, is found by minimizing $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{B}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Since $\mathbf{B}$ is positive definite, there is a nonsingular matrix $\mathbf{H}$ such that $\mathbf{B} = (\mathbf{H}'\mathbf{H})^{-1}$ and $\mathbf{B}^{-1} = \mathbf{H}'\mathbf{H}$. Thus minimizing $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{B}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is equivalent to minimizing $(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})$ via ordinary least-squares, where $\mathbf{y}^* = \mathbf{H}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{H}\mathbf{X}$. It follows that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_G &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^* \\ &= (\mathbf{X}'\mathbf{B}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}^{-1}\mathbf{y}.\end{aligned} \tag{5}$$

For the AR(1) error structure, the transformation $\mathbf{H}$ that permits ordinary least-squares estimation is

$$\mathbf{H} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}. \tag{6}$$

In (5), one could have used $\mathbf{A}^{-1}$ or $\boldsymbol{\Sigma}^{-1}$ in place of $\mathbf{B}^{-1}$ since the scalars cancel out.

When ρ *is not known*, Judge et al. [7] point out that three procedures are available for parameter estimation: estimated generalized least-squares, nonlinear least-squares, and maximum likelihood*.

Let $\hat{\boldsymbol{\beta}}_E$ denote the estimated generalized least-squares estimator of $\boldsymbol{\beta}$. $\hat{\boldsymbol{\beta}}_E$ is obtained by using the estimator in (5) after estimating

ρ. Thus these procedures are called two-step procedures. Several methods are available for estimating ρ. These include:

1. The Cochrane–Orcutt procedure [3], where $\hat{\rho}_1 = \sum_{t=2}^{N} \hat{\epsilon}_t \hat{\epsilon}_{t-1} / \sum_{t=1}^{N} \hat{\epsilon}_t^2$ and the $\hat{\epsilon}_t$'s are obtained by using ordinary least-squares on $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. More precisely, this is termed the Prais–Winsten [10] procedure, since all N elements of $\mathbf{y}$ and all N rows of $\mathbf{X}$ were affected by the $\mathbf{H}$ transformation in obtaining $\mathbf{y}^*$ and $\mathbf{X}^*$. Theil [11] proposed a modification of $\hat{\rho}_1$: namely, $(N-p)/(N-1)\hat{\rho}_1$.
2. The estimate of ρ obtained from the Durbin–Watson* statistic, D. Specifically, $\hat{\rho}_2 = 1 - D/2$. Theil and Nagar [12] give the following modification of $\hat{\rho}_2 : (N^2\hat{\rho}_2 + p^2)/(N^2 - p^2)$.
3. The estimate of ρ obtained from the Durbin procedure [4]. Let $\mathbf{H}_0$ denote the $(N-1) \times N$ matrix obtained by deleting the first row of $\mathbf{H}$ in (6). Let $\hat{\rho}_3$ denote the estimated coefficient of y_{t-1} in the model: $\mathbf{H}_0\mathbf{y} = \mathbf{H}_0\mathbf{X}\boldsymbol{\beta} + \mathbf{H}_0\boldsymbol{\epsilon}$. For the simple linear regression* model, we have

$$\begin{aligned} y_t &= \beta_0(1-\rho) + \rho y_{t-1} \\ &\quad + \beta_1 x_t - \beta_1 \rho x_{t-1} + a_t, \quad t = 2, \ldots, N. \end{aligned}$$

By using this method, Maddala [8] points out that one is ignoring the constraint that (coefficient of x_{t-1}) $= -$(coefficient of x_t). (coefficient of y_{t-1}).

In the nonlinear least-squares procedure, one needs to find those estimates of $\boldsymbol{\beta}$ and ρ that simultaneously minimize $(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})'$ $(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})$. Although nonlinear optimization algorithms can be used, Hildreth and Lu [6] suggested a search procedure. For values of ρ from -1.0 to 1.0 in increments of 0.1, calculate $\hat{\boldsymbol{\beta}}_G$ as stipulated in (5) and the corresponding sum of squares, $(\mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}_G)'(\mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}_G)$. Choose that value of ρ which minimizes this sum of squares. Higher decimal accuracy can be obtained by finding the sum of squares for several additional values of ρ near the minimizing value. Although the Hildreth–Lu method is not computationally efficient, the minimum sum of squares obtained should be global rather than local if some care is exercised in the search procedure. Obviously, the value of ρ so obtained need not equal any of the values used in the estimated generalized least-squares procedure.

Under the assumption that the a_t's are normally distributed, the maximum-likelihood procedure can be used. Judge et al. [7] show that maximizing the concentrated likelihood function is equivalent to minimizing $(1-\rho^2)^{-1/N}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})$; this differs from the nonlinear least-squares procedure by the $(1-\rho^2)^{-1/N}$ factor. Algorithms for maximizing the concentrated likelihood function are presented in Hildreth and Dent [5] and in Beach and MacKinnon [1], although a search procedure similar to the Hildreth–Lu method could be utilized.

Empirical results for some of the procedures discussed above are presented in Maddala [8]. For annual data from 1935 to 1954, and 10 different firms, Maddala regresses gross investment on two independent variables: value of the firm, and stock of plant and equipment. The results are presented in Table 1. Inspection of the entries in Table 1 reveals that the maximum-likelihood and Hildreth–Lu estimates are always in the same neighborhood, with Durbin's estimates differing substantially from these two.

Pindyck and Rubinfeld [9] also present two numerical examples using the Hildreth–Lu scanning method.

Although the name Hildreth–Lu has been reserved to refer to the search procedure for first-order autoregressive error, a similar search procedure could be employed for any ARMA error structure, as discussed in Judge et al. [7].

REFERENCES

1. Beach, C. M. and MacKinnon, J. G. (1978). *Econometrica*, **46**, 51–58.
2. Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.
3. Cochrane, D. and Orcutt, G. H. (1949). *J. Amer. Statist. Ass.*, **44**, 32–61.
4. Durbin, J. (1960). *J.R. Statist. Soc. B*, **22**, 139–153.
5. Hildreth C. and Dent W. (1974). In *Econometrics and Economic Theory: Essays in Honor of*

Table 1. Estimates of First-Order Autocorrelation Coefficient by Different Methods

	Method			
Firm	Cochrane–Orcutt	Durbin	Hildreth–Lu	Maximum Likelihood
GM	0.458	0.816	0.67	0.64
U.S. Steel	0.481	0.874	0.74	0.69
GE	0.461	1.061	0.50	0.47
Chrysler	−0.020	−0.346	−0.05	−0.04
Atlantic-Richfield	−0.236	−0.737	−0.22	−0.21
IBM	0.114	0.624	0.18	0.17
Union Oil	0.098	0.125	0.12	0.11
Westinghouse	0.241	0.297	0.30	0.28
Goodyear	0.246	0.706	0.39	0.36
Diamond Match	0.402	0.385	0.65	0.57

Jan Tinbergen, W. Sellekaert, ed. Macmillan, London, pp. 3–25.

6. Hildreth C. and Lu, J. Y. (1960). *Demand Relations with Autocorrelated Disturbances. Mich. State Univ. Agric. Exp. Stn. Tech. Bull. 276*, East Lansing, Mich.
7. Judge, G. G., Griffiths, W. E., Hill, R. C., and Lee, T. -C. (1980). *The Theory and Practice of Econometrics*. Wiley, New York.
8. Maddala, G. S. (1977). *Econometrics*. McGraw-Hill, New York.
9. Pindyck, R. S. and Rubinfeld, D. L. (1976). *Econometric Models and Economic Forecasts*. McGraw-Hill, New York.
10. Prais, S. J. and Winsten, C. B. (1954). *Trend Estimators and Serial Correlation. Cowles Comm. Discuss. Paper No. 383*, Chicago.
11. Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.
12. Theil, H. and Nagar, A. L. (1961). *J. Amer. Statist. Ass.*, **56**, 793–806.

See also Autoregressive–Moving Average (ARMA) Models; Least Squares; and Serial Correlation.

Frank B. Alt

AUTOREGRESSIVE–INTEGRATED MOVING AVERAGE (ARIMA) MODELS

An important class of models for describing a single time series*z_t is the class of autoregressive–moving average models* referred to as ARMA(p, q)* models.

$$(z_t - \mu) = \phi_1(z_{t-1} - \mu) + \cdots + \phi_p(z_t - \mu) + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \quad (1)$$

where the notation in (1) implies that (a) z_t is the original time series, or some suitable nonlinear transformation of it (such as a logarithm or a square root); (b) z_t is a stationary* time series with a fixed mean μ; (c) a_t is a random residual series, which can also be interpreted as the series of one-step-ahead forecast errors; and (d) $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \mu$ are parameters to be estimated from the data. Alternatively, autoregressive–moving average (ARMA) models* may be written in terms of the backward-shift operator* B, such that $B^j z_t = z_{t-j}$, $B^j a_t = a_{t-j}$, as follows:

$$(z_t - \mu) = \frac{1 - \theta_1 B - \ldots - \theta_q B^q}{1 - \phi_1 B - \ldots - \phi_p B^p} a_t \quad (2)$$

$$= \frac{\theta(B)}{\phi(B)} a_t \quad (3)$$

Thus the ARMA(p, q) model represents the time series, or a suitable nonlinear transformation, as the output from a linear filter whose input is a random series and whose *transfer function** is a rational function of the backward-shift operator B.

The model (1) is not of immediate practical use because very few real-world time series are stationary time series in statistical equilibrium about a fixed mean μ. Instead, they are characterized by random changes in their level, slope, etc., and by the presence of seasonal patterns which also evolve with time. Traditional methods of handling such diverse behavior involve the decomposition of the time series into a "trend"*, a "seasonal component"*, and a "residual component." After removal of the trend and seasonal component, it is customary to describe the residual component by means of a stationary ARMA(p,q) model of the form (1). Such an approach suffers from the following disadvantages: (a) it is arbitrary as to what is called a trend and a seasonal component; (b) removal of the trend and seasonal component introduces additional autocorrelation into the residual component; and (c) the assumptions normally made about the behavior of the trend and seasonal component are unrealistic.

To overcome these difficulties a new class of models, called autoregressive–integrated moving average models, referred to as ARIMA models, has been developed to describe, *under the umbrella of one model*, trends, seasonality, and residual random behavior [1]. Moreover, such models for describing nonstationary* time series contain flexible structures which allow the trend and seasonal component to be nondeterministic, i.e., their statistical properties evolve in time. In addition, iterative methods have been developed [1] for identifying (or specifying), estimating* (or fitting*) and checking (or criticizing) such models given the data.

NONSEASONAL ARIMA MODELS

Consider the first-order autoregressive model

$$(z_t - \mu) = \phi(z_{t-1} - \mu) + a_t \tag{4}$$

or, in backward-shift-operator notation $(B^j)z_t = z_{t-j}$,

$$(1 - \phi B)(z_t - \mu) = a_t,$$

which is stationary if $|\phi| < 1$. The solution of the difference equation* (4) may be written as the sum of the complementary function, i.e., the solution of $(1 - \phi B)(z_t - \mu) = 0$, and a particular integral, i.e., any function that satisfies (4). Relative to a time origin $t = 0$, the solution of (4) thus becomes

$$(z_t - \mu) = \phi^t(z_0 - \mu) + \sum_{j=1}^{t} \phi^{j-1} a_j. \tag{5}$$

If $|\phi| > 1$, the first term in (5) dominates and the growth of the series is explosive. Although such explosive nonstationarity occurs in some situations (such as bacterial growth), for most practical situations it is convenient to work with a less severe form of nonstationarity. This can be achieved by setting $\phi = 1$ in (4) and (5), which then become

$$(1 - B)z_t = z_t - z_{t-1} = a_t, \tag{6}$$

$$z_t = z_0 + \sum_{j=1}^{t} a_j, \tag{7}$$

i.e., z_t is a random walk* model. More generally, we consider a nonstationary ARMA(p,q) model

$$\phi'(B)(z_t - \mu) = \theta(B)a_t, \tag{8}$$

where $\phi'(B)$ is a nonstationary autoregressive operator. To prevent explosive nonstationarity, we impose the restriction that d of the factors of $\phi'(B)$ are unity, i.e., $\phi'(B) = \phi(B)(1 - B)^d$. Model (8) then becomes

$$\phi(B)(1 - B)^d z_t = \theta(B)a_t \tag{9}$$

where $\phi(B)$ is a stationary autoregressive* operator and $\theta(B)$ is an invertible moving average* operator, as in a stationary autoregressive–moving average model. Since $(1 - B)z_t = z_t - z_{t-1} = \nabla z_t$, where ∇ is the backward difference operator, the model (9) can also be written

$$\phi(B)\nabla^d z_t = \theta(B)a_t \qquad (\nabla^0 = 1). \tag{9}$$

Model (9) implies that whereas z_t is a nonstationary series, its dth difference $w_t = \nabla^d z_t$ is stationary and can be described by an autoregressive-moving average model. The model (9) is called an *autoregressive-integrated moving average model* or ARIMA(p,d,q) model, where p is the number of parameters in the autoregressive operator, d is the number of times that the series has to be differenced to induce stationarity, and q is the number of parameters in the moving average operator. Provided that the series does not contain seasonality, the ARIMA model (9) with small values of p, d, and q is capable of describing a wide range of practically occurring time series. When $d > 0$, the stationary series $w_t = \nabla^d z_t$ will usually have a zero mean. However, a useful generalization of (9) can be obtained by allowing w_t to have a nonzero mean, i.e.,

$$\phi(B)(\nabla z_t - \mu) = \theta(B)a_t. \tag{10}$$

With $d > 0$, model (10) is capable of describing a *deterministic* polynomial trend of degree d as well as a stochastic nonstationary component. For example, when $d = 1$, the model is capable of describing nonstationary stochastic behavior over and above an underlying linear growth rate.

