CHAPTER 1

Preliminary Information

Introduction

This work contains descriptions of many different distributions used in statistical theory and applications, each with its own pecularities distinguishing it from others. The book is intended primarily for reference. We have included a large number of formulas and results. Also we have tried to give adequate bibliographical notes and references to enable interested readers to pursue topics in greater depth.

The same general ideas will be used repeatedly, so it is convenient to collect the appropriate definitions and methods in one place. This chapter does just that. The collection serves the additional purpose of allowing us to explain the sense in which we use various terms throughout the work. Only those properties likely to be useful in the discussion of statistical distributions are described. Definitions of exponential, logarithmic, trigonometric, and hyperbolic functions are not given. Except where stated otherwise, we are using real (not complex) variables, and "log," like "ln," means natural logarithm (i.e., to base e).

A further feature of this chapter is material relating to formulas that will be used only occasionally; where appropriate, comparisons are made with other notations used elsewhere in the literature. In subsequent chapters the reader should refer back to this chapter when an unfamiliar and apparently undefined symbol is encountered.

1.1 MATHEMATICAL PRELIMINARIES

1.1.1 Factorial and Combinatorial Conventions

The number of different orderings of n elements is the product of n with all the positive integers less than n; it is denoted by the familiar symbol $n!$ (*factorial* n),

$$n! = n(n-1)(n-2)\cdots 1 = \prod_{j=0}^{n-1}(n-j). \tag{1.1}$$

Univariate Discrete Distributions, Third Edition.
By Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz

The less familiar semifactorial symbol $k!!$ means

$$(2n)!! = 2n(2n-2)\cdots 2,$$

where $k = 2n$.

The product of a positive integer with the next $k-1$ smaller positive integers is called a *descending (falling) factorial*; it will in places be denoted by

$$\begin{aligned} n^{(k)} &= n(n-1)\cdots(n-k+1) \\ &= \prod_{j=0}^{k-1}(n-j) = \frac{n!}{(n-k)!}, \end{aligned} \tag{1.2}$$

in accordance with earlier editions of this book. Note that there are k terms in the product and that $n^{(k)} = 0$ for $k > n$, where n is a positive integer. Readers are WARNED that there is no universal notation for descending factorials in the statistical literature. For example, Mood, Graybill, and Boes (1974) use the symbol $(n)_k$ in the sense $(n)_k = n(n-1)\cdots(n-k+1)$, while Stuart and Ord (1987) write $n^{[k]} = n(n-1)\cdots(n-k+1)$; Wimmer and Altmann (1999) use $x_{(n)} = x(x-1)(x-2)\cdots(x-n+1)$, $x \in \mathbb{R}$, $n \in \mathbb{N}$.

Similarly there is more than one notation in the statistical literature for *ascending (rising) factorials*; for instance, Wimmer and Altmann (1999) use $x^{(n)} = x(x+1)(x+2)\cdots(x+n-1)$, $x \in \mathbb{R}$, $n \in \mathbb{N}$. In the first edition of this book we used

$$\begin{aligned} n^{[k]} &= n(n+1)\cdots(n+k-1) \\ &= \prod_{j=0}^{k-1}(n+j) \ = \frac{(n+k-1)!}{(n-1)!}. \end{aligned} \tag{1.3}$$

There is, however, a standard notation in the mathematical literature, where the symbol $(n)_k$ is known as *Pochhammer's symbol* after the German mathematician L. A. Pochhammer [1841–1920]; it is used to denote

$$(n)_k = n(n+1)\cdots(n+k-1) \tag{1.4}$$

[this definition of $(n)_k$ differs from that of Mood et al. (1974)]. We will use Pochhammer's symbol, meaning (1.4) except where it conflicts with the use of (1.3) in earlier editions.

The *binomial coefficient* $\binom{n}{r}$ denotes the number of different possible combinations of r items from n different items. We have

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{n}{n-r}; \tag{1.5}$$

also

$$\binom{n}{0} = \binom{n}{n} = 1 \qquad \text{and} \qquad \binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}. \tag{1.6}$$

It is usual to define $\binom{n}{r} = 0$ if $r < 0$ or $r > n$. However,

$$\begin{aligned}\binom{-n}{r} &= \frac{(-n)(-n-1)\cdots(-n-r+1)}{r!} \\ &= (-1)^r \binom{n+r-1}{r}.\end{aligned} \tag{1.7}$$

The *binomial theorem* for a positive integer power n is

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^{n-j} b^j. \tag{1.8}$$

Putting $a = b = 1$ gives

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$$

and putting $a = 1$, $b = -1$ gives

$$\binom{n}{0} - \binom{n}{1} + \cdots + (-1)^n \binom{n}{n} = 0.$$

More generally, for any real power k

$$(1+b)^k = \sum_{j=0}^{\infty} \binom{k}{j} a^j, \qquad -1 < b < 1. \tag{1.9}$$

By equating coefficients of x in $(1+x)^{a+b} = (1+x)^a (1+x)^b$, we obtain the well-known and useful identity known as *Vandermonde's theorem* (A. T. Vandermonde [1735–1796]):

$$\binom{a+b}{n} = \sum_{j=0}^{n} \binom{a}{j} \binom{b}{n-j}. \tag{1.10}$$

Hence

$$\binom{2n}{n} = \binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2.$$

The *multinomial coefficient* is

$$\binom{n}{r_1, r_2, \ldots, r_k} = \frac{n!}{r_1! r_2! \cdots r_k!}, \tag{1.11}$$

where $r_1 + r_2 + \cdots + r_k = n$.

The *multinomial theorem* is a generalization of the binomial theorem:

$$\left(\sum_{j=1}^{k} a_j\right)^n = \sum \left(\frac{n! \prod_{i=1}^{k} a_i^{n_i}}{\prod_{i=1}^{k} n_i!}\right), \tag{1.12}$$

where summation is over all sets of nonnegative integers $n_1, n_2, \ldots, n_k$ that sum to n.

There are four ways in which a sample of k elements can be selected from a set of n distinguishable elements:

Order Important?	Repetitions Allowed?	Name of Sample	Number of Ways to Select Sample
No	No	k-Combination	$C(n, k)$
Yes	No	k-Permutation	$P(n, k)$
No	Yes	k-Combination with replacement	$C^R(n, k)$
Yes	Yes	k-Permutation with replacement	$P^R(n, k)$

where

$$\begin{aligned} C(n, k) &= \frac{n!}{k!(n-k)!}, & P(n, k) &= \frac{n!}{(n-k)!}, \\ C^R(n, k) &= \frac{(n+k-1)!}{k!(n-1)!}, & P^R(n, k) &= n^k. \end{aligned} \tag{1.13}$$

The number of ways to arrange n distinguishable items in a row is $P(n, n) = n!$ (the number of permutations of n items).

The number of ways to arrange n items in a row, assuming that there are k types of items with n_i nondistinguishable items of type i, $i = 1, 2, \ldots, k$, is the multinomial coefficient $\binom{n}{n_1, n_2, \ldots, n_k}$.

The number of derangements of n items (permutations of n items in which item i is not in the ith position) is

$$D_n = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots (-1)^n \frac{1}{n!}\right).$$

The signum function, $\operatorname{sgn}(\cdot)$, shows whether an argument is greater or less than zero:

$$\operatorname{sgn}(x) = 1 \quad \text{when} \quad x > 0; \quad \operatorname{sgn}(0) = 0; \quad \operatorname{sgn}(x) = -1 \quad \text{when} \quad x < 0.$$

The ceiling function, $\lceil x \rceil$, is the least integer that is not smaller than x, for example,

$$\lceil e \rceil = 3, \qquad \lceil 7 \rceil = 7, \qquad \lceil -2.4 \rceil = -2.$$

The floor function, $\lfloor x \rfloor$, is the greatest integer that is not greater than x, for example,

$$\lfloor e \rfloor = 2, \qquad \lfloor 7 \rfloor = 7, \qquad \lfloor -2.4 \rfloor = -3.$$

The notation $[\cdot] = \lfloor \cdot \rfloor$ is called the integer part.

$$\pi = 4 \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} = 3.1415926536,$$

$$e = \sum_{j=0}^{\infty} \frac{1}{j!} = 2.7182818285,$$

$$\ln 2 = \sum_{j=0}^{\infty} \frac{(-1)^{j-1}}{j} = 0.6931471806.$$

1.1.2 Gamma and Beta Functions

When n is real but is *not* a positive integer, meaning can be given to $n!$, and hence to (1.2), (1.3), (1.5), (1.7), and (1.11), by defining

$$(n-1)! = \Gamma(n), \qquad n \in \mathbb{R}^+, \tag{1.14}$$

where $\Gamma(n)$ is the *gamma function*.

The binomial theorem can thereby be shown to hold for any real power.

There are three equivalent definitions of the gamma function, due to L. Euler [1707–1783], C. F. Gauss [1777–1855], and K. Weierstrass [1815–1897], respectively:

Definition 1 *(Euler)*:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt, \qquad x > 0. \tag{1.15}$$

Definition 2 *(Gauss)*:

$$\Gamma(x) = \lim_{n\to\infty}\left[\frac{n!n^x}{x(x+1)\cdots(x+n)}\right], \qquad x \neq 0, -1, -2, \ldots. \tag{1.16}$$

Definition 3 *(Weierstrass)*:

$$\frac{1}{\Gamma(x)} = xe^{\gamma x}\prod_{n=1}^{\infty}\left[\left(1+\frac{x}{n}\right)\exp\left(-\frac{x}{n}\right)\right], \qquad x > 0, \tag{1.17}$$

where γ is *Euler's constant*

$$\gamma = \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\ln n\right) \cong 0.5772156649\ldots. \tag{1.18}$$

From Definition 1, $\Gamma(1) = 0! = 1$.
Using integration by parts, Definition 1 gives the recurrence relation for $\Gamma(x)$:

$$\Gamma(x+1) = x\Gamma(x) \tag{1.19}$$

[when x is a positive integer, $\Gamma(x+1) = x!$]. This enables us to define $\Gamma(x)$ over the entire real line, except where x is zero or a negative integer, as

$$\Gamma(x) = \begin{cases} \displaystyle\int_0^{\infty} t^{x-1}e^{-t}\,dt, & x > 0, \\ x^{-1}\Gamma(x+1), & x < 0, \qquad x \neq -1, -2, \ldots. \end{cases} \tag{1.20}$$

From Definition 3 it can be shown that $\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}$; this implies that

$$\int_0^{\infty}\frac{e^{-t}}{t^{1/2}}\,dt = \sqrt{\pi};$$

hence, by taking $t = u^2$, we obtain

$$\int_0^{\infty}\exp\left(\frac{-u^2}{2}\right)du = \sqrt{\frac{\pi}{2}}. \tag{1.21}$$

Also, from $\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}$, we have

$$\Gamma\left(n+\tfrac{1}{2}\right) = \frac{(2n)!\pi^{1/2}}{n!2^{2n}}, \tag{1.22}$$

Definition 3 and the product formula

$$\sin(\pi x) = \pi x \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2}\right) \tag{1.23}$$

together imply that

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}, \qquad x \neq 0, -1, -2, \ldots. \tag{1.24}$$

Legendre's duplication formula [A.-M. Legendre, 1752–1833] is

$$\sqrt{\pi}\Gamma(2x) = 2^{2x-1}\Gamma(x)\Gamma\left(x + \tfrac{1}{2}\right), \qquad x \neq 0, -\tfrac{1}{2}, -1, -\tfrac{3}{2}, \ldots. \tag{1.25}$$

Gauss's multiplication theorem is

$$\Gamma(mx) = (2\pi)^{(1-m)/2} m^{mx-1/2} \prod_{j=1}^{m} \Gamma\left(x + \frac{j-1}{m}\right),$$
$$x \neq 0, -\frac{1}{m}, -\frac{2}{m}, -\frac{3}{m}, \ldots, \tag{1.26}$$

where $m = 1, 2, 3, \ldots$. This clearly reduces to Legendre's duplication formula when $m = 2$.

Many approximations for probabilities and cumulative probabilities have been obtained using various forms of *Stirling's expansion* [J. Stirling, 1692–1770] for the gamma function:

$$\Gamma(x+1) \sim (2\pi)^{1/2}(x+1)^{x+1/2}e^{-x-1} \times \exp\left(\frac{1}{12(x+1)} - \frac{1}{360(x+1)^3} + \frac{1}{1260(x+1)^5} - \cdots\right), \tag{1.27}$$

$$\Gamma(x+1) \sim (2\pi)^{1/2}x^{x+1/2}e^{-x} \times \exp\left(\frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \frac{1}{1680x^7} + \cdots\right), \tag{1.28}$$

$$\Gamma(x+1) \sim (2\pi)^{1/2}(x+1)^{x+1/2}e^{-x-1} \times \left(1 + \frac{1}{12(x+1)} + \frac{1}{288(x+1)^2} - \cdots\right), \tag{1.29}$$

$$\Gamma(x+1) \sim (2\pi)^{1/2}x^{x+1/2}e^{-x} \times \left(1 + \frac{1}{12x} + \frac{1}{288x^2} - \frac{139}{51{,}840x^3} - \frac{571}{2{,}488{,}320x^4} + \cdots\right). \tag{1.30}$$

These are divergent asymptotic expansions, yielding extremely good approximations. The remainder terms for (1.27) and (1.28) are each less in absolute value than the first term that is neglected, and they have the same sign.

Barnes's expansion [E. W. Barnes, 1874–1953] is less well known, but it is useful for half integers:

$$\Gamma\left(x+\tfrac{1}{2}\right) \sim (2\pi)^{1/2} x^x e^{-x} \exp\left(-\frac{1}{24x} + \frac{7}{2880x^3} - \frac{31}{40320x^5} + \cdots\right). \quad (1.31)$$

Also

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} \sim x^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2x} + \cdots\right). \quad (1.32)$$

These also are divergent asymptotic expansions. Series (1.31) has accuracy comparable to (1.27) and (1.28).

The *beta function* $B(a, b)$ is defined by the *Eulerian integral of the first kind*:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt, \qquad a > 0, \qquad b > 0. \quad (1.33)$$

Clearly $B(a, b) = B(b, a)$. Putting $t = u/(1+u)$ gives

$$B(a,b) = \int_0^\infty \frac{u^{a-1}\,du}{(1+u)^{a+b}}\,du, \qquad a > 0, \qquad b > 0. \quad (1.34)$$

The relationship between the beta and gamma functions is

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \qquad a, b \neq 0, -1, -2 \ldots. \quad (1.35)$$

The derivatives of the logarithm of $\Gamma(a)$ are also useful, though they are not needed as often as the gamma function itself. The function

$$\psi(x) = \frac{d}{dx}[\ln \Gamma(x)] = \frac{\Gamma'(x)}{\Gamma(x)} \quad (1.36)$$

is called the *digamma function* (with argument x) or the *psi function*. Similarly

$$\psi'(x) = \frac{d}{dx}[\psi(x)] = \frac{d^2}{dx^2}[\ln \Gamma(x)]$$

is called the *trigamma function*, and generally

$$\psi^{(s)}(x) = \frac{d^s}{dx^s}[\psi(x)] = \frac{d^{s+1}}{dx^{s+1}}[\ln \Gamma(x)] \tag{1.37}$$

is called the $(s+2)$-*gamma function*. Extensive tables of the digamma, trigamma, tetragamma, pentagamma, and hexagamma functions are contained in Davis (1933, 1935). Shorter tables are in Abramowitz and Stegun (1965).

The recurrence formula (1.19) for the gamma function yields the following recurrence formulas for the psi function:

$$\psi(x+1) = \psi(x) + x^{-1}$$

and

$$\psi(x+n) = \psi(x) + \sum_{j=1}^{n} (x+j-1)^{-1}, \qquad n = 1, 2, 3, \ldots. \tag{1.38}$$

Also

$$\psi(x) = \lim_{n\to\infty} \left[\ln(n) - \sum_{j=0}^{n} (x+j)^{-1} \right]$$

$$= -\gamma - \frac{1}{x} + \sum_{j=1}^{\infty} \frac{x}{j(x+j)} \tag{1.39}$$

$$= -\gamma + (x-1) \sum_{j=0}^{\infty} [(j+1)(j+x)]^{-1} \tag{1.40}$$

and

$$\psi(mx) = \ln(m) + \frac{1}{m} \sum_{j=0}^{m-1} \psi\left(x + \frac{j}{m}\right), \qquad m = 1, 2, 3, \ldots, \tag{1.41}$$

where γ is Euler's constant ($\cong 0.5772156649\ldots$).