SPECIAL CASES OF NONSEASONAL ARIMA (*P, D, Q*) MODELS

With $p = 0$, $d = 1$, $q = 1$, $\mu = 0$, model (10) becomes

$$\nabla z_t = (1 - \theta B)a_t \tag{11}$$

i.e., a nonstationary series whose first difference is stationary and can be represented as a first-order moving average model. Model (11) can be inverted to give

$$z_t = (1 - \theta)(z_{t-1} + \theta z_{t-2} + \theta^2 z_{t-3} + \cdots) + a_t. \tag{12}$$

Thus the one-step-ahead forecast of z_t from origin $(t - 1)$ is an exponentially weighted moving average of past values of the series. By solving the difference equation, model (11) may also be written

$$\begin{aligned} z_t &= a_t + (1 - \theta)(a_{t-1} + a_{t-2} + \cdots) \\ &= a_t + l_{t-1} \end{aligned} \tag{13}$$

Although the series has no fixed mean, at a given time it has a local level l_{t-1} which is updated from time $(t - 1)$ to time t according to

$$l_t = l_{t-1} + (1 - \theta)a_t.$$

Thus when the random shock a_t occurs, a proportion $(1 - \theta)a_t$ of it is absorbed into the "level" of the series and the remaining proportion θa_t is "lost" from the system.

With $p = 0, d = 1, q = 1, \mu \neq 0$, model (10) becomes

$$\nabla z_t - \mu = (1 - \theta B)a_t. \tag{14}$$

The solution of the difference equation (14) may be written as a complementary function (the solution of $\nabla z_t - \mu = 0$) and the particular integral (13), i.e.,

$$\begin{aligned} z_t &= c + \mu t + a_t \\ &\quad + (1 - \theta)(a_{t-1} + a_{t-2} + \cdots) \end{aligned} \tag{15}$$

and thus contains a "deterministic drift" term.

With $p = 0, d = 2, q = 2, \mu = 0$, model (10) becomes

$$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t. \tag{16}$$

Thus z_t and ∇z_t are nonstationary series, and second-order differencing $\nabla^2 z_t$ is necessary to induce stationarity. It may be shown [1, p. 111] that model (16) implies that the series has a local "level" l_t and a local "slope" s_t which are updated from time $t - 1$ to time t by the new random shock a_t according to

$$\begin{aligned} l_t &= l_{t-1} + s_{t-1} + (1 + \theta_2)a_t, \\ s_t &= s_{t-1} + (1 - \theta_1 - \theta_2)a_t. \end{aligned}$$

SEASONAL ARIMA MODELS

One of the deficiencies in handling seasonal time series in the past has been the absence of parametric models to describe seasonal behavior. A new class of models [3] for describing seasonality as well as nonstationary trends can be obtained by modification of the nonseasonal ARIMA model (10). Suppose that data become available at monthly intervals and that they are set out in the form of a two-way table in which the columns denote months and the rows denote years. The series for a particular column, say March, may contain a trend but is not seasonal in its behavior. Hence it is reasonable to link the observation for March in this year to observations in previous Marches by an ARIMA (P, D, Q) model of the form (9):

$$\Phi(B^S)\nabla_S^D z_t = \Theta(B^S)\alpha_t, \tag{17}$$

where the autoregressive and moving average operators are now polynomials in B^s of degrees P and Q, respectively, and s is the seasonal period and equals 12 in this case. Also, the nonseasonal difference operator ∇ in (9) is replaced by the seasonal difference operator $\nabla_s z_t = z_t - z_{t-s}$ in (17) and ∇_s^D denotes the Dth seasonal difference. In general, the error terms α_t in models of the form (17) fitted to each month separately would not be random since the behavior of the series in March of this year will usually depend not only on what happened in previous Marches but also on the behavior of the series in February, January, etc., of this year. To describe this monthly dependence, we can use the nonseasonal ARIMA model

$$\phi(B)\nabla^d \alpha_t = \theta(B)a_t, \qquad (18)$$

where a_t is now a random series and $\phi(B)$ and $\theta(B)$ are polynomials in B of degrees p and q, respectively. Substituting (18) in (17), and allowing for the possibility that the differenced series may have a nonzero mean μ, we obtain the ARIMA (p, d, q) $\times(P, D, Q)$ multiplicative model

$$\begin{aligned}\phi_p(B)\Phi_P(B^s)(\nabla^d\nabla_s^D z_t - \mu)\\ = \theta_q(B)\Theta_Q(B^s)a_t,\end{aligned} \qquad (19)$$

where the subscripts on the operators denote the degrees of the polynomials involved.

In some cases, it may be better to work with a nonmultiplicative model in which the autoregressive operator, or the moving average operator, or both operators, cannot be factorized into a product of nonseasonal and seasonal operators, as in (19); for example, the right-hand side of (19) might take the form

$$(1 - \theta_1 B - \theta_{12}B^{12} - \theta_{13}B^{13})a_t.$$

Seasonal models of the form (19) may be fitted to data with a range of seasonal periods: e.g., daily data ($s = 7$), weekly data ($s = 12$), and quarterly data ($s = 4$). Moreover, several seasonal periods may occur simultaneously; e.g., hourly traffic data may display a cycle over a day ($s = 24$) and a further cycle over a week ($s = 168$). In such examples it may be necessary to add further seasonal autoregressive, moving average, and differencing operators to the model.

BUILDING ARIMA MODELS

Figure 1a shows part of a series consisting of the logarithms of the electricity consumption in one country. The series contains an upward trend and an annual cycle. ARIMA models of the form (19) may be fitted to data, using an iterative cycle of identification, estimation, and checking, as described below.

Initial analysis [4] suggested that to achieve a homoscedastic* distribution of the residuals a_t, it is necessary to apply a logarithmic transformation $\ln z_t$ to the data before fitting a model of the form (19). Alongside the plot of $\ln z_t$ shown in Fig. 1a is a plot of the autocorrelation function r_k of $\ln z_t$ as a function of the lag k. The autocorrelation function fails to damp out with the lag k and is indicative of nonstationarity [1]. Figure 1b shows the autocorrelation function of the nonseasonal difference $\nabla \ln z_t$. This autocorrelation function has peaks at $12, 24, 36, \ldots$, indicating nonstationarity with respect to the seasonal behavior and suggesting that further seasonal differencing is needed. Figure 1c shows the autocorrelation function of $\nabla\nabla_{12} \ln z_t$. This function contains no obvious trends, implying that the differenced and transformed series $w_t = \nabla\nabla_{12} \ln z_t$ is stationary. The next step is to arrive at an initial guess of the seasonal and nonseasonal autoregressive and moving average structure needed to explain the autocorrelation function of w_t. The autocorrelation functions of autoregressive-moving average models are characterized by a discrete number of spikes corresponding to the moving average part of the model and damped exponentials and/or damped sine waves corresponding to the autoregressive part of the model. The largest autocorrelations r_k in Figure 1c occurs at lags 1 and 12, suggesting an initial model of the form

$$\nabla\nabla_{12} \ln z_t = (1 - \theta B)(1 - \Theta B^{12})a_t, \qquad (20)$$

where we may take as initial estimates of the parameters, $\theta = 0.30$ (based on $r_1 = -0.27$) and $\hat{\Theta} = 0.35$ (based on $r_{12} = -0.33$), using a procedure described in Box and Jenkins [1].

The initial model structure (20) may be fitted to the data by iterative calculation of the maximum-likelihood* estimates, starting

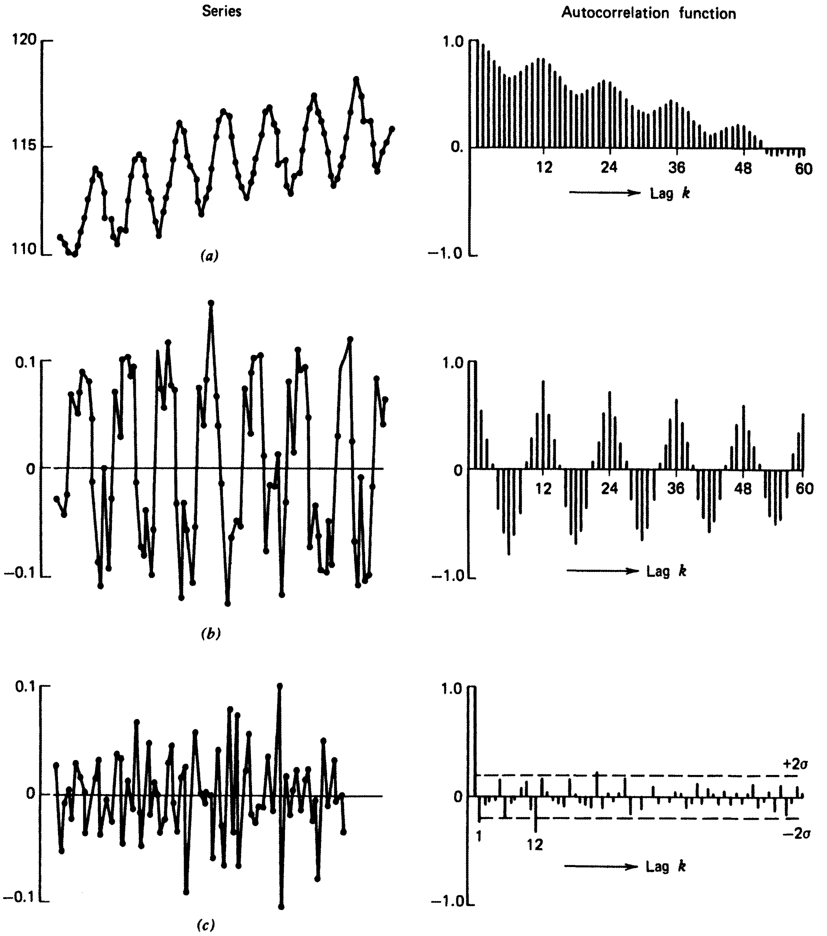


Figure 1. Various differences of the logarithms of national electricity consumption series, together with their corresponding autocorrelation functions: (a) $\ln Y_t$; (b) $\nabla \ln Y_t$; (c) $\nabla\nabla_{12} \ln Y_t$. Reproduced with the permission of GJP Publications from *Practical Experiences with Modeling and Forecasting Time Series* by Gwilym M. Jenkins.

from the initial values given above (see ref. 1, pp. 269–284, for the likelihood* function of autoregressive-moving average models). The fitted model, based on $N = 96$ observations, was

$$\nabla\nabla_{12} \ln Y_t = \begin{array}{cc} (1 - 0.73B) & (1 - 0.83B^{12}) \\ \pm 0.08 & a_t \pm 0.05 \end{array} \qquad (21)$$

with estimated residual variance $\sigma_a^2 = 0.0006481 (\sigma_a = 0.0255)$. The ± values underneath the estimated parameters denote the 1-standard-error* limits.