An asymptotic expansion for $\psi(x)$ is

$$\psi(x) \sim \ln x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6} + \cdots, \tag{1.42}$$

and hence a very good approximation for $\psi(x)$ is $\psi(x) \approx \ln(x - 0.5)$, provided that $x \geq 2$. Particular values of $\psi(x)$ are

$$\psi(1) = -\gamma, \qquad \psi\left(\tfrac{1}{2}\right) = -\gamma - 2\ln(2) \approx -1.963510\ldots.$$

1.1.3 Finite Difference Calculus

The *displacement operator* E increases the argument of a function by unity:

$$E[f(x)] = f(x+1),$$
$$E[E[f(x)]] = E[f(x+1)] = f(x+2).$$

More generally,

$$E^n[f(x)] = f(x+n) \tag{1.43}$$

for any positive integer n, and we interpret $E^h[f(x)]$ as $f(x+h)$ for any real h.

The *forward-difference operator* Δ is defined by

$$\Delta f(x) = f(x+1) - f(x). \tag{1.44}$$

Noting that

$$f(x+1) - f(x) = E[f(x)] - f(x) = (E-1)f(x),$$

we have the *symbolic* (or *operational*) relation

$$\Delta \equiv E - 1. \tag{1.45}$$

If n is an integer, then the nth *forward difference* of $f(x)$ is

$$\Delta^n f(x) = (E-1)^n f(x) = \sum_{j=0}^{n} \binom{n}{j} (-1)^j E^{n-j} f(x)$$
$$= \sum_{j=0}^{n} \binom{n}{j} (-1)^j f(x+n-j). \tag{1.46}$$

Also, rewriting (1.45) as $E = 1 + \Delta$, we have

$$f(x+n) = (1+\Delta)^n f(x) = \sum_{j=0}^{n} \binom{n}{j} \Delta^j f(x). \tag{1.47}$$

Newton's forward-difference (interpolation) formula [I. Newton, 1642–1727] is obtained by replacing n by h, where h may be any real number, and using the interpretation of $E^h[f(x)]$ as $f(x+h)$:

$$f(x+h) = (1+\Delta)^h = f(x) + h\,\Delta f(x) + \frac{h(h-1)}{2!}\Delta^2 f(x) + \cdots. \tag{1.48}$$

The series on the right-hand side need not terminate. However, if h is small and $\Delta^n f(x)$ decreases rapidly enough as n increases, then a good approximation to $f(x+h)$ may be obtained with but few terms of the expansion. This expansion may then be used to interpolate values of $f(x+h)$, given values $f(x)$, $f(x+1), \ldots$, at unit intervals.

The *backward-difference operator* ∇ is defined similarly, by the equation

$$\nabla f(x) = f(x) - f(x-1) = (1 - E^{-1}) f(x). \tag{1.49}$$

Note that $\nabla \equiv \Delta E^{-1} \equiv E^{-1}\Delta$. There is a backward-difference interpolation formula analogous to Newton's forward-difference formula.

The *central-difference operator* δ is defined by

$$\begin{aligned} \delta f(x) &= f\left(x + \tfrac{1}{2}\right) - f\left(x - \tfrac{1}{2}\right) \\ &= (E^{1/2} - E^{-1/2}) f(x). \end{aligned} \tag{1.50}$$

Note that $\delta \equiv \Delta E^{-1/2} \equiv E^{-1/2}\Delta$. *Everett's central-difference interpolation formula* [W. N. Everett, 1924–]

$$\begin{aligned} f(x+h) = (1-h) f(x) + hf(x+1) - \tfrac{1}{6}(1-h)[1-(1-h)^2]\delta^2 f(x) \\ - \tfrac{1}{6}h(1-h^2)\delta^2 f(x+1) + \cdots \end{aligned}$$

is especially useful for computation.

Newton's forward-difference formula (1.48) can be rewritten as

$$f(x+h) = \sum_{j=0}^{\infty} \binom{h}{j} \Delta^j f(x). \tag{1.51}$$

If $f(x)$ is a polynomial of degree N, this expansion ends with the term containing $\Delta^N f(x)$.

Applying the difference operator Δ to the descending factorial $x^{(N)}$ gives

$$\begin{aligned} \Delta x^{(N)} &= (x+1)^{(N)} - x^{(N)} \\ &= (x+1)x(x-1)\cdots(x-N+2) - x(x-1)(x-2)\cdots(x-N+1) \\ &= [(x+1) - (x-N+1)]x(x-1)\cdots(x-N+2) \\ &= Nx^{(N-1)}. \end{aligned} \tag{1.52}$$

Repeating the operation, we have

$$\Delta^j x^{(N)} = N^{(j)} x^{(N-j)}, \qquad j \le N. \tag{1.53}$$

For $j > N$ we have $\Delta^j x^{(N)} = 0$.

Putting $x = 0$, $h = x$, and $f(x) = x^n$ in (1.51) gives

$$x^n = \sum_{k=0}^{n} \binom{x}{k} \Delta^k 0^n = \sum_{k=0}^{n} \frac{S(n,k)x!}{(x-k)!}, \tag{1.54}$$

where $\Delta^k 0^n / k!$ in (1.54) means $\Delta^k x^n / k!$ evaluated at $x = 0$ and is called a *difference of zero*. The multiplier $S(n,k) = \Delta^k 0^n / k!$ of the descending factorials in (1.54) is called a *Stirling number of the second kind.*

Equation (1.54) can be inverted to give the descending factorials as polynomials in x with coefficients called *Stirling numbers of the first kind*:

$$\frac{x!}{(x-n)!} = \sum_{j=0}^{n} s(n,j)x^j. \tag{1.55}$$

These notations for the Stirling numbers of the first and second kinds have won wide acceptance in the statistical literature. However, there are no standard symbols in the mathematical literature. Other notations for the Stirling numbers are as follows:

First Kind	Second Kind	Reference
$s(n,j)$	$S(n,k)$	Riordan (1958)
$\binom{n-1}{j-1} B_{n-j}^{(n)}$	$\binom{n}{n} k B_{n-k}^{(-k)}$	Milne-Thompson (1933)
	$\Delta^k 0^n / k!$	David and Barton (1962)
$S_n^{(j)}$	$\mathbf{S}_n^{(m)}$	Abramowitz and Stegun (1965)
S_n^j	$\mathfrak{S}_k^n$	Jordan (1950)
S_n^j	σ_n^k	Patil et al. (1984)
$S(n,j)$	$Z(n,k)$	Wimmer and Altmann (1999)

Both sets of numbers are nonzero only for $j = 0, 1, 2, \ldots, n$, $k = 0, 1, 2, \ldots, n$, $n > 0$. For given n or given k, the Stirling numbers of the first kind alternate in sign. The Stirling numbers of the second kind are always positive. An extensive tabulation of the numbers and details of their properties appear in Abramowitz and Stegun (1965) and in Goldberg et al. (1976). The numbers increase very rapidly as their parameters increase.

Useful properties are

$$[\ln(1+x)]^j = j! \sum_{n=j}^{\infty} \frac{s(n,j)x^n}{n!}, \tag{1.56}$$

$$(e^x - 1)^k = k! \sum_{n=k}^{\infty} \frac{S(n,k)x^n}{n!}. \tag{1.57}$$

Also

$$s(n+1, j) = s(n, j-1) - ns(n, j), \tag{1.58}$$

$$S(n+1, k) = kS(n, k) + S(n, k-1), \tag{1.59}$$

and

$$\sum_{j=m}^{n} S(n, j)s(j, m) = \sum_{j=m}^{n} s(n, j)S(j, m) = \delta_{m,n}, \tag{1.60}$$

where $\delta_{m,n}$ is *Kronecker delta* [L. Kronecker, 1823–1891]; that is, $\delta_{m,n} = 1$ for $m = n$ and zero otherwise.

Charalambides and Singh (1988) have written a useful review and bibliography concerning the Stirling numbers and their generalizations. Charalambides's (2002) book deals in depth with many types of special numbers that occur in combinatorics, including generalizations and modifications of the Stirling numbers and the Carlitz, Carlitz–Riordan, Eulerian, and Lah numbers.