Examination of the residuals a_t in (21) showed that none of the residuals was large compared with their standard deviation $\sigma_a = 0.0255$ and that 5 residuals out of 83 fell outside $\pm 2\sigma_a$, in reasonable accord with expectation. The largest autocorrelations $r_a(k)$ of the residuals a_t occurred at lags 6 and 24, suggesting some evidence of model inadequacy. However, further elaboration of the model revealed little improvement on model (21). Further details of how this model was built, and how it was elaborated to a transfer function* model relating electricity consumption to temperature, have been given by Jenkins [4].

Model (21) may be used to forecast future values z_{t+l} for each lead time $l = 1, 2, 3, \ldots$

from the current origin t by writing it at time $t + l$ in the form

$$\begin{aligned}\ln z_{t+l} - \ln z_{t+l-1} - \ln z_{t+l-12} + \ln z_{t+l-13} \\ = a_{t+l} - 0.73a_{t+l-1} - 0.83a_{t+l-12} \\ + 0.61a_{t+l-13}\end{aligned} \qquad (22)$$

and then taking conditional expectations at time t, bearing in mind that the conditional expectation of future values of the random series a_{t+l} for $l > 0$ are zero. For example, when $l = 1$, the one-step-ahead forecast $\hat{z}_t(1)$ can be calculated from

$$\begin{aligned}\ln \hat{z}_t(1) = \ln z_t + \ln z_{t-11} - \ln z_{t-12} \\ - 0.73a_t - 0.83a_{t-11} + 0.61a_{t-12},\end{aligned} \qquad (23)$$

where $a_{t-j} = \ln z_{t-j} - \ln \hat{z}_{t-j-1}(1)$ for $j \geqslant 0$. Thus the forecasts for each lead time l can be generated recursively, together with the standard deviations of the forecast errors $e_t(l) = \ln \hat{z}_{t+l} - \ln z(l)$ (see ref. 1). In the example above, further improvements in forecasting accuracy could be expected by introducing into the model other variables which are related with electricity consumption: e.g., temperature, industrial production, price. Such *transfer function* models are discussed in Box and Jenkins [1] and in Jenkins [4].

MULTIVARIATE ARIMA MODELS

Univariate ARIMA models may be generalized to deal with mutual interaction between several nonstationary time series. To illustrate the possibilities, consider two time series z_{1t} and z_{2t}. First, nonlinear transformation and nonseasonal differencing may be needed to produce stationary time series

$$w_{1t} = \nabla^{d_1} z_{1t}, \qquad w_{2t} = \nabla^{d_2} z_{2t}. \qquad (24)$$

Then it might be possible to describe the resulting stationary vector by a multivariate autoregressive–moving average model*

$$\begin{bmatrix} \phi_{11}(B) & \phi_{12}(B) \\ \phi_{21}(B) & \phi_{22}(B) \end{bmatrix} \begin{bmatrix} w_{1t} - \mu_1 \\ w_{2t} - \mu_2 \end{bmatrix} = \begin{bmatrix} \theta_{11}(B) & \theta_{12}(B) \\ \theta_{21}(B) & \theta_{22}(B) \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}, \qquad (25)$$

where a_{1t} and a_{2t} are the one-step-ahead forecast errors or residuals for z_{1t} and z_{2t}, respectively. If the forecasts are to be optimal, a_{1t} must be a random series, a_{2t} a random series, and a_{1t} and a_{2t} mutually uncorrelated series except possibly at simultaneous times. The model defined by (24) and (25) is an example of an ARIMA($\boldsymbol{P}$, $\boldsymbol{d}$, $\boldsymbol{Q}$) model, where $\mathbf{P}$ is a matrix whose elements (p_{ij}) define the degrees of the polynomials $\phi_{ij}(B)$ in the autoregressive matrix, the vector $\mathbf{d}$ has elements d_i which define the degrees of differencing needed to induce stationarity of the time series*, and $\mathbf{Q}$ is a matrix whose elements (q_{ij}) define the degrees of the polynomials $\theta_{ij}(B)$ in the moving average matrix. The foregoing models may also be generalized to deal with seasonality [2,4], and may be generalized by introducing explanatory variables* to explain the simultaneous behavior of the vector $\mathbf{z}_t$ of time series, leading to multiple output–multiple input transfer function models [4].

REFERENCES

1. Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco (2nd ed., 1976).
2. Box, G. E. P., Hillmer, S. C., and Tiao, G. C. (1976). *NBER Census Conf. Seasonal Time Ser.*, Washington, D. C.
3. Box, G. E. P., Jenkins, G. M., and Bacon, D. W. (1967). In *Advanced Seminar on Spectral Analysis of Time Series*, *B. Harris*, ed., Wiley, New York, pp. 271–311.
4. Jenkins, G. M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*. GJP Publications, St. Helier, N.J.

Further Reading

The basic theoretical properties of ARIMA models are given in *Time Series Analysis: Forecasting and Control* by G. E. P. Box and G. M. Jenkins (Holden-Day, San Francisco, 1970), together with practical procedures for identifying, fitting, and checking such models. Further accounts are given in *Applied Time Series Analysis for Managerial Forecasting* by C. R. Nelson (Holden-Day, San Francisco, 1973), *Time Series Analysis and Forecasting: The Box-Jenkins Approach* by O. D. Anderson (Butterworth, London, 1975), and *Forecasting Economic Time Series* by C. W. J. Granger and P. Newbold (Academic Press, New York, 1977). Theoretical and practical accounts of multivariate ARIMA models are given in *Practical Experiences with Modelling and Forecasting Time Series* by G. M. Jenkins (GJP Publications, St. Helier, N.J., 1979). Numerous papers in the field are published in *Statistical Literature* and are listed in *Current Index to Statistics** (CIS).

See also AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS; PREDICTION AND FORECASTING; SEASONALITY; TIME SERIES; and TRANSFER FUNCTION MODEL.

G. M. JENKINS

AUTOREGRESSIVE MODELS.

See AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS

AUTOREGRESSIVE–MOVING AVERAGE (ARMA) MODELS

Data occurring in the form of time series* occur in many branches of the physical sciences, social sciences, and engineering. Figure 1 shows an example of a single time series* in which the value z_t of a certain variable (the average number of spots on the sun's surface) is plotted against time t. Earlier approaches to analyzing time series were based on decomposing the variance of the series into components associated with different frequencies, based on Fourier analysis* and leading more recently to methods based on spectral analysis*.

Alternative historical approaches were based on building a model for the time series in the time domain. The main motivation for building such a model was to forecast future values of the time series given its past history. However, such a model could be used (a) to gain a better understanding of the mechanisms generating the series, (b) to smooth the random variation in the series, and (c) to allow for dependence in the series when the data were used for other statistical purposes, such as testing the differences between the means of two sets of data or relating one time series to another as in some form of regression* analysis. The first practically useful models for describing time series were the autoregressive models introduced by G. U. Yule and G. Walker (see below) and the moving average* models introduced by Slutsky, Wold, and others. Later, it was recognized that more general structures could be obtained by combining autoregressive and moving average models, leading to autoregressive-moving average (ARMA) models.

GENERAL LINEAR MODEL*

When forecasting* an observed time series z_t from a knowledge of its past history, it is natural to think of the forecast as being obtained by applying a set of weights to past values of the time series. Thus, the one-step-ahead forecast of z_t made from *origin* $(t-1)$ may be written

$$\text{forecast} = \pi_1 z_{t-1} + \pi_2 z_{t-2} + \pi_3 z_{t-3} + \cdots, \tag{1}$$

where π_j is the weight applied to the previous observation z_{t-j} in order to forecast z_t. When the future observation z_t comes to hand, it follows that

$$z_t = \text{forecast} + \text{forecast error} \tag{2}$$

Substituting for the forecast from (1) and denoting the forecast error by a_t, (2) becomes

$$z_t = \pi_1 z_{t-1} + \pi_2 z_{t-2} + \pi_3 z_{t-3} + \cdots + a_t. \tag{3}$$

If the forecast one step ahead of z_t is the best possible, then the forecast errors a_t, a_{t-1}, $a_{t-2} \ldots$ should be a *random series**, or *white noise**. If not, it should be possible to forecast the forecast errors and add this forecast to forecast (1) to obtain a better forecast. Model (3), with a_t a random series, is called a linear model*. In practice, the forecast errors a_t may depend on the level of the series, in which case a better representation is obtained by using a nonlinear transformation of z_t in (3), such as a log or square-root transformation*. From now on it will be assumed that the notation used in (3) denotes a representation for z_t or a suitable nonlinear transformation of z_t chosen so as to make the forecast errors a_t homoscedastic*. Although the representation (3) provides a useful general way

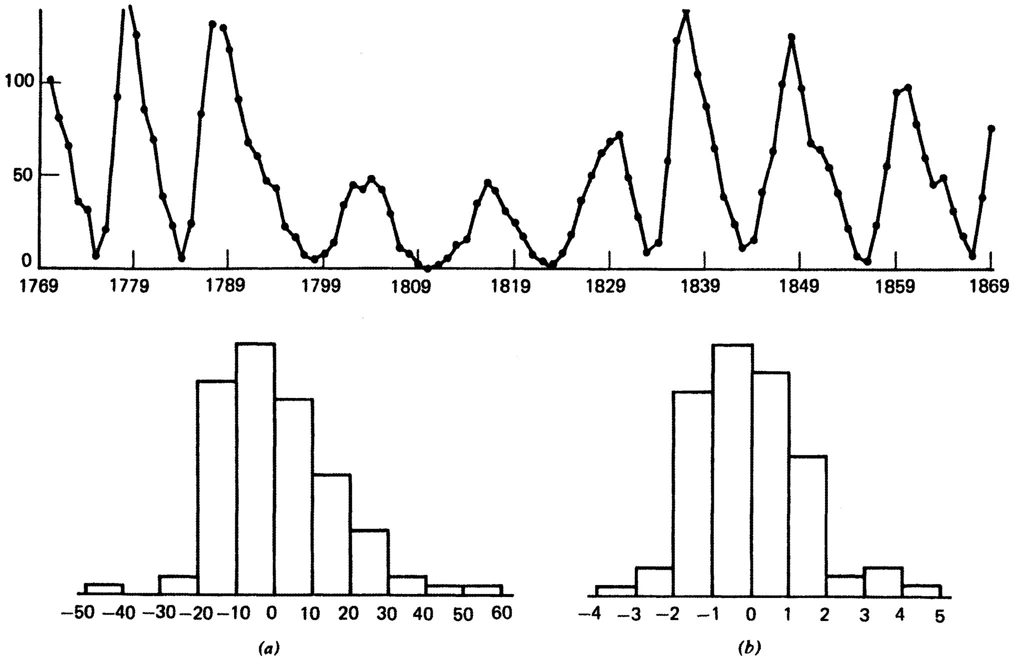


Figure 1. Plot of annual Wölfer sunspot numbers (1770–1869), together with histograms of residuals a_t from: (a) model $(z_t - 46.9) = 1.42(z_{t-1} - 46.9) - 0.73(z_{t-2} - 46.9) + a_t$; ($b$) model $(\sqrt{z_t} - 7.4) + 1.41(\sqrt{z_{t-1}} - 7.4) - 0.70(\sqrt{z_{t-2}} - 7.4) + a_t$.

of modeling a time series, it suffers from the disadvantage that it contains a large (potentially infinite) number of weights or parameters π_i. Since it would be impossible to estimate very accurately such a large number of weights, a practical solution to time-series problems requires a more parsimonious representation, containing as few parameters as possible. Such economy in parameterization can be achieved using autoregressive and moving average models.

PURE AUTOREGRESSIVE MODELS

From now on it is assumed that the time series is *stationary**, i.e., that it is in statistical equilibrium about a fixed mean μ and that it possesses, among other properties, a constant variance and a covariance structure which depends only on the difference k (or lag) between two time points. Suppose also that the weights π_i applied to past observations in the representation (3) are zero beyond a certain point p. Then, writing the series as a deviation about its mean μ, (3) becomes

$$\begin{aligned}(z_t - \mu) &= \phi_1(z_{t-1} - \mu) + \phi_2(z_{t-2} - \mu) \\ &\quad + \cdots + \phi_p(z_{t-p} - \mu) + a_t, \qquad (4)\end{aligned}$$

where the finite set of weights or parameters ϕ_i may be estimated from the data. In words, (4) implies that the current deviation of the time series from its mean is a linear combination of the p previous deviations plus a random residual* a_t. The analogy between (4) and a multiple regression model* should be noted. Because the regressor variables in (4) are lagged values of the series itself and not distinct variables, (4) is called an *autoregressive* model of order p, or an AR(p). Model (4) also implies that the best forecast of z_t made from origin $(t - 1)$ is a linear combination of the past p-values of the series.