The *Bell numbers* are partial sums of Stirling numbers of the second kind,

$$B_m = \sum_{j=0}^{m} S(m, j).$$

The *Catalan numbers* are

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

The *Fibonacci numbers* are

$$\begin{aligned} F_0 &= F_1 = 1, \\ F_2 &= F_0 + F_1 = 2, \\ F_3 &= F_1 + F_2 = 3, \\ F_4 &= F_2 + F_3 = 5, \\ &\vdots \end{aligned}$$

Their generating function is $g(t) = 1/(1 - t - t^2)$.

The *Narayana numbers* are

$$N(n, k) = \frac{1}{n}\binom{n}{k}\binom{n}{k-1}.$$

1.1.4 Differential Calculus

Next we introduce from the differential calculus the *differential operator* D, defined by

$$Df(x) = f'(x) = \frac{df(x)}{dx}. \tag{1.61}$$

More generally

$$D^j x^N = N^{(j)} x^{N-j}, \qquad j \leq N. \tag{1.62}$$

Note the analogy between (1.53) and (1.62). If the function $f(x)$ can be expressed in terms of a Taylor series, then the Taylor series is

$$f(x+h) = \sum_{j=0}^{\infty} \left(\frac{h^j}{j!}\right) D^j f(x). \tag{1.63}$$

The operator D acting on $f(x)$ formally satisfies

$$\sum_{j=0}^{\infty} \frac{(hD)^j}{j!} \equiv e^{hD}. \tag{1.64}$$

Comparing (1.48) with (1.63), we have (again formally)

$$e^{hD} \equiv (1+\Delta)^h \qquad \text{and} \qquad e^D \equiv 1 + \Delta. \tag{1.65}$$

Although this is only a *formal* relation between operators, it gives exact results when $f(x)$ is a polynomial of finite order; it gives useful approximations in many other cases, especially when $D^j f(x)$ and $\Delta^j f(x)$ decrease rapidly as j increases.

Rewriting $e^D \equiv 1 + \Delta$ as $D \equiv \ln(1+\Delta)$, we obtain a *numerical differentiation* formula

$$f'(x) = Df(x) = \Delta f(x) - \tfrac{1}{2}\Delta^2 f(x) + \tfrac{1}{3}\Delta^3 f(x) - \cdots. \tag{1.66}$$

(This is not the only numerical differentiation formula. There are others that are sometimes more accurate. This one is quoted as an example.)

Given a change of variable, $x = (1+t)$, we have

$$[D^k f(x)]_{x=1+t} = D^k f(1+t). \tag{1.67}$$

Consider now the *differential operator* θ, defined by

$$\theta f(x) = xDf(x) = xf'(x) = x\frac{df(x)}{dx}. \tag{1.68}$$

This satisfies

$$\theta^k f(x) = \sum_{j=1}^{k} S(k, j) x^j D^j f(x) \tag{1.69}$$

and

$$x^k D^k f(x) = \theta(\theta - 1) \cdots (\theta - k + 1) f(x). \tag{1.70}$$

Also

$$[\theta^k f(x)]_{x=e^t} = D^k f(e^t), \tag{1.71}$$

$$e^{-ct}[\theta^k f(x)]_{x=e^t} = (D + c)^k [e^{-ct} f(e^t)], \tag{1.72}$$

and

$$\begin{aligned} x^c \theta^k [x^{-c} f(x)] &= [e^{ct} D^k \{e^{-ct} f(e^t)\}]_{e^t = x} \\ &= [(D - c)^k f(e^t)]_{e^t = x} \\ &= (\theta - c)^k f(x). \end{aligned} \tag{1.73}$$

The D and θ operators are useful for handling moment properties of distributions.

Lagrange's expansion [J. L. Lagrange, 1736–1813] for the reversal of a power series assumes that if (1) $y = f(x)$, where $f(x)$ is regular in the neighborhood of x_0, (2) $y_0 = f(x_0)$, and (3) $f'(x_0) \neq 0$, then

$$x = x_0 + \sum_{k=1}^{\infty} \frac{(y - y_0)^k}{k!} \left[\frac{d^{k-1}}{dx^{k-1}} \left(\frac{x - x_0}{f(x) - y_0} \right)^k \right]_{x=x_0}. \tag{1.74}$$

More generally

$$h(x) = h(x_0) + \sum_{k=1}^{\infty} \frac{(y - y_0)^k}{k!} \left[\frac{d^{k-1}}{dx^{k-1}} \left\{ h'(x) \left(\frac{x - x_0}{f(x) - y_0} \right)^k \right\} \right]_{x=x_0}, \tag{1.75}$$

where $h(x)$ is infinitely differentiable. (This expansion plays an important role in the theory of Lagrangian distributions; see Section 2.5.)

L'Hôpital's rule [G. F. A. de L'Hôpital, 1661–1704] is useful for finding the limit of an indeterminate form. If $f(x)$ and $g(x)$ are functions of x for which $\lim_{x \to b} f(x) = \lim_{x \to b} g(x) = 0$, and if $\lim_{x \to b} [f'(x)/g'(x)]$ exists, then

$$\lim_{x \to b} \frac{f(x)}{g(x)} = \lim_{x \to b} \frac{f'(x)}{g'(x)}. \tag{1.76}$$

The use of the O, o notation (*Landau's notation*) [E. Landau, 1877–1938] is standard. We say that

$$f(x) = o(g(x)) \qquad \text{as } x \to \infty \qquad \text{if } \lim_{x \to \infty} \left(\frac{f(x)}{g(x)} \right) = 0$$

and

$$f(x) = O(g(x)) \qquad \text{as } x \to \infty \qquad \text{if } \left|\frac{f(x)}{g(x)}\right| < C \tag{1.77}$$

for some constant C and large x.

1.1.5 Incomplete Gamma and Beta Functions and Other Gamma-Related Functions

In statistical work we often encounter the *incomplete gamma function* $\gamma(a, x)$ and its complement $\Gamma(a, x)$; see Khamis (1960) for a discussion of incomplete gamma function expansions of statistical distribution functions. These functions are defined by

$$\begin{aligned} \gamma(a, x) &= \int_0^x t^{a-1} e^{-t}\, dt, \\ \Gamma(a, x) &= \int_x^\infty t^{a-1} e^{-t}\, dt, \qquad x > 0; \end{aligned} \tag{1.78}$$

that is,

$$\gamma(a, x) + \Gamma(a, x) = \Gamma(a).$$

The notation $\Gamma_x(a) = \gamma(a, x)$ is also in use.

Infinite-series formulas are

$$\begin{aligned} \gamma(a, x) &= a^{-1} x^a\, {}_1F_1[a; a+1; -x] = \sum_{n=0}^{\infty} \frac{(-1)^n x^{a+n}}{n!(a+n)} \\ &= a^{-1} x^a e^{-x}\, {}_1F_1[1; a+1; x] = e^{-x} \sum_{n=0}^{\infty} \frac{(a-1)! x^{a+n}}{(a+n)!}, \end{aligned} \tag{1.79}$$

$a \neq 0, -1, -2, \ldots$, where ${}_1F_1[\cdot]$ is a confluent hypergeometric function; see Section 1.1.7.

The following recursion formulas are useful:

$$\begin{aligned} \gamma(a+1, x) &= a\gamma(a, x) - x^a e^{-x}, \\ \Gamma(a+1, x) &= a\Gamma(a, x) + x^a e^{-x}. \end{aligned} \tag{1.80}$$

For x real, $x \to \infty$,

$$\Gamma(a, x) \sim x^{a-1} e^{-x} \left[1 + \frac{a-1}{x} + \frac{(a-1)(a-2)}{x^2} + \cdots\right].$$

The *incomplete gamma function ratio*

$$\frac{\Gamma_x(a)}{\Gamma(a)} = \frac{\gamma(a, x)}{\Gamma(a)}$$

is used in the statistical literature more than $\Gamma_x(a) = \gamma(a, x)$ itself. (The word "ratio" is, alas, sometimes omitted.)

The function tabulated in Pearson's (1922) tables is

$$I(u, p) = \frac{\Gamma_{u\sqrt{p+1}}(p+1)}{\Gamma(p+1)}; \tag{1.81}$$

it is given to seven decimal places for $p = -1(0.05)0(0.1)5(0.2)50$, with u at intervals of 0.1. Harter (1964) gave $I(u, p)$ to nine decimal places for $p = -0.5(0.5)74(1)164$ and u at intervals of 0.1. We note also the extensive tables of Khamis and Rudert (1965).