Introducing the backward-shift operator* B, such that $Bz_t = z_{t-1}$, $B^j z_t = z_{t-j}$, (4) may be written in the alternative form

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(z_t - \mu) = a_t. \qquad (5)$$

Thus an AR(p) model is characterized by an operator

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p), \quad (6)$$

which is a polynomial of degree p in the backward-shift operator B. The polynomial (6) may have real factors of the form $(1 - G_i B)$ or complex factors corresponding to complex roots of $\phi(B) = 0$. Complex factors indicate the presence of a quasi-cyclic* component in the data. Such cycles* do not have fixed periods, as in a sine wave, but are subject to random changes in amplitude, phase, and period. The fact that complex roots in (6) produce quasi-cyclical behavior in z_t may be seen by noting that if $p = 2$ and $a_t = 0$, the solution of the difference equation* (6) is a damped sine wave, as in the motion of a damped simple pendulum. When the zero on the right-hand side is replaced by a random series a_t, the sine wave is prevented from damping out by a series of random shocks, producing randomly disturbed sinusoidal behavior.

Autoregressive models were first suggested by G. U. Yule [9] who used a second-order model to describe the annual series of Wölfer sunspot numbers. Figure 1 shows a plot of this series, based on the annual average of daily readings, for the period 1770–1869. The fitted model* is

$$\begin{aligned}(z_t - 46.9) = {} & \underset{\pm 0.07}{1.42}(z_{t-1} - 46.9) \\ & - \underset{\pm 0.07}{0.73}(z_{t-2} - 46.9) + a_t, \end{aligned} \quad (7)$$

where the ± values underneath the estimated parameters are their estimated standard error limits. The variance of the residuals a_t can be estimated together with the parameters μ, ϕ_1, and ϕ_2 [2] and was $\sigma_a^2 = 228.0$ in this example. The operator $(1 - 1.42B + 0.73B^2)$ corresponding to (7) has complex factors with a period p that can be calculated from

$$\cos \frac{2\pi}{p} = \frac{\phi_1}{2\sqrt{-\phi_2}} = \frac{1.42}{2\sqrt{0.73}}$$

and is 10.65 years. Figure 1 also shows the histogram* of the residuals a_t corresponding to model (7). The distribution is skew, suggesting that a transformation of the data is needed before fitting a model. Using an approach due to Box and Cox [1] (*see* BOX–COX TRANSFORMATION—I), it may be shown that a better representation is obtained using the following model based on a square-root transformation*:

$$\begin{aligned}(\sqrt{z_t} - 7.4) = {} & \frac{1.41}{\pm 0.07}(\sqrt{z_{t-1}} - 7.4) \\ & - \frac{0.70}{\pm 0.07}(\sqrt{z_{t-2}} - 7.4) + a_t \end{aligned} \quad (8)$$

with $\sigma_a^2 = 1.994$. Note that the parameter estimates are changed only very slightly by transformation. Its main affect is to shrink the peaks and stretch the troughs in the series, resulting in a more symmetric distribution* of the residuals, as shown in Fig. 1. The estimate of the average period corresponding to model (8) is 11.05 years, much closer than the previous value of 10.65 years to the period quoted by meteorologists for this series.

PURE MOVING AVERAGE MODELS

For autoregressive models the π-weights in the representation (8) have a cut-off after p, where p is the order of the model. In some situations it may be more appropriate to apply steadily declining weights to generate the forecasts rather than weights which have an abrupt cut-off. Such a pattern of weights may be obtained, e.g., by using a moving average model

$$z_t - \mu = a_t - \theta a_{t-1} = (1 - \theta B)a_t. \quad (9)$$

Inverting (9), we obtain

$$a_t = \frac{1}{1 - \theta B}(z_t - \mu)$$

and provided that $|\theta| < 1$ and $|B| < 1$, this expression may be expanded to give

$$a_t = (1 + \theta B + \theta^2 B^2 + \cdots)(z_t - \mu), \quad (10)$$

so that the π-weights decay exponentially. More generally, a moving average model of order q, or MA(q) is defined by

$$\begin{aligned}(z_t - \mu) = {} & a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \\ & - \cdots - \theta_q a_{t-q} \end{aligned} \quad (11)$$

where a_t is a random series. The model (11) contains q parameters $\theta_1, \theta_2, \ldots, \theta_q$, which can be estimated from the data. It implies that the current deviation of the series z_t from its mean μ is a linear combination of the current and q previous random shocks a_t (or one-step-ahead forecast errors) which have entered the system. In backward-shift notation, (11) may be written as $(z_t - \mu) = \theta(B)a_t$, where the moving average operator $\theta(B)$ is given by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q. \quad (12)$$

The model (11) has π-weights consisting of a mixture of real exponentials, corresponding to real factors of $\theta(B)$, and of damped sine waves, corresponding to complex factors of $\theta(B)$.

MIXED AUTOREGRESSIVE–MOVING AVERAGE MODELS

Result (10) shows that an MA(1) can be written as an autoregressive model of infinite order. If θ is small, say $\theta = 0.3$, then from a practical point of view the infinite series in (10) can be truncated after the term in B since it would require a long length of series to detect the parameter 0.09 in the next term $0.09B^2$ in the expansion. However, if θ is moderate or large, several terms would be needed in (10) to provide an adequate approximation to the single-parameter model (9). Thus if the moving average model were incorrectly specified as an autoregressive model, involving several parameters, the estimates of the parameters in the autoregressive representation would tend to have high standard errors and be highly correlated.

Conversely, the AR(1) model

$$(1 - \phi B)(z_t - \mu) = a_t \quad (13)$$

can be written as an infinite-order moving average model

$$(z_t - \mu) = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots, \quad (14)$$

and hence estimation problems will be encountered if an autoregressive model is incorrectly specified as a moving average model. To achieve parsimony in parameterization in a given practical situation, it may be necessary to include both autoregressive and moving average terms in the model. Thus the mixed autoregressive–moving average model, or ARMA(p, q), is defined by [8]

$$\begin{aligned}(z_t - \mu) &= \phi_1(z_{t-1} - \mu) + \cdots \\ &\quad + \phi_p(z_{t-p} - \mu) + a_t - \theta_1 a_{t-1} \\ &\quad - \cdots - \theta_q a_{t-q}. \quad (15)\end{aligned}$$

Written in terms of the backward shift operator, (15) becomes

$$(z_t - \mu) = \frac{1 - \theta_1 B - \cdots - \theta_q B^q}{1 - \phi_1 B - \cdots - \phi_p B^p} a_t. \quad (16)$$

The form (15) represents the time series z_t (or an appropriate nonlinear transformation of z_t) as the output from a linear filter whose input is a random series and whose *transfer function** is a rational function of the backward-shift operator B. In words, (15) implies that the current deviation of the time series from its mean is a linear combination of the p previous deviations and of the current and q previous residuals a_t (or one-step-ahead forecast errors). The ARMA(p, q) model (15) is capable of generating π-weights in (1), the first p of which follow no fixed pattern and the remainder of which lie on a curve that is a mixture of damped exponentials and sine waves. Table 1 shows special cases of the general ARMA(p, q) model of the kind that frequently arise in practice.

STATIONARITY AND INVERTIBILITY CONDITIONS

The parameters in the ARMA(p, q) model (16) must satisfy the following two conditions. (a) For z_t to be written in the π-weight form (1), i.e.,

$$\begin{aligned}\left(1 - \sum_{j=1}^{\infty} \pi_j B^j\right)(z_t - \mu) &= \theta^{-1}(B)\phi(B)z_t \\ &= a_t \\ &\quad \text{for } |B| < 1,\end{aligned}$$

the factors of $\theta(B)$ must be less than unity in modulus (the invertibility condition). This condition implies that the forecast weights

π_j die out; i.e., the forecast depends less on what has happened in the distant past than in the recent past. (b) For z_t to be written in the ψ-weight form

$$(z_t - \mu) = \left(1 + \sum_{j=1}^{\infty} \psi_j B^j\right) a_t = \phi^{-1}(B)\theta(B)a_t \quad \text{for } |B| < 1,$$

the factors of $\theta(B)$ must be less than unity in modulus (the stationarity condition). This condition implies that the series is stationary with finite variance. Table 2 shows the characteristic shapes of the π-weights and ψ-weights for the AR(p), MA(q), and AR-MA(p, q) models.

AUTOCORRELATION FUNCTIONS

The autocovariance functions $\gamma_k = E[(z_t - \mu)(z_{t+k} - \mu)]$ shows how the dependence between neighboring values of the series varies with the lag* k. It may be calculated from the autocovariance generating function

$$\begin{aligned}\gamma(B) &= \sum_{k=-\infty}^{\infty} \gamma_k B^k \\ &= \sigma_a^2 \phi^{-1}(B)\phi^{-1}(B^{-1})\theta(B)\theta(B^{-1}). \quad (17)\end{aligned}$$

Table 2 shows the characteristic patterns of the autocorrelation functions $\rho_k = \gamma_k/\gamma_0$ of AR(p), MA(q), and ARMA (p, q) models. Such patterns may be used to provide an initial guess of the structure of an observed time series [2].

PARTIAL AUTOCORRELATION FUNCTIONS

A complementary tool to the autocorrelation function for identifying the structure of an ARMA(p, q) model is the partial autocorrelation function s_k [2]. The partial autocorrelation function may be estimated by fitting autoregressive models of orders $1, 2, 3, \ldots, k$ to a time series and picking out the estimates $s_1, s_2, \ldots, s_k$ of the *last* parameter in the model. Table 2 shows the partial autocorrelation function shapes corresponding to AR(p), MA(q), and ARMA(p, q) models. The duality in the properties of autoregressive and moving average models should be noted.

MULTIVARIATE AUTOREGRESSIVE–MOVING AVERAGE MODELS

If $\mathbf{z}_t$ denotes an m-vector of mutually interacting time series, the univariate ARMA(p, q) model (16) may be generalized to

$$\boldsymbol{\phi}(B)(\mathbf{z}_t - \boldsymbol{\mu}) = \boldsymbol{\theta}(B)\mathbf{a}_t, \quad (18)$$

where $\boldsymbol{\phi}(B)$ is an autoregressive matrix whose elements $\phi_{ij}(B)$ are autoregressive operators, $\boldsymbol{\mu}$ a vector of mean values, $\boldsymbol{\theta}(B)$ a moving average matrix with elements $\theta_{ij}(B)$, and $\mathbf{a}_t$ a vector of random series that are mutually uncorrelated. For further discussion of the properties of multivariate ARMA models, the reader is referred to Quenouille [7], Hannan [4], Box and Tiao [3], and Jenkins [5,6].

Since time series occurring in practice are rarely stationary with fixed means, ARMA models are of limited use in describing practical situations. The modifications necessary

Table 1. Some Simple Special Cases of the Autoregressive–Moving Average Model

(p, q)	Nature of Model	Mathematical Form of Model	Backward-Shift-Operator Form of Model
(1, 0)	First-order autoregressive	$z_t - \mu = \phi_1(z_{t-1} - \mu) + a_t$	$z_t - \mu = \frac{1}{1 - \phi_1 B} a_t$
(2, 0)	Second-order autoregressive	$z_t - \mu = \phi_1(z_{t-1} - \mu) + \phi_2(z_{t-2} - \mu) + a_t$	$z_t - \mu = \frac{1}{1 - \phi_1 B - \phi_2 B^2} a_t$
(0, 1)	First-order moving average	$z_t - \mu = a_t - \theta_1 a_{t-1}$	$z_t - \mu = (1 - \theta_1 B)a_t$
(0, 2)	Second-order moving average	$z_t - \mu = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$	$z_t - \mu = (1 - \theta_1 B - \theta_2 B^2)a_t$
(1, 1)	First-order autoregressive, first-order moving average	$z_t - \mu = \phi_1(z_{t-1} - \mu) + a_t - \theta_1 a_{t-1}$	$z_t - \mu = \frac{1 - \theta_1 B}{1 - \phi_1 B} a_t$

Table 2. Summary of Properties of AR, MA, and ARMA Models

	AR(p) Models	MA(q) Models	ARMA(p, q) Models
π-weights and partial autocorrelations function	Cutoff after p; follow no fixed pattern	Infinite; mixture of damped exponentials and sine waves	First p values follow no fixed pattern; thereafter, mixture of damped exponentials and sine waves
ψ-weights and autocorrelation function	Infinite; mixture of damped exponentials and sine waves	Cutoff after q; follow no fixed pattern	First q values follow no fixed pattern; thereafter, mixture of damped exponentials and sine waves

to make them practically useful are discussed under autoregressive-integrated moving average (ARIMA) models*.