Pearson and Hartley (1976) [see also Abramowitz and Stegun (1965)] tabulated the function

$$Q(\chi^2|\nu) = \frac{\Gamma(\nu/2, \chi^2/2)}{\Gamma(\nu/2)}$$

(the upper tail of a χ^2 distribution) for

$$\chi^2 = 0.001(0.001)0.01(0.01)0.1(0.1)2(0.2)10(0.5)20(1)40(2)76,$$
$$\nu = 1(1)30$$

to five decimal places.

Just as we often need the incomplete gamma function, so we need also the *incomplete beta function*

$$B_p(a, b) = \int_0^p t^{a-1}(1-t)^{b-1}\,dt, \qquad 0 < p < 1, \tag{1.82}$$

and the *incomplete beta function ratio*

$$I_p(a, b) = \frac{B_p(a, b)}{B(a, b)}. \tag{1.83}$$

Again the word "ratio" is often omitted. In terms of the hypergeometric function ${}_2F_1[\cdot]$ (cf. Section 1.1.6) we have

$$B_p(a, b) = a^{-1}p^a\,{}_2F_1[a, 1-b; a+1; p] = \sum_{n=0}^{\infty} \frac{(1-b)_n p^{a+n}}{n!(a+n)}. \tag{1.84}$$

The incomplete beta function ratio $I_p(a, b)$ has the following properties:

$$
\begin{aligned}
I_p(a, b) &= 1 - I_{1-p}(b, a), \\
I_p(k, n-k+1) &= \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}, \qquad 1 \le k \le n, \\
I_p(a, b) &= pI_p(a-1, b) + (1-p)I_p(a, b-1), \qquad (1.85) \\
(a+b-ap)I_p(a, b) &= a(1-p)I_p(a+1, b-1) + bI_p(a, b+1), \\
(a+b)I_p(a, b) &= aI_p(a+1, b) + bI_p(a, b+1).
\end{aligned}
$$

Extensive tables of $I_p(a, b)$ to seven decimal places are contained in Pearson (1934) for $p = 0.01(0.01)1$; $a, b = 0.5(0.5)11(1)50$, $a \ge b$. These may be supplemented for small values of a by the tables of Vogler (1964). Both Pearson and Vogler give values for the complete beta function $B(a, b)$.

Pearson and Hartley (1976) have tabulated the percentage points of the F distribution with upper tail

$$
Q(F|\nu_1, \nu_2) = I_p\left(\tfrac{1}{2}\nu_2, \tfrac{1}{2}\nu_1\right),
$$

where $p = \nu_2/(\nu_2 + \nu_1 F)$ for

$$
\begin{aligned}
Q(F|\nu_1, \nu_2) &= 0.001, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, \\
\nu_1 &= 1(1)6, 8, 12, 15, 20, 30, 60, \infty, \\
\nu_2 &= 1(1)30, 40, 60, 120, \infty,
\end{aligned}
$$

to at least three significant digits; this table is quoted in Abramowitz and Stegun (1965).

The *Laplace transform* [P. S. Laplace, 1749–1827] of a function $f(t)$ is defined as

$$
F(p) = \int_0^\infty f(t)e^{-pt}\,dt. \qquad (1.86)
$$

The *error function* $\operatorname{erf}(x)$ is defined by

$$
\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,dt. \qquad (1.87)
$$

It is closely related to the *normal distribution function*,

$$
\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt = 0.5\left[1 + \operatorname{erf}\left(\frac{\mathrm{x}}{\sqrt{2}}\right)\right]. \qquad (1.88)
$$

Its complement is $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$. Sometimes one sees

$$\operatorname{Erf}(x) = 0.5\sqrt{\pi}\operatorname{erf}(x), \qquad \operatorname{Erfc}(x) = 0.5\sqrt{\pi}\operatorname{erfc}(x).$$

The *Bessel function of the first kind* $J_\nu(x)$ is

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{j=0}^{\infty} \frac{(-x^2/4)^j}{j!\Gamma(\nu + j + 1)}, \tag{1.89}$$

where ν is the order of the function. The *modified Bessel function of the first kind* is

$$I_\nu(x) = (-i)^\nu J_\nu(ix) = \left(\frac{x}{2}\right)^\nu \sum_{j=0}^{\infty} \frac{(x^2/4)^j}{j!\Gamma(\nu + j + 1)}, \tag{1.90}$$

where $i = \sqrt{-1}$.

The *modified Bessel function of the third kind*, $K_\nu(y)$, is defined as

$$K_\nu(y) = \frac{\pi}{2} \cdot \frac{I_{-\nu}(y) - I_\nu(y)}{\sin(\nu\pi)} \tag{1.91}$$

when ν is not an integer or zero. When ν is an integer or zero, the right-hand side of this definition is replaced by its limiting value; see, for example, Abramowitz and Stegun (1965). Sometimes $K_\nu(y)$ is called the modified Bessel function of the *second* kind in the statistical literature.

Useful properties are

$$K_{-\nu}(y) = K_\nu(y) \tag{1.92}$$

and the recurrence relation

$$K_{\nu+1}(y) = \frac{2\nu}{y} K_\nu(y) + K_{\nu-1}(y). \tag{1.93}$$

The *Riemann zeta function* [G. F. B. Riemann, 1826–1866] is defined by the equation

$$\zeta(x) = \sum_{j=1}^{\infty} j^{-x}. \tag{1.94}$$

The series is convergent for $x > 1$, and it is only for these values of x that we shall use the function. A *generalized form of the Riemann zeta function* is defined by

$$\zeta(x, a) = \sum_{j=1}^{\infty} (j + a)^{-x}, \tag{1.95}$$

where $x > 1$ and $a > 0$.

An approximate formula for $\zeta(x)$ is

$$\zeta(x) \approx 1 + \frac{2x^2 + 8.4x + 21.6}{(x-1)(x+7)2^{x+1}}.$$

Particular values are

$$\zeta(2) = \frac{\pi^2}{6} \qquad \text{and} \qquad \zeta(4) = \frac{\pi^4}{90}.$$

Values of $\zeta(n)$ for $n = 2(1)42$ to 20 decimal places are given in Abramowitz and Stegun (1965).

A general formula for even values of the argument is

$$\zeta(2r) = \frac{(2\pi)^{2r}}{2[(2r)!]}|B_{2r}|, \tag{1.96}$$

where B_{2r} is a Bernoulli number (see Section 1.1.9).

The *Lerch function* is

$$\Phi(z, s, \nu) = \sum_{j=0}^{\infty} \frac{z^j}{(\nu + j)^s}, \qquad \nu, z, s \text{ real}, \qquad \nu \neq 0, -1, -2, \ldots. \tag{1.97}$$

1.1.6 Gaussian Hypergeometric Functions

The *hypergeometric function*, or more precisely the *Gaussian hypergeometric function*, has the form

$$\begin{aligned} {}_2F_1[a, b; c; x] &= 1 + \frac{ab}{c\ 1!}x + \frac{a(a+1)b(b+1)}{c(c+1)\ 2!}x^2 + \cdots \\ &= \sum_{j=0}^{\infty} \frac{(a)_j (b)_j x^j}{(c)_j\ j!}, \qquad c \neq 0, -1, -2, \ldots, \end{aligned} \tag{1.98}$$

where $(a)_j$ is Pochhammer's symbol (1.4). The suffixes refer to the numbers of numerator and denominator parameters—there are two numerator parameters and one denominator parameter. Clearly ${}_2F_1[b, a; c; x] = {}_2F_1[a, b; c; x]$.

We will only be interested in the case where a, b, c, and x are real. If a is a nonpositive integer, then $(a)_j$ is zero for $j > -a$, and the series terminates. When the series is infinite, it is absolutely convergent for $|x| < 1$ and divergent for $|x| > 1$. For $|x| = 1$, it is

1. absolutely convergent if $c - a - b > 0$;
2. conditionally convergent if $-1 < c - a - b \leq 0$, $x = -1$; and
3. divergent if $c - a - b \leq -1$.

When $a = 1$ and $b = c$ (or $b = 1$ and $a = c$), the series becomes $1 + x + x^2 + \cdots$; hence the name "hypergeometric."