REFERENCES

1. Box, G. E. P. and Cox, D. R. (1964). *J. Roy. Statist. Soc. Ser. B*, **26**, 211–252.
2. Box, G. E. P. and Jenkins, G. M., (1970). *Time Series Analysis; Forecasting and Control*. Holden-Day, San Francisco (2nd ed., 1976).
3. Box, G. E. P. and Tiao, G. C. (1977). *Biometrika*, **64**, 355–366.
4. Hannan, E. J. (1970). *Multiple Time Series*, Wiley, New York.
5. Jenkins, G. M. (1975). *Proc. 8th Int. Biometric Conf.*, Constanta, Romania, 1974. Editura Academici Republicii Socialists Romania, 53.
6. Jenkins, G. M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*, GJP Publications, St. Helier, N. J.
7. Quenouille, M. H. (1957). *The Analysis of Multiple Time Series*, Charles Griffin, London.
8. Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Stockholm (2nd ed., 1953).
9. Yule, G. U. (1927). *Philos. Trans. Roy. Soc. A.*, **226**, 267.

FURTHER READING

The first balanced account of theoretical and practical aspects of autoregressive–moving average models is given in the book by Wold [8]. Box and Jenkins [2] summarize the properties of these models and also give practical guidelines for identifying, fitting, and checking such models given the data. Pioneering work on multivariate models is to be found in Quenouille [7], and Hannan [4] discusses the theoretical background for both univariate and multivariate models. Practical guidelines for building multivariate autogressive-moving average models have been given by Jenkins [6].

See also Autoregressive–Integrated Moving Average (ARIMA) Models; Moving Averages; Prediction and Forecasting; Time Series; and Transfer Function Model.

G. M. Jenkins

AVAILABILITY

Availability is a property of a system, defined as the proportion of time the system is functioning (properly). If failure and repair times are each distributed exponentially* with expected values θ and ϕ, then the availability is $\theta/(\theta + \phi)$. Sometimes the availability is defined generally in this way with $\theta = E$ [time to failure], $\phi = E$ [repair time].

BIBLIOGRAPHY

Gray, H. L. and Lewis, T. (1967). *Technometrics*, **9**, 465–471.

Nelson, W. (1968). *Technometrics*, **10**, 594–596.

See also Quality Control, Statistical and Reliability, Probabilistic.

AVERAGE CRITICAL VALUE METHOD

Geary's [1] average critical value (ACV) method applied to a statistic T used to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$ determines what the difference between θ_0 to θ_1 should be for $E[T|\theta_1]$ to fall on the boundary of the critical region of the test. Test statistics with small differences correspond to tests with high efficiency*. The advantage of this method is that it requires only the calculation of the expected value $E[T|\theta_1]$ rather than the distribution of T. (The latter is required for power* function calculations.) Examples are given by Geary [1] and Stuart [2]. Stuart in 1967 showed that the ACV method of gauging the efficiency of tests can usually be represented as an approximation to the use of the asymptotic relative efficiency* of the tests.

REFERENCES

1. Geary, R. C. (1966). *Biometrika*, **53**, 109–119.
2. Stuart, A. (1967). *Biometrika*, **54**, 308–310.

See also EMPIRICAL DISTRIBUTION FUNCTION (EDF) STATISTICS.

AVERAGE EXTRA DEFECTIVE LIMIT (AEDL)

The average extra defective limit (AEDL) is a concept introduced by Hillier [1] as a measure of effectiveness of a continuous sampling plan* in adjusting to a process that has gone out of control.

Assume that a process has been operating in control at a quality level p_0 and then instantaneously deteriorates after the mth item to a level p_1, where $0 \leqslant p_0 < p_1 \leqslant 1$. Let D be the number of *uninspected* defectives among the next L items after the mth item is observed. The expected value of D, $E\ (D)$, is a well-defined quantity for a specific sampling plan. An average extra defective limit is the smallest number denoted by AEDL satisfying

$$E(D) \leqslant \text{ AEDL } + L \times A$$

for all possible values of L, p_0, or p_1, where A is the average outgoing quality limit (AOQL)* of the plan.

Equivalently, let

$$X_m = \begin{cases} 1 & \text{if the } m\text{th item is defective} \\ & \quad \text{but not inspected,} \\ 0 & \text{otherwise.} \end{cases}$$

Let the m_0th item be the last item before the shift in quality from p_0 to p_1; then

$$\text{AEDL} = \sup_{(p_0,p_1)} \sup_{L} \sum_{m=m_0+1}^{m_0+L} [E(X_m) - A].$$

Intuitively, AEDL is the upper limit to the expected number of "extra" defectives that will be left among outgoing items when the process goes out of control regardless of L, p_0, or p_1. Additional interpretations of AEDL, its uses, and methods of computation for continuous sampling plans have been discussed by Hillier [1].

REFERENCE

1. Hillier, F. S. (1964). *Technometrics*, **6**, 161–178.

See also AVERAGE OUTGOING QUALITY LIMIT (AOQL); CONTINUOUS SAMPLING PLANS; and QUALITY CONTROL, STATISTICAL.

AVERAGE OUTGOING QUALITY (AOQ)

The definition of this concept suggested by the Standards Committee of ASQC [3] is: "the expected quality of the outgoing product following the use of an acceptance sampling plan* for a given value of incoming product quality." It is basically a ratio of defective items to total items, i.e., the total number of defectives in the lots accepted divided by the total number of items in those lots. Two other formal (but not equivalent) definitions are: (1) the average fraction defective in all lots after rejected lots have been sorted and cleared of defects—this is an average based on practically perfect lots (those sorted) and lots still with fraction of defectives approximately p (it is assumed that lots of stable quality are offered), and (2) the expected fraction of defectives, after substituting good items for bad ones in rejected lots, and in samples taken from accepted lots.

The AOQ serves as a performance measure associated with an (attribute) acceptance sampling plan* when the same sampling plan is used repeatedly.

Wortham and Mogg [4] present formulas for calculating AOQ for nine different ways of carrying out rectifying inspection*. For example, if the defective items are replaced by good ones (thereby returning the lot to its original size N), then

$$\text{AOQ} = \frac{P_a \times p \times (N - n)}{n},$$

where P_a is the probability of acceptance using the given acceptance plan, p the fraction defective, N the lot size, and n the sample size.

REFERENCES

1. Burr, I. W. (1953). *Engineering Statistics and Quality Control*, McGraw-Hill, New York.
2. Juran, J. M. ed. (1951). *Quality Control Handbook*. McGraw-Hill, New York.
3. Standards Committee of ASQC (1978). *Terms, Symbols and Definitions for Acceptance Sampling. ASQC Standard A*3.
4. Worthman, A. W. and Mogg, J. W. (1970). *J. Quality Tech.*, **2**(1), 30–31.

See also Average Outgoing Quality Limit (AOQL) and Quality Control, Statistical.

AVERAGE OUTGOING QUALITY LIMIT (AOQL)

The current “official” definition of this concept as suggested by the Standards Committee of ASQC [3] reads: “For a given acceptance sampling plan* AOQL is the maximum AOQ over all possible levels of incoming quality.”

Originally, AOQL was defined by Dodge [1] as the upper limit to the percent of defective units that remain in the output after inspection, given that the process is in statistical control (i.e., the proportion of defectives being produced is constant). In other words, it represents the worst average quality the consumer will accept under a particular rectifying inspection* scheme.

REFERENCES

1. Dodge, H. F. (1943). *Ann. Math. Statist.*, **14**, 264–279.
2. Sackrowitz, H. (1975). *J. Quality Tech.*, **7**, 77–80.
3. Standards Committee of ASQC (1978). *Terms, Symbols and Definitions for Acceptance Sampling*, *ASQC Standard A*3.
4. Wald, A. and Wolfowitz, J. (1945). *Ann. Math. Statist.*, **16**, 30–49.

See also Average Outgoing Quality (AOQ) and Sampling Plans.

AVERAGE RUN LENGTH (ARL)

The length of time the process must run, on the average, before a control chart* will indicate a shift in the process level* is called the average run length (ARL). It is, of course, desirable that the ARL should be long when no shift has occurred, but short when a shift has occurred.

The ARL is usually measured in terms of the number of consecutive points plotted on the control chart.

BIBLIOGRAPHY

Standards Committee of ASQC (1978). ASQC Standard AL.

See also Cumulative Sum Control Charts.

AVERAGE SAMPLE NUMBER (ASN)

In a sequential test* $\mathscr{S}$, the final size of the sample (N) required by the test is a random variable. If the sample sequential test $\mathscr{S}$ is carried out repeatedly, N will generally assume different values in successive repetitions of the test. The average amount of sampling per test that would result from the use of $\mathscr{S}$ is measured by the expected value of N and is called the average sampling number (ASN) of the test. If the test relates to the value of a parameter θ, we have formally

$$E[N; \theta] = \sum_{n=1}^{\infty} np(n; \theta),$$

where $p(n;\theta) = \Pr[N = n|\theta]$ is the probability of reaching the terminal decision at sample size n. A graph showing $E[N;\theta]$ against various values of θ is called the ASN curve or ASN surface, according as θ is a scalar or a vector, respectively.

BIBLIOGRAPHY

Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, Reading, Mass.

See also SEQUENTIAL ANALYSIS and STOPPING NUMBERS AND STOPPING TIMES.

AVERAGED SHIFTED HISTOGRAM

For a random univariate sample of size n, the classic histogram* takes the value

$$\hat{f}(x) = \frac{\text{bin count}}{nh},$$

where h is the width of a bin. Important details of proper histogram construction are usually overlooked. In practice, continuous data have a finite number of significant digits, with resultant accuracy $\pm\delta/2$. In this view, the raw data have been rounded and assume values at the midpoints of a finer histogram mesh with bin width δ. Let the kth such bin be denoted by $B_k = [t_k, t_{k+1})$, $-\infty < k < \infty$, where the bin edges $\{t_k\}$ satisfy $t_{k+1} - t_k = \delta$ for all k. If ν_k is the bin count for B_k, then $\sum_k \nu_k = n$. Thus the classic histogram should obey two constraints: first, the bin width should be a multiple of δ, that is, $h = m\delta$ for some integer m; and second, the bins of the histogram should conform to the finer mesh, for example, $[t_k, t_{k+m})$. The bin count for this wider bin, $[t_k, t_{k+m})$, would be computed as $\sum_{i=0}^{m-1} \nu_{k+i}$.

If the point of estimation, x, falls in the (narrower) bin B_k, precisely which histogram with bin width h should be selected? Clearly, there are exactly m such *shifted histograms*, with explicit bin intervals ranging from $[t_{k-m+1}, t_{k+1})$ to $[t_k, t_{k+m})$. The ordinary *averaged shifted histogram* (ASH) estimates the density at $x \in B_k$ as the arithmetic mean of these m shifted histogram estimates. A simple calculation shows that the ordinary ASH is given by

$$\begin{aligned}\hat{f}(x) &= \frac{1}{m}\sum_{i=1-m}^{m-1} \frac{(m-|i|)\nu_{k+i}}{nh}\\ &= \frac{1}{nh}\sum_{i=1-m}^{m-1}\left(1 - \frac{|i|}{m}\right)\nu_{k+i},\\ &\quad x \in B_k.\end{aligned}$$

As $m \to \infty$, Scott [3] showed that $\hat{f}(x)$ converges to the kernel density estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

with the triangle kernel $K(t) = 1 - |t|$ on $(-1, 1)$. The ASH may be generalized to mimic any kernel $K(t)$ defined on $(-1, 1)$ such that $K(\pm 1) = 0$ by

$$\hat{f}(x) = \frac{1}{n\delta}\sum_{i=1-m}^{m-1} w_m(i)\nu_{k+i}, \qquad x \in B_k,$$

by defining the weights

$$w_m(i) = \frac{K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)}.$$

In this form, the ASH is seen as a discrete convolution of adjacent bin counts. In practice, popular kernels are the biweight and triweight, which equal $\frac{15}{16}(1 - t^2)^2$ and $\frac{35}{32}(1 - t^2)^3$ on $(-1, 1)$, respectively. Silverman [6] provided a fast-Fourier-transform procedure for the Normal kernel, which does not have finite support. Fan and Marron [1] compare many algorithms.

MULTIVARIATE ASH

Binning* seems an effective device in dimensions up to four or five. For example, with bivariate data, (x_1, x_2), construct a fine mesh of size $\delta_1 \times \delta_2$. Then construct bivariate histograms with bins of size $h_1 \times h_2$, where $h_1 = m_1\delta_1$ and $h_2 = m_2\delta_2$. Then the bivariate

ASH is the average of the $m_1 \times m_2$ shifted histograms, and for $(x_1, x_2) \in B_{kl}$,

$$\hat{f}(x_1, x_2) = \frac{1}{n\delta_1\delta_2} \sum_{i=1-m_1}^{m_1-1} \sum_{j=1-m_2}^{m_2-1} w_{m_1}(i) w_{m_2}(j) \nu_{k+i,l+j}$$

with obvious extension to more dimensions. Scott [4] discusses visualization of trivariate and quadrivariate densities with applications in clustering and discrimination.

COMPUTATION AND ASYMPTOTICS

The computational advantage of the ASH derives from the fact that the smoothing is applied to perhaps 50–500 bin counts, rather than to the raw data themselves. ASH software is available from *Statlib* or from ftp.stat.rice.edu by anonymous ftp.

From the point of view of statistical efficiency, the ordinary ASH represents a compromise between the histogram and triangle-kernel estimator*. For the univariate ASH, the asymptotic mean integrated squared error (AMISE) is

$$\text{AMISE}(h, m) = \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) + \frac{h^2}{12m^2} R(f') + \frac{h^4}{144}\left(1 - \frac{2}{m^2} + \frac{3}{5m^2}\right) R(f''),$$

where $R(\phi) = \int \phi(x)^2\, dx$. When $m = 1$, the bias has the usual $O(h^2)$ behavior of the histogram, whereas the bias is of order $O(h^4)$ as $m \to \infty$ as for the positive kernel estimator. If a piecewise linear interpolant of the ASH midpoints is used, then the bias is $O(h^4)$ for all m. Thus, linear interpolation is almost always recommended for the ASH.

The exact optimal choices of h and m depend on the unknown density through $R(f'')$. However, Terrell and Scott [7] provide useful upper bounds for h using any reasonable estimate of the standard deviation: for the ordinary ASH with $m = 1$, $h < 3.73\sigma n^{-1/3}$; for the ASH with $m \to \infty$, $h < 2.78\sigma n^{-1/5}$. These estimates converge at the rates $O(n^{-2/3})$ and $O(n^{-4/5})$, respectively. However, if the linear interpolant is used, the bound for the ASH with $m = 1$ is $h < 2.33\sigma n^{-1/5}$; the $m \to \infty$ bound is unchanged. If δ is fixed, these formulas give upper bounds for m. A discussion of algorithms for choosing h may be found in Scott and Terrell [5] and Scott [4]. A simple table of factors to obtain smoothing parameters for use with other kernel weights is given on p. 142 in Scott [4].

ASH REGRESSION

An efficient nonparametric regression* estimate may be constructed by using the multivariate ASH. For example, with data (x, y), construct the bivariate ASH and then compute the conditional mean of the estimate. The result is particularly simple as $\delta_y \to 0$:

$$\hat{m}(x) = \frac{\sum_{i=1-m_1}^{m_1-1} w_{m_1}(i) \nu_{k+i} \bar{y}_{k+i}}{\sum_{i=1-m_1}^{m_1-1} w_{m_1}(i) \nu_{k+i}}, \quad x \in B_k,$$

where $\bar{y}_k$ is the average of the y_i's for those points (x_i, y_i), where $x_i \in B_k$ and ν_k is the univariate bin count. This estimator was introduced by Härdle and Scott [2]; two- and three-dimensional versions with application to spatial estimation from sample surveys of agriculture data are given in Whittaker and Scott [8].

EXAMPLES

Bradford Brown (see ref. [4], Appendix B.4) measured the thickness of 90 U.S. pennies to the nearest tenth of a mil. Two shifted histograms plus an ASH are shown in Fig. 1, all with smoothing parameter $h = 1.5$ mils. The visual impression of the two histograms is markedly different. The ASH essentially removes the effect of the choice of bin origin, which is a nuisance parameter. With a smaller smoothing parameter $(h = 0.9)$, two extra modes appear (rather than only one) at 55.9 and 57.1 mils.

A fuller view of these data may be obtained by plotting them as a time series as in Fig. 2. Contours of the bivariate ASH, together with the regression ASH, are also displayed. Apparently, the thickness of pennies has changed more than once since World War II.

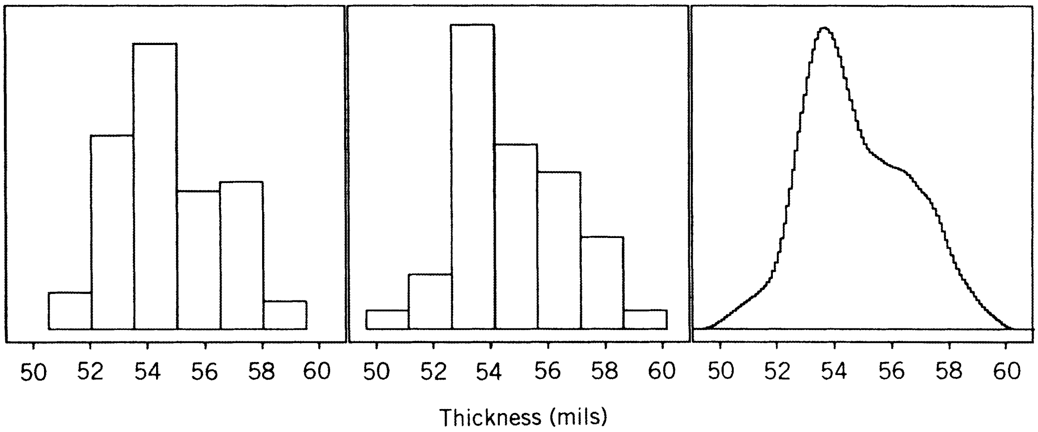


Figure 1. Two shifted histograms and the ASH of the penny thickness data. The bin origins for the histograms were 50.55 and 49.65. The triweight kernel was used for the ASH with $\delta = 0.1$ and $m = 1$.

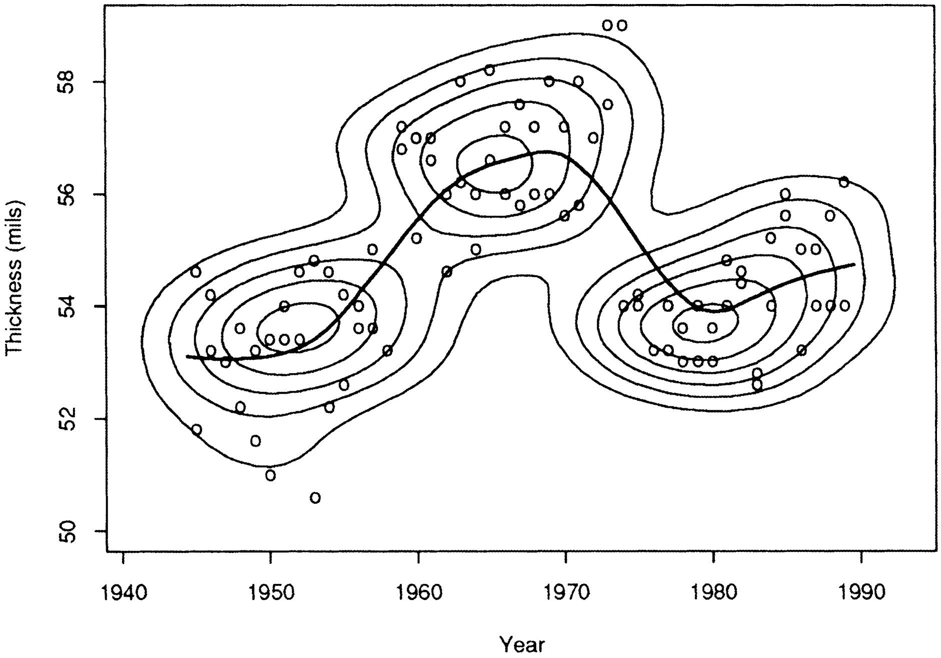


Figure 2. Bivariate ASH density and regression estimates of the penny data.

REFERENCES

1. Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comput. Graph. Statist.*, **3**, 35–56.
2. Härdle, W. and Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Comput. Statist.*, **7**, 97–128.
3. Scott, D. W. (1985). Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *Ann. Statist.*, **13**, 1024–1040.
4. Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley, New York.
5. Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density

estimation. *J. Amer. Statist. Ass.*, **82**, 1131–1146.

6. Silverman, B. W. (1982). Kernel density estimation using the fast Fourier transform, *J. R. Statist. Soc. B*, **31**, 93–97.
7. Terrell, G. R. and Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *J. Amer. Statist. Ass.*, **80**, 209–214.
8. Whittaker, G. and Scott, D. W. (1994). Spatial estimation and presentation of regression surfaces in several variables via the averaged shifted histogram. *Comput. Sci. Statist.*, **26**, 8–17.

See also Graphical Representation of Data; Histograms; and Kernel Estimators.

David W. Scott

AXIAL DISTRIBUTIONS. See Directional Distributions

AXIOMS OF PROBABILITY

THE AXIOMATIC METHOD

It is perhaps no coincidence that the axiomatic method in mathematics became prominent somewhat before the "use" theory of meaning became prominent in philosophy. An axiomatic system aims to capture and formalize some way of using language and it sidesteps the difficulties of explicit definitions.

The advantage of the axiomatic method is that theorems can in principle be deduced mathematically from the axioms without new assumptions creeping in surreptitiously, and without necessary philosophical commitment. In other words, the theorems can be proved rigorously according to the usual standards of pure mathematics without involvement in the controversial problems of application to the real world. In practice it is difficult to be totally rigorous, as has been found by philosophers of mathematics. An early example of the axiomatic method is in the geometry of Euclid, although his axioms do not satisfy most modern pure mathematicians. The value of a more precise axiomatic approach was emphasized by David Hilbert near the beginning of the twentieth century. The approach has become a paradigm for pure mathematics, but less so for applied mathematics and for physics because it can lead to rigor mortis (to quote Henry Margenau's joke). Probability theory is both pure and applied*, so that different specialities put more or less emphasis on axiomatic systems.

NOTATION

Many theories of probability have been proposed, and many different notations have been used. In this article we use notations such as $P(E|F)$, which can be read as the probability of E given (or assuming, or conditional on) F. Here, depending on the theory or the context, E and F might denote propositions, events*, hypotheses*, scientific theories, or sets, or might even be abstract symbols, such as those in abstract algebra, without ordinary definitions but subject only to some axioms. We can regard $P(\cdot|\cdot)$ as a function of two variables, and the domains of E and F are not necessarily identical. The notation $P(E)$ is read "the probability of E" and is used either when F is taken for granted or, in some theories, not as a conditional probability* but as a so-called "absolute probability" in which officially nothing is "given" or "assumed" other than logic and mathematics (were that possible).

When a theory of probability is expressed axiomatically there will usually be axioms satisfied by such symbols as E and F and further axioms satisfied by the "probabilities" themselves. Some theories of probability are formulated as theories of rationality, and then the set of axioms needs to mention either decisions or "utilities" (="desirabilities"). *See* Decision Theory.