Gauss's summation theorem states that, when $x = 1$,

$$ {}_2F_1[a, b; c; x] = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} = \frac{B(c, c-a-b)}{B(c-a, c-b)}, \tag{1.99} $$

where $c - a - b > 0$, $c \neq 0, -1, -2, \ldots$.

When a is a nonpositive integer, $a = -n$ say, and $b = -u$, $c = v - n + 1$, this becomes *Vandermonde's theorem* (see Section 1.1.1):

$$ \sum_{j=0}^{n} \binom{u}{j}\binom{v}{n-j} = \binom{u+v}{n}. \tag{1.100} $$

The Gaussian hypergeometric function satisfies the second-order linear differential equation

$$ x(1-x)\frac{d^2y}{dx^2} + [c - (a+b+1)x]\frac{dy}{dx} - aby = 0, \tag{1.101} $$

or, equivalently,

$$ [\theta(\theta + c - 1) - x(\theta + a)(\theta + b)]y = 0, \tag{1.102} $$

where θ is the differential operator $x(d/dx)$; see Section 1.1.4.

The Gaussian hypergeometric function has been described as "the wooden plough of the nineteenth century"; it occurs frequently in mathematical applications because every linear differential equation of the second order, whose singularities are regular and at most three in number, can be transformed into the hypergeometric equation.

The derivatives are

$$ \begin{aligned} \frac{d}{dx}\,{}_2F_1[a, b; c; x] &= \frac{ab}{c}\,{}_2F_1[a+1, b+1; c+1; x], \\ D^n\,{}_2F_1[a, b; c; x] &= \frac{(a)_n(b)_n}{(c)_n}\,{}_2F_1[a+n, b+n; c+n; x]. \end{aligned} \tag{1.103} $$

Euler's integral for the function is

$$ {}_2F_1[a, b; c; x] = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 u^{a-1}(1-u)^{c-a-1}(1-xu)^{-b}\,du, \tag{1.104} $$

where $c > a > 0$. The function is also a Laplace transform:

$$ {}_2F_1\left[a, b; c; \frac{k}{s}\right] = \frac{s^b}{\Gamma(b)} \int_0^\infty e^{-su}u^{b-1}\,{}_1F_1[a; c; ku]\,du. \tag{1.105} $$

The *Euler transformations* are

$$\begin{aligned} {}_2F_1[a,b;c;x] &= (1-x)^{-a}\ {}_2F_1\left[a,c-b;c;\frac{x}{x-1}\right] \\ &= (1-x)^{-b}\ {}_2F_1\left[c-a,b;c;\frac{x}{x-1}\right] \\ &= (1-x)^{c-a-b}\ {}_2F_1[c-a,c-b;c;x]. \end{aligned} \tag{1.106}$$

Hypergeometric representations of elementary functions are

$$\begin{aligned} (1-x)^{-a} &= {}_2F_1[a,b;b;x], \\ \ln(1+x) &= x\ {}_2F_1[1,1;2;-x], \\ \ln\left(\frac{1+x}{1-x}\right) &= 2x\ {}_2F_1\left[\frac{1}{2},1;\frac{3}{2};x^2\right], \\ \arcsin(x) &= x\ {}_2F_1\left[\frac{1}{2},\frac{1}{2};\frac{3}{2};x^2\right], \\ \arctan(x) &= x\ {}_2F_1\left[\frac{1}{2},1;\frac{3}{2};-x^2\right], \\ \left(\frac{1}{2}+\frac{\sqrt{1-x}}{2}\right)^{1-2a} &= {}_2F_1\left[a,a-\frac{1}{2};2a;x\right], \\ &= (1-x)^{1/2}\ {}_2F_1\left[a,a+\frac{1}{2};2a;x\right], \\ (1+x)^{-2a}+(1-x)^{-2a} &= 2\ {}_2F_1\left[a,a+\frac{1}{2};\frac{1}{2};x^2\right], \\ (1-x)^{-2a-1}(1+x) &= {}_2F_1[a+1,2a;a;x]. \end{aligned} \tag{1.107}$$

A large number of special functions can also be represented as Gaussian hypergeometric functions. The incomplete beta function is

$$B_p(a,b) = a^{-1}p^a\ {}_2F_1[a,1-b;a+1;p], \tag{1.108}$$

the *Legendre polynomials* are

$$P_n(x) = {}_2F_1\left[-n,n+1;1;\tfrac{1}{2}(1-x)\right], \tag{1.109}$$

the *Chebyshev polynomials* [P. L. Chebyshev, 1821–1894] are

$$T_n(x) = {}_2F_1\left[-n, n; \tfrac{1}{2}; \tfrac{1}{2}(1-x)\right], \tag{1.110}$$

$$U_n(x) = (n+1)\,{}_2F_1\left[-n, n+2; \tfrac{3}{2}; \tfrac{1}{2}(1-x)\right], \tag{1.111}$$

and the *Jacobi polynomials* [C. G. J. Jacobi, 1804–1851] are

$$P_n^{(a,b)}(x) = \binom{a+n}{n}\,{}_2F_1\left[-n, a+b+n+1; a+1; \tfrac{1}{2}(1-x)\right]. \tag{1.112}$$

For detailed studies of the Gaussian hypergeometric function, including recurrence relationships between contiguous functions, see Bailey (1935), Erdélyi et al. (1953, Vol. 1), Slater (1966), and Luke (1975).

1.1.7 Confluent Hypergeometric Functions (Kummer's Functions)

Notations vary for the *confluent hypergeometric function* (also known as *Kummer's series* [E. E. Kummer, 1810–1893]). We have

$$\begin{aligned} {}_1F_1[a; c; x] &= 1 + \frac{a}{c1!}x + \frac{a(a+1)}{c(c+1)2!}x^2 + \cdots \\ &= \sum_{j=0}^{\infty} \frac{(a)_j x^j}{(c)_j\, j!} \\ &= \lim_{|b|\to\infty} {}_2F_1\left[a, b; c; \frac{x}{b}\right], \qquad c \neq 0, -1, -2, \ldots, \end{aligned} \tag{1.113}$$

where $(a)_j$ is Pochhammer's symbol. Other notations for ${}_1F_1[a; c; x]$ are $M(a; c; x)$ and $\phi(a; c; x)$. The suffixes in ${}_1F_1[a; c; x]$ emphasize that there is one numerator parameter and one denominator parameter. If a is a nonpositive integer, the series terminates. The series converges for all real values of a, c, and x, provided that c is not a nonpositive integer. When $a = c$, $c > 0$, the series becomes the exponential series $1 + x + x^2/2! + x^3/3! + \cdots$.

The confluent hypergeometric function satisfies Kummer's differential equation

$$x\frac{d^2y}{dx^2} + (c - x)\frac{dy}{dx} - ay = 0, \tag{1.114}$$

that is,

$$[\theta(\theta + c - 1) - x(\theta + a)]y = 0, \tag{1.115}$$

where $\theta \equiv x(d/dx)$.

The derivatives of the confluent hypergeometric function are

$$\begin{aligned} \frac{d}{dx}\,{}_1F_1[a;c;x] &= \frac{a}{c}\,{}_1F_1[a+1;c+1;x], \\ D^n\,{}_1F_1[a;c;x] &= \frac{(a)_n}{(c)_n}\,{}_1F_1[a+n;c+n;x]. \end{aligned} \tag{1.116}$$

The following integral representation is useful:

$${}_1F_1[a;c;x] = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\int_0^1 u^{a-1}(1-u)^{c-a-1}e^{xu}\,du, \tag{1.117}$$

where $c > a > 0$.

Kummer's first theorem yields the transformation

$${}_1F_1[a;c;x] = e^x\,{}_1F_1[c-a;c;-x]. \tag{1.118}$$

Kummer's second theorem is

$$e^{-x}\,{}_1F_1[a;2a;2x] = {}_0F_1\left[\ ;a+\tfrac{1}{2};\tfrac{1}{4}x^2\right], \tag{1.119}$$

where $a + \frac{1}{2}$ is not a negative integer and

$${}_0F_1[\ ;c;x] = \lim_{|a|\to\infty}\,{}_1F_1\left[a;c;\frac{x}{a}\right] = \sum_{j=0}^{\infty}\frac{x^j}{(c)_j\ j!},$$

where $c \neq 0, 1, 2, \ldots$ and x is finite.