ARGUMENTS OF $P(E|F)$

A theory in which E and F denote sets or are abstract symbols can be regarded as a branch of pure mathematics, but propositions, events, and hypotheses are not purely mathematical concepts when they are interpreted as ordinary English words. We shall not try to define the English meanings of events, hypotheses, and theories, but the meaning of "proposition" is especially controversial: see, for example, Gale [11]. Perhaps the best definition is that a proposition is "the meaning of a clear statement." By ruling out unclear statements we are adequately justified in assuming that each proposition is capable of being either true or false, although really there are degrees of meaningfulness because statements can be more or less vague.

Moreover, a statement can be either empirically or mathematically meaningful, a point that is relevant to the choice of axioms. For example, to say that a measurement of a continuous variable lies between 5.25 and 5.35 inches is often empirically meaningful, whereas to say that it is exactly 5.30 inches, with no error at all, is at best mathematically meaningful within an idealized mathematical model. Again, to say that "the limit of the proportion of time that a coin comes up heads is approximately 0.5 in an infinite sequence of tosses" can be fully meaningful only within pure mathematics, because all sequences of tosses in the real world are of finite length.

AXIOMS FOR PROPOSITIONS AND SETS

The *conjunction* of propositions E and F is the proposition E & F and is denoted in this article by EF. The *disjunction* of E and F is denoted by $E \vee F$. This proposition asserts that E or F or both are true. The negation of E is denoted by $\tilde{E}$ or by $\sim E$. If E and F denote the same proposition, then we write $E = F$. (Other notations are in use.)

Some axioms for propositions are:

A1 If E is a proposition, then $\tilde{E}$ is also. This axiom might not be accepted by those who define a proposition as (scientifically) meaningful only if it is refutable if false. This first axiom, if applied to scientific propositions, forces us to the view that a proposition is also scientifically meaningful when it is confirmable if true. There are, however, degrees in these matters: see Good [15, pp. 492–494].

A2 $\sim (\sim E) = E$.

A3 If E and F are both propositions, then so is EF.

A4 *Commutative law*. $EF = FE$.

A5 *Associative law*. $E(FG) = (EF)G$.

A6 *De Morgan's law*. $\sim (EF) = \tilde{E} \vee \tilde{F}$. From this we can prove that the commutative and associative laws apply also to disjunctions.

A7 *Distributive laws*

$$E(F \vee G) = (EF) \vee (EG)$$

and

$$E \vee (FG) = (E \vee F)(E \vee G),$$

of which the second law can be inferred from the first by means of de Morgan's law*.

To these seven axioms, which are essentially the axioms of Boolean algebra, we can appendthe optional axiom (A8) and perhaps (A9):

A8 The conjunction and disjunction of a countably infinite number of propositions are propositions, with a corresponding "de Morgan law,"

$$\sim (E_1E_2E_3 \cdots) = \tilde{E}_1 \vee \tilde{E}_2 \vee \tilde{E}_3 \vee \cdots .$$

A9 The conjunction and disjunction of any infinite number of propositions are propositions, with yet another "de Morgan law": The negation of the conjunction is the disjunction of the negations.[1]

ORIGINS OF THE AXIOMS

In most theories of probability the probabilities lie in some sense between 0 and 1 and satisfy axioms somewhat resembling in appearance the addition and product axioms, namely:

$$P(A \vee B) = P(A) + P(B)$$

when A and B are mutually exclusive* and

$$P(AB) = P(A) \cdot P(B|A).$$

These comments will be clarified in what follows.

The addition and product axioms were known at least implicitly to Fermat* and Pascal* in 1654 and perhaps to Cardano* in the sixteenth century. But it is more convenient to express the axioms explicitly, formally, and completely.

Axioms are seldom slapped down arbitrarily; a system of axioms should be chosen either as a convenient form of other axioms or should be constructed to capture some intuitive ideas about the world or about mathematics, so that the system has some prior justification. The axioms can also be justified by their practical and philosophical implications; e.g., they should not be seen to lead to an irresolvable contradiction. Before the axioms can have practical meaning, some formal rules of application to the real world must be provided. Moreover, in practice a set of axioms and rules of application are still not sufficient: a theory needs to become to some extent a technique if it is to be useful. One needs informal suggestions of how to apply the theory, although these suggestions are not logically essential. In this article no more will be said about such practical suggestions because such matters belong properly to a discussion of the relationship between probability and statistics or between probability and practical decision making. *See* DECISION THEORY.

The prior justification of a set of axioms must depend on some concept of probability, however vague. One of the earliest concepts of probability was derived from games of chance*, such as those depending on coin spinning, dice throwing, and card drawing. In such games there are some symmetry properties that suggest that some outcomes are at least approximately equally probable, and this is so even for people who have not much idea of what probability means. In the context of such games one might be ready to accept Laplace's* definition of the probability of an event E as k/m, where m is the number of "equally possible" cases (meaning equally probable cases) that could occur, and k is the number of those cases that constitute E. When this "definition" is applicable, it leads to a familiar set of axioms. The main disadvantage of this approach is that a clearly exhaustive set of "equally probable cases" cannot usually be specified with reasonable objectivity in scientific applications. Also, the definition is somewhat circular.

In many experiments or observational circumstances the kind of symmetry required for the direct application of the classical definition of probability is lacking. To get around this difficulty, or for other reasons, a definition in terms of long-run proportional frequency of "successes" was explicitly proposed by Leslie Ellis and Cournot* in 1843, and developed in much detail by Venn in 1866. (For these references and further history, see Keynes [21, pp. 92–93]). As usual with simple ideas, "frequentism" had been to some extent foreshadowed long before, e.g., by Aristotle, who said that the probable is what usually happens, or Greek words to that effect, but a self-respecting kudologist would not on that account attribute the theory to Aristotle alone. The frequentist definition is associated with physical probability rather than with logical or subjective (= personal) probability. For discussions of kinds of probability, *see* BELIEF, DEGREES OF and its references, and PROBABILITY, FOUNDATIONS OF—I.

It is by no means simple to construct a satisfactory definition of physical probability based on limiting frequencies. Consider, e.g., the following naive approach. By a "trial" we mean an experiment whose outcome is either some event E, or is the negation of E, a "failure" F. For example, a trial might be the tossing of a coin or the throw of a die and E might denote "heads" or "a six." Let an infinite sequence of such trials be performed under "essentially equivalent" conditions. Then the proportion of successes in the first n trials might tend to a limit p when $n \to \infty$. If so, then p might be called the probability of a success.

This naive definition of physical probability by long-run or limiting frequency* has some disadvantages. Even if we admit the possibility of an infinite sequence of trials in the real world, as some kind of approximation, the definition says nothing about whether the sequence of outcomes is in any sense random. A more sophisticated long-run-frequency definition of probability was proposed by von Mises [24] based on the prior notion of a random sequence or irregular Kollektiv*. This approach requires axioms for random sequences and again has severe mathematical and logical difficulties, although it can be presented fairly convincingly to the intuition in terms of generalized decimals [17, and references therein]. The theory of von Mises can be made logically rigorous and then leads to a familiar set of axioms for probability [6,23]. For a discussion of randomness, with further references, see also Coffa et al. [5] and the article on RANDOMNESS, BDS TEST FOR in the present encyclopedia.

An approach to the axioms via sharp absolute probabilities, when there are fewer than 26^M possible mutually exclusive propositions, where $M = 10^{1000}$, is to argue as follows. Suppose that the N mutually exclusive possible propositions $E_1, E_2, \ldots, E_N$ of interest have sharp probabilities approximated to ν places of decimals by $p_1, p_2, \ldots, p_N$, where ν is large, so that $p_i = m_i 10^{-\nu}$, where m_i is a positive integer, and $\sum m_i = 10^\nu$. For each i, take a well-shuffled pack of cards containing m_i equiprobable cards and use it to break E_i into m_i mutually exclusive propositions each of probability $10^{-\nu}$. This leads to 10^ν equally probable propositions and the classical definition can now be used to arrive at a familiar set of axioms. (Compare ref. 12, p. 33, where the argument was expressed somewhat differently.)

An approach that again assumes that probabilities mean something and can be expressed numerically was apparently first suggested by S. N. Bernstein* [3]. It depends on ideas such as that $P((E \vee F) \vee G)$ must equal $P(E \vee (F \vee G))$. It is assumed further that, when E and F are mutually exclusive, then $P(E \vee F)$ is some function of $P(E)$ and $P(F)$. The assumptions lead to functional equations that must be satisfied by probabilities, and these equations can be used to justify the axioms. The idea was developed independently by Schrödinger [32], Barnard and Good [12, pp. 107–108] and especially by R. T. Cox [7,8]. Cox's assumptions were weakened by Aczél [1]. The approach again leads to a familiar set of axioms, and seems to be the most convincing justification of these axioms for numerical subjective probability among those approaches that makes no reference to decisions or to gambles.

An advantage of bringing decisions or gambles into the discussion is that a prior intuitive concept of probability is then less necessary in arriving at axioms for subjective probabilities. We are then led to a behavioral approach that many people find more convincing than a more purely linguistic approach. With some ingenuity the behavioral approach can be developed to the point where no explicit definition of either probability or utility* is assumed, but only preferences between acts. This approach was adopted by F. P. Ramsey [27], B. de Finetti [9], and L. J. Savage [31]. They assumed that "your" preferences between acts can be completely ordered, and that the preferences satisfy desiderata for rationality that many people find compelling once the complete ordering is granted. These desiderata lead to the conclusion that if you were perfectly rational, you would behave as if you had a set of probabilities (degrees of belief*) satisfying familiar axioms, and a set of utilities, and that you would always prefer the act of maximum expected utility. In this approach the concepts of probability and utility are not separately defined, nor are they taken for granted. In fact, a perfectly rational person might not know the concepts of probability and utility, but these concepts can be used by some one else to describe the rational person. We can *imagine* a doctor or warrior, for example, who always makes the best decisions, although never having heard of probabilities.

THE PURELY MATHEMATICAL APPROACH

Since most of the philosophical approaches lead to somewhat similar formal theories, it is natural for a mathematician to choose a set of axioms based on earlier formalisms. By separating the symbols E, F, etc., from their concrete meanings, the mathematician can avoid philosophical controversies and get on with the job. This approach was adopted by A. N. Kolmogorov* [22], following some earlier writers. His axioms were expressed in the language of sets and measure theory. Borel–Lebesgue measure was introduced at the turn of the century: see, e.g., Carathéodory [4, p. 702]. Before 1890 set theory was not regarded as mathematically respectable, but by 1930 it was regarded as part of the foundation of pure mathematics.

Kolmogorov stated that he agreed with von Mises's frequency interpretation* of probability, but his axioms do not presuppose this interpretation. To begin with, he assumes that there is a set Ω of "elementary events" ω, but the concept of an elementary event requires no definition as long as we are concerned only with the mathematics, and each ω can instead be called a "point" if this helps the imagination. A class $\mathscr{S}$ of subsets of Ω are called "events," not yet to be interpreted in the ordinary sense; and it is assumed that if the subset S is an element of $\mathscr{S}$, then so is its complement $\Omega - S$. Furthermore, it is assumed that if S and T both belong to Ω, then so does the union of S and T. "Events" satisfying these assumptions are said to constitute an *algebra of events** or field of events. Note the similarity to the axioms for propositions given earlier. If the union of any countable infinity of events is also in $\mathscr{S}$, then $\mathscr{S}$ is unfortunately said to be a σ-algebra. We shall soon see why the condition of countability is assumed.

A symbol $P(S)$ is introduced and is called the (absolute) probability of S. It is assumed that $P(S)$ is a real number and lies in the closed interval [0,1], also that $\Omega \in \mathscr{S}$ and that $P(\Omega) = 1$. Finally, if a countable class of sets $S_1, S_2, S_3, \ldots$ are disjoint, then $P(S_1 \cup S_2 \cup \cdots) = P(S_1) + P(S_2) + \cdots$, where $\cup$ denotes union. This last assumption is called the axiom of complete additivity. A weaker axiom asserts the property only for two sets (which implies the property for any finite number of sets) instead of for a countable infinity of sets. The axiom of complete additivity is the main feature of Kolmogorov's system and makes his system highly reminiscent of Lebesgue measure.

The product axiom is introduced through the back door by defining the conditional probability $P(S|T)$ by the quotient $P(S \cap T)/P(T)$ when $P(T) \neq 0$, where $\cap$ denotes the intersection of sets.

The theory is applied by interpreting "events" as meaning physical events.