Kummer's differential equation is also satisfied by

$$\Psi(a,c;x) = x^{-a}\,{}_2F_0\left[a, a-c+1;\ ;-\frac{1}{x}\right], \tag{1.120}$$

$$= \frac{1}{\Gamma(a)}\int_0^\infty e^{-xu}u^{a-1}(1+u)^{c-a-1}\,du, \tag{1.121}$$

where $a > 0$, $x > 0$.

The following relationship holds:

$$\begin{aligned} \Psi(a,c;x) = {} & \frac{\Gamma(1-c)}{\Gamma(a-c+1)}\,{}_1F_1[a;c;x] \\ & + \frac{\Gamma(c-1)x^{1-c}}{\Gamma(a)}\,{}_1F_1[a-c+1;2-c;x], \end{aligned} \tag{1.122}$$

provided that $c \neq 0, \pm 1, \pm 2, \ldots$. Also

$$\Psi(a, c; x) = x^{1-c}\Psi(a - c + 1, 2 - c; x). \tag{1.123}$$

Many functions that are important in distribution theory can be expressed in terms of the confluent hypergeometric function; for example, the incomplete gamma functions are

$$\gamma(a, x) = a^{-1}x^a \, {}_1F_1[a; a + 1; -x], \tag{1.124}$$

$$\Gamma(a, x) = \Gamma(a) - a^{-1}x^a \, {}_1F_1[a; a + 1; -x], \tag{1.125}$$

and the *error functions* are

$$\begin{aligned} \mathrm{Erf}(x) &= \frac{\sqrt{\pi}}{2}\mathrm{erf(x)} = 0.5\ \gamma\left(\frac{1}{2}, x^2\right) \\ &= x \, {}_1F_1\left[\frac{1}{2}; \frac{3}{2}; -x^2\right], \end{aligned} \tag{1.126}$$

$$\mathrm{Erfc}(x) = \frac{\sqrt{\pi}}{2}\mathrm{erfc(x)} = \frac{\sqrt{\pi}}{2} - x \, {}_1F_1\left[\frac{1}{2}; \frac{3}{2}; -x^2\right]. \tag{1.127}$$

The *Hermite polynomials* [Ch. Hermite, 1822–1901] as used in statistics are defined as

$$H_n(x) = \sum_{j=0}^{[n/2]} \frac{(-1)^j n! x^{n-2j}}{(n - 2j)! j! 2^j}, \tag{1.128}$$

where $[\cdot]$ denotes the integer part. Hence

$$\begin{aligned} H_{2n}(x) &= \frac{(-1)^n (2n)!}{n! 2^n} \, {}_1F_1\left[-n; \frac{1}{2}; \frac{x^2}{2}\right], \\ H_{2n+1}(x) &= \frac{(-1)^n (2n + 1)! x}{n! 2^n} \, {}_1F_1\left[-n; \frac{3}{2}; \frac{x^2}{2}\right]; \end{aligned} \tag{1.129}$$

see Stuart and Ord (1987, Sections 6.14–6.15). Fisher (1951, p. xxxi) used the "modified" Hermite polynomials

$$H_n^*(x) = i^{-n} H_n(ix), \qquad \text{where } i = \sqrt{-1}. \tag{1.130}$$

The Bessel functions $J_\nu(x)$, $I_\nu(x)$, and $K_\nu(x)$ [F. W. Bessel, 1784–1846], Whittaker functions [E. T. Whittaker, 1873–1956], Laguerre functions and

polynomials [E. N. Laguerre, 1834–1886], and Poisson–Charlier polynomials (S. D. Poisson [1781–1840] and C. L. Charlier [1862–1939]) can also all be represented as confluent hypergeometric functions.

Further details concerning some of these functions can be found in Section 1.1.11. Thorough coverage is in Erdélyi et al. (1953, Vols. 1 and 2) and in the book devoted to confluent hypergeometric functions by Slater (1960). Readers are WARNED, however, that most mathematical texts, including those by Erdélyi and by Abramowitz and Stegun, use slightly different notations for the Hermite polynomials (differing by powers of 2). Slater (1960), Rushton and Lang (1954), and Abramowitz and Stegun (1965) give useful tables.

1.1.8 Generalized Hypergeometric Functions

The *generalized hypergeometric function* is a natural generalization of the Gaussian hypergeometric function. The series is defined as

$$
\begin{aligned}
{}_pF_q[a_1,\ldots,a_p;b_1,\ldots,b_q;\ x] &= {}_pF_q\left[\begin{matrix}a_1,\ldots,a_p;\ x\\ b_1,\ldots,b_q\end{matrix}\right]\\
&= \sum_{j=0}^{\infty}\frac{(a_1)_j\ldots(a_p)_j x^j}{(b_1)_j\ldots(b_q)_j\, j!},
\end{aligned}
\tag{1.131}
$$

where $b_i \neq 0, -1, -2, \ldots,\ i = 1,\ldots,q$.

There are p numerator parameters and q denominator parameters. Clearly the orderings of the numerator parameters and of the denominator parameters are immaterial. The simplest generalized hypergeometric series is

$$
{}_0F_0[-;-;x] = {}_0F_0[\ ;\ ;x] = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = e^x \tag{1.132}
$$

(a blank indicates the absence of a parameter).

If one of the numerator parameters a_i, $i = 1,\ldots,p$, is a negative integer, $a_1 = -n$ say, the series terminates and

$$
\begin{aligned}
&{}_pF_q\left[\begin{matrix}-n,a_2,\ldots,a_p;x\\ b_1,\ldots,b_q\end{matrix}\right]\\
&\quad= \sum_{j=0}^{n}\frac{(-n)_j(a_2)_j\ldots(a_p)_j x^j}{(b_1)_j\ldots(b_q)_j\, j!}, \qquad (1.133)\\
&\quad= \frac{(a_2)_n\ldots(a_p)_n(-x)^n}{(b_1)_n\ldots(b_q)_n}\\
&\qquad\times {}_{q+1}F_{p-1}\left[\begin{matrix}-n,1-b_1-n,\ldots,1-b_q-n;(-1)^{p+q-1}x^{-1}\\ 1-a_2-n,\ldots,1-a_p-n\end{matrix}\right]. \qquad (1.134)
\end{aligned}
$$

When the series is infinite, it converges for $|x| < \infty$ if $p \le q$, it converges for $|x| < 1$ if $p = q + 1$, and it diverges for all x, $x \neq 0$ if $p > q + 1$. Furthermore, if

$$s = \sum_{i=1}^{q} b_i - \sum_{i=1}^{p} a_i,$$

then the series with $p = q + 1$ is absolutely convergent for $|x| = 1$ if $s > 0$, is conditionally convergent for $|x| = 1$, $x \neq 1$ if $-1 < s \le 0$, and is divergent for $|x| = 1$ if $s \le -1$.

The function is characterized as a power series $\sum_{j=0}^{\infty} A_j x^j$ by the property that A_{j+1}/A_j is a rational function of j.

The function satisfies the differential equation

$$\theta(\theta + b_1 - 1) \cdots (\theta + b_q - 1)y = x(\theta + a_1) \cdots (\theta + a_p)y, \tag{1.135}$$

where θ is the differential operator $x(d/dx)$.

The derivatives are

$$\frac{d}{dx}\, {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; x] = \frac{a_1 \ldots a_p}{b_1 \ldots b_q}\, {}_pF_q[a_1 + 1, \ldots, a_p + 1; b_1 + 1, \ldots, b_q + 1; x], \tag{1.136}$$

$$D^n\, {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; x] = \frac{(a_1)_n \ldots (a_p)_n}{(b_1)_n \ldots (b_q)_n}\, {}_pF_q[a_1 + n, \ldots, a_p + n; b_1 + n, \ldots, b_q + n; x]. \tag{1.137}$$

The *Eulerian integral* generalizes to

$${}_{p+1}F_{q+1}[a_1, \ldots, a_p, c; b_1, \ldots, b_q, d; x] = \frac{\Gamma(d)}{\Gamma(c)\Gamma(d - c)} \int_0^1 u^{c-1}(1 - u)^{d-c-1}\, {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; xu]\, du. \tag{1.138}$$

Also

$${}_{p+1}F_q[a_1, \ldots, a_p, c; b_1, \ldots, b_q; x] = \frac{1}{\Gamma(c)} \int_0^\infty e^{-u} u^{c-1}\, {}_pF_q[a_1, \ldots, a_p; b_1, \ldots, b_q; xu]\, du. \tag{1.139}$$

The product of two generalized hypergeometric functions can be expressed as a series in other generalized hypergeometric functions. So can generalized hypergeometric functions with arguments of the form $x = y + z$.