Kolmogorov's axioms, perhaps better called the measure-theoretic axioms, are the most popular among mathematical statisticians at present. He did not pretend that he had no predecessors, and Rényi [29, p. 55] cites Borel (1909), Lomnicki (1923), Lévy (1925), Steinhaus (1923), and Jordan (1925). As Rényi says, the measure-theoretic approach leads to a rigorous mathematical theory of stochastic processes.

The analogy with Lebesgue measure makes it clear why the axiom of complete additivity is stated only for a countable infinity of sets: the Lebesgue measure of a unit interval is unity, but this measure can hardly be expressed as the sum of the noncountable number of zero measures of the points in the interval. A similar objection has been raised by de Finetti [10, p. 124] against the axiom of complete additivity itself. For consider an infinite sequence in which it is known that there is precisely one "success" but where this success might be anywhere. In the von Mises theory the probability of a success would apparently be zero, and the axiom of complete additivity appears then to lead to the contradiction $1 = \sum 0 = 0$. Perhaps the resolution of this difficulty is to deny that the foregoing sequences count as Kollektivs and say that one models zero probability by a Kollektiv in which the limiting proportional frequency of "successes" is zero. Then there are a noncountable number of such Kollektivs and no paradox arises. (How you could recognize such Kollektivs in practice from finite initial segments is another problem.)

As indicated earlier, from a strictly practical point of view it makes no sense to choose a precise real number at random; in fact, to do so would be like selecting the infinite sequence of its decimal digits. When E and F denote meaningful practical propositions the measure-theoretic approach is not essential, but the approach compensates for this by its valuable mathematical convenience. Then again a price is paid because the mathematics becomes advanced.

The approach in terms of propositions appears more general than in terms of sets, because there need be no concept of an "elementary proposition." It would, however, be possible, at least if the total number of propositions if finite, to define an elementary proposition as a conjunction of propositions that is (a) not strictly impossible while (b) if it can be made any less probable by being conjoined with another proposition, then it becomes impossible. An elementary proposition would be an interpretation of an elementary event

ω. In this manner one might be able to subsume the propositional approach under the measure-theory umbrella, but if there were only a finite number of propositions, the axiom of complete additivity would be unnecessary.

AXIOMS EXPRESSED IN TERMS OF CONDITIONAL PROBABILITIES

Since probabilities in practice are always conditional, absolute probabilities do not capture anything concrete, so several writers have proposed sets of axioms stated directly in terms of conditional probabilities; e.g., Wrinch and Jeffreys [33], Keynes [21], Reichenbach [28], Good [12], Popper [26], and Rényi [29]. Some of these writers expressed the axioms in terms of the probabilities of propositions. We give here an example of such a system of axioms based on Good [12,13]. In these axioms, the symbol H does not necessarily denote a hypothesis, but in many applications it does.

- **A1** $P(E|F)$ *is a real number if* F *is not self-contradictory*. (A similar caveat applies in the remaining axioms.)
- **A2** $0 \leqslant P(E|F) \leqslant 1$.
- **A3** *If* $P(EF|H) = 0$, *then* $P(E \vee F|H) = P(E|H) + P(F|H)$.
- **A4** *If* H *logically implies* E (i.e., if $\overline{H} \vee E$ is a tautology), *then* $P(E|H) = 1$ (but not conversely).
- **A5** (Axiom of equivalence.) *If neither HE nor HF is self-contradictory and HE implies F and HF implies E, then* $P(E|H) = P(F|H)$.
- **A6** $P(EF|H) = P(E|H)P(F|EH)$.
- **A7** $P(H^*|H^*) \neq 0$, *where* H^* *denotes the basic assumptions of logic and pure mathematics*.
- **A8** $P(E^*|H^*) = 0$ *for some proposition* E^*.
- **A9** (Complete additivity: optional.) *If* $P(E_iE_j|H) = 0 (i < j; i,j = 12, \ldots)$, *then* $P(E_1 \vee E_2 \vee \cdots |H) = \sum P(E_i|H)$.
- **A10** (The principle of cogent reason: optional: see Keynes [21, p. 56], Russell [30, p. 397], Good [12, p. 37].) *Let* ϕ *and* ψ *be propositional functions. Then, for all a and b for which the functions are defined, we have*

$$P(\phi(a)|\psi(a)) = P(\phi(b)|\psi(b)).$$

For example, the probability of getting 7 hearts in a whist hand, given only that the pack contains 13 hearts, is the same if we change "hearts" to "diamonds."

In this theory, the main rule of application is obtained by thinking of the axioms as the workings of a black box* into which judgments of probability inequalities can be plugged and from which discernments of new inequalities can be read. This black-box theory is explained in more detail in the article BELIEF, DEGREES OF. Following Keynes [21], who, however, dealt with logical probabilities, Good assumes that (subjective) probabilities are only partially ordered and the use of axioms for sharp probabilities is only a device for expressing the theory in a highly intelligible form. From this form of the theory it is shown by Good [14] that one can derive axioms for the upper* and lower* probabilities themselves. For example, the product axiom splits into six axioms, one of which is

$$P_*(EF|H) \leqslant P^*(E|H) \cdot P_*(F|EH).$$

One can think of upper and lower probabilities as exterior and interior measures, and in a frequency theory they might correspond to upper and lower limits.

If one wishes to talk meaningfully about the probability of a mathematical theorem, as is desirable for the formalizing of "plausible reasoning" (see, e.g., Pólya [25]), then it is necessary as in Good [12, p. 49] to replace the axiom of equivalence by something like A5′. *If at time t you have seen that E and F are equivalent, then* $P_t(E|H) = P_t(F|H)$ *and* $P_t(H|E) = P_t(H|F)$, *where the subscript t is self-explanatory*. [Judgments are needed to decide whether $P_t(G|K) = P_s(G|K)$, where $t \neq s$.] This axiom allows subjective probabilities to vary as time passes, without changing ordinary empirical evidence. This is not, of course, the same as the elementary fact that $P(E|GH)$ is not in general equal to $P(E|H)$. By allowing probabilities to have the dynamic feature of varying as a consequence of calculations and thinking, such as the variables in FORTRAN, one can say meaningfully and quantitatively that a mathematical theorem conveys information, and one can also solve some otherwise intractable philosophical problems concerning scientific induction. (For example, see Good [16,18,19], where these varying probabilities are called "evolving" or "dynamic.")

The theory for partially ordered or comparative probability, as just discussed, extends immediately to a theory of rational behavior, by introducing utilities. For details, see Good [13].

A difficulty in theories of subjective probability, pointed out by Richard Jeffrey [20, p. 154], is that a subjective probability can change as a consequence of an experience that "you" cannot express in words. As a matter of fact, badly remembered experiences cause the same difficulty. Although you might have experiences that you cannot personally express fully in words, you can describe them as experiences that occurred at a certain time, and the meaning of this description can be regarded as a proposition in an extended sense. By allowing the meaning of "proposition" to be extended in this manner, the difficulty seems to be overcome without the need for any new axioms. The difficulty would not be overcome within a theory of logical rather than subjective probability.

A distinctive feature of Jeffrey [20, p. 83] is connected with utilities. Previous theories had led to the conclusion that if preference rankings* are sufficiently extensive, probabilities can be uniquely determined but utilities can be determined only up to linear transformations; that is, if a set of (expected) utilities is given, then each element u of this set can be replaced by $au + b$, where a and b are constants, and this substitution will have no effects on any recommended decisions. In Jeffrey's theory, in which both probabilities and utilities refer to propositions, the probabilities and utilities can undergo a class of transformations, the transformation of the utility being of the form $(au + b)/(cu + d)$. He attributes this result to independent personal communications from Kurt Gödel and Ethan Bolker.

In summary, distinct purposes and distinct philosophies can be associated with distinct systems of axioms of probability, although these systems fortunately have much in common.

NOTE

1. All these axioms and comments are applicable *mutatis mutandis* to sets as well as to propositions. For propositional functions, with "quantifiers" such as "for all" and "there exists," further axioms are necessary, but we shall not labor this point.

REFERENCES

1. Aczél, J. (1963). *Ann. Univ. Sci. Budap. Rolando Eötvös Nominatae Sect. Math.*, **6**, 3–11.
2. Barnard, G. A. (1949). *J. R. Statist. Soc. B*, **11**, 115–139.
3. Bernstein, S. N. (1917). An attempt at an axiomatic foundation for the calculus of probability (in Russian). *Khar'kov Univ. Kar'kovskoi mat. obshch. Soobshcheniia*, **15**, 209–274. Abstract in German by Bernstein in *Jb. Math.*, **48**, (1920–1921), 596–599.
4. Carathéodory, C. (1972). *Vorlesungen über Reelle Funktionen*, 2nd ed. Teubner, Leipzig. (Reprint: Chelsea, New York, 1948.)
5. Coffa, J. A., Good, I. J., and Kyburg, H. E. (1974). *PSA 1972* (Proc. 1972 Bienn. Meet. Philos. Sci. Ass.), K. F. Schaffner and R. S. Cohen, eds. D. Reidel, Dordrecht, pp. 103–149.
6. Copeland, A. H. (1937). *Trans. Amer. Math. Soc.*, **42**, 333–357.
7. Cox, R. T. (1946). *Amer. J. Phys.*, **14**, 1–13.
8. Cox, R. T. (1961). *The Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore, Md.
9. Finetti, B. de (1937). *Ann. Inst. Henri Poincaré*, **7**, 1–68. English translation in *Studies in Subjective Probability*, H. E. Kyburg and H. E. Smokler, eds. Wiley, New York, 1964, pp. 95–158.
10. Finetti, B. de (1974). *Theory of Probability*, Vol. 1. Wiley, New York.
11. Gale, R. M. (1967). In *The Encyclopedia of Philosophy*, Vol. 5, Paul Edwards, ed. Macmillan/The Free Press, New York, pp. 494–505.

12. Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London/Hafner, New York.
13. Good, I. J. (1952). *J. R. Statist. Soc. B*, **14**, 107–114.
14. Good, I. J. (1962). In *Logic, Methodology, and Philosophy of Science*, E. Nagel, P. Suppes, and A. Tarski, eds. Stanford University Press, Stanford, Calif., pp. 319–329.
15. Good, I. J. (1962/1966). In *Theories of the Mind*, J. Scher, ed. Glencoe Free Press/Macmillan, New York, pp. 490–518. Misprints corrected in second edition.
16. Good, I. J. (1968). *Brit. J. Philos. Sci.*, **19**, 123–143.
17. Good, I. J. (1974). In the symposium cited in ref. 5, pp. 117–135.
18. Good, I. J. (1975). *Synthése*, **30**, 39–73.
19. Good, I. J. (1977). In *Machine Intelligence*, Vol. 8, E. W. Elcock and D. Michie, eds. Wiley, New York, pp. 139–150.
20. Jeffrey, R. (1965). *The Logic of Decision*. McGraw-Hill, New York.
21. Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan, London (2nd ed., 1929).
22. Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin (English translation: Chelsea, New York, 1950).
23. Martin-Löf, P. (1969). *Theoria*, **35**, 12–37.
24. Mises, R. von (1919). *Math. Zeit.*, **5**, 52–99.
25. Pólya, G. (1954). *Mathematics and Plausible Reasoning*, 2 vols. Princeton University Press, Princeton, N. J.
26. Popper, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson, London.
27. Ramsey, F. P. (1926/1931). *The Foundations of Mathematics and Other Logical Essays*. Kegan Paul, London; Harcourt Brace, New York.
28. Reichenbach, H. (1949). *The Theory of Probability*. University of California Press, Berkeley, Calif.
29. Rényi, A. (1970). *Foundations of Probability*. Holden-Day, San Francisco.
30. Russell, B. (1948). *Human Knowledge, Its Scope and Limitations*. Routledge & Kegan Paul, London.
31. Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York. (2nd ed., Dover, New York).
32. Schrödinger, E. (1947). *Proc. R. Irish Acad.*, **51A**, 51–66, 141–146.
33. Wrinch, D. and Jeffreys, H. (1919). *Philos. Mag., 6th Ser.*, **38**, 715–731.

See also Belief, Degrees of; Chance—II; Decision Theory; Probability, Foundations of—I; Probability, History of; and Probability Theory: An Outline.

I. J. Good

AZZALINI'S FAMILY.

See Skew-Normal Family of Distributions