A generalized hypergeometric series tail truncated after $m+1$ terms can be represented as

$$ {}_{p+1}F_{q+1}[a_1, \ldots, a_p, -m; b_1, \ldots, b_q, -m; x]. $$

Head truncation of the first k terms gives

$$ \frac{(a_1)_k \cdots (a_p)_k x^k}{(b_1)_k \cdots (b_q)_k k!}\ {}_{p+1}F_{q+1}[a_1+k, \ldots, a_p+k, 1; b_1+k, \ldots, b_q+k, 1+k; x]. \tag{1.140} $$

Generalized hypergeometric representations of elementary functions include

$$ \begin{aligned} e^x &= {}_0F_0[-; -; x] = {}_0F_0[\ ;\ ; x], \\ (1-x)^{-a} &= {}_1F_0[a; -; x] = {}_1F_0[a;\ ; x], \\ \cos(x) &= {}_0F_1\left[-; \tfrac{1}{2}; -\tfrac{1}{4}x^2\right] = {}_0F_1\left[\ ; \tfrac{1}{2}; -\tfrac{1}{4}x^2\right], \\ \sin(x) &= x\,{}_0F_1\left[-; \tfrac{3}{2}; -\tfrac{1}{4}x^2\right] = x\,{}_0F_1\left[\ ; \tfrac{3}{2}; -\tfrac{1}{4}x^2\right], \\ \arctan(x) &= x\,{}_2F_1\left[\tfrac{1}{2}, 1; \tfrac{3}{2}; -x^2\right]. \end{aligned} \tag{1.141} $$

Bessel functions can also be stated this way; for example,

$$ \begin{aligned} J_\nu(x) &= \frac{(x/2)^\nu}{\Gamma(\nu+1)}\ {}_0F_1\left[\ ; \nu+1; -\frac{x^2}{4}\right], \\ I_\nu(x) &= \frac{(x/2)^\nu}{\Gamma(\nu+1)}\ {}_0F_1\left[\ ; \nu+1; \frac{x^2}{4}\right] \end{aligned} \tag{1.142} $$

(see also Sections 1.1.5 and 1.1.7).

The *Horn–Appell functions* are generalized hypergeometric functions in two variables; they include

$$ \begin{aligned} F_1(a, b, b'; c; x, y) &= \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_m(b')_n}{(c)_{m+n}} \frac{x^m y^n}{m!n!} \\ \Phi_1(a, b; c; x, y) &= \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_n}{(c)_{m+n}} \frac{x^m y^n}{m!n!}, \qquad |x| < 1, \\ \Psi_1(a, b; c, c'; x, y) &= \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_m}{(c)_m(c')_n} \frac{x^m y^n}{m!n!}, \qquad |x| < 1, \\ \Xi_1(a, a', b; c; x, y) &= \sum_{m,n=0}^{\infty} \frac{(a)_m(a')_n(b)_m}{(c)_{m+n}} \frac{x^m y^n}{m!n!}, \qquad |x| < 1, \end{aligned} \tag{1.143} $$

where $(a)_m$ is Pochhammer's symbol.

Extensive treatments of generalized hypergeometric functions (including further references) are provided in the books by Erdélyi et al. (1953, Vol. 1), Rainville (1960), and Slater (1966). Certain useful integrals are in Erdélyi et al. (1954, Vols. 1 and 2) and Exton (1978). More advanced special functions and their statistical applications have been studied by Mathai and Saxena (1973, 1978).

1.1.9 Bernoulli and Euler Numbers and Polynomials

The *Bernoulli numbers* [J. Bernoulli, 1654–1705] $B_0, B_1, \ldots, B_r, \ldots$ are defined by the identity

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n t^n}{n!}, \tag{1.144}$$

giving

$$B_0 = 1, \quad B_1 = -\tfrac{1}{2}, \quad B_2 = \tfrac{1}{6}, \quad B_4 = -\tfrac{1}{30}, \quad B_6 = \tfrac{1}{42}, \quad B_8 = -\tfrac{1}{30},$$

with $B_{2r+1} = 0$ for $r > 0$.

The *Bernoulli polynomials* $B_0(x), B_1(x), \ldots, B_r(x), \ldots$ are defined by the identity

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n(x)t^n}{n!}. \tag{1.145}$$

Clearly $B_r(0) = B_r$. A useful formula is

$$\sum_{j=1}^{n} j^r = (r+1)^{-1}[B_{r+1}(n+1) - B_{r+1}]. \tag{1.146}$$

The polynomials have the properties that

$$\begin{aligned} \frac{dB_r(x)}{dx} &= rB_{r-1}(x), \\ B_r(x+h) &= \sum_{j=0}^{r} \binom{r}{j} B_j(x) h^{r-j} \end{aligned} \tag{1.147}$$

[symbolically $B_r(x+h) = (E+h)^r B_0(x)$ with the displacement operator E applying to the subscript].

The first seven Bernoulli polynomials are

$$\begin{aligned} B_0(x) &= 1, \\ B_1(x) &= x - \tfrac{1}{2}, \\ B_2(x) &= x^2 - x + \tfrac{1}{6}, \\ B_3(x) &= x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x, \\ B_4(x) &= x^4 - 2x^3 + x^2 - \tfrac{1}{30}, \\ B_5(x) &= x^5 - \tfrac{5}{2}x^4 + \tfrac{5}{3}x^3 - \tfrac{1}{6}x, \\ B_6(x) &= x^6 - 3x^5 + \tfrac{5}{2}x^4 - \tfrac{1}{2}x^2 + \tfrac{1}{42}. \end{aligned} \tag{1.148}$$

David et al. (1966) have tabulated the Bernoulli polynomials $B_n(x)$ for $n = 0(1)12$ and the Bernoulli numbers B_n for $n = 1(1)12$.

Let $T_k(n) = 1^k + 2^k + \cdots + n^k = \sum_{j=1}^{n} j^k$. Then

$$T_k(n) = \frac{B_{k+1}(n+1) - B_{k+1}(0)}{k+1}.$$

In particular,

$$\begin{aligned} T_1(n) &= \tfrac{1}{2}n(n+1), \\ T_2(n) &= \tfrac{1}{6}n(n+1)(2n+1) = \tfrac{1}{3}T_1(n)(2n+1), \\ T_3(n) &= \tfrac{1}{4}n^2(n+1)^2 = \tfrac{1}{2}T_1(n)(n^2+n), \\ T_4(n) &= \tfrac{1}{5}T_2(n)(3n^2+3n-1), \\ T_5(n) &= \tfrac{1}{3}T_3(n)(2n^2+2n-1), \\ T_6(n) &= \tfrac{1}{7}T_2(n)(3n^4+6n^3-3n+1). \end{aligned}$$

The *Euler numbers* E_r are defined by the identity

$$\frac{2e^x}{e^{2x}+1} = \sum_{n=0}^{\infty} \frac{E_n x^n}{n!}. \tag{1.149}$$

They satisfy the symbolic formula

$$(E+1)^n + (E-1)^n = 0, \tag{1.150}$$

with powers of E^m replaced by E_m. We find that $E_{2n+1} = 0$ and that the Euler numbers are all integers for r even:

$$\begin{aligned} E_0 &= 1, \\ E_2 &= -1, \\ E_4 &= 5, \\ E_6 &= -61, \\ E_8 &= 1{,}385, \\ E_{10} &= -50{,}521 \\ &\vdots \end{aligned}$$

Further values are given in Abramowitz and Stegun (1965).

The *Euler polynomials* $E_r(x)$ are defined by the identity

$$\frac{2e^{tx}}{e^t + 1} \equiv \sum_{j=0}^{\infty} E_j(x) \frac{t^j}{j!}. \tag{1.151}$$

Their properties include

$$E_n(x) + E_n(x+1) = 2x^n, \tag{1.152}$$

$$\frac{dE_n(x)}{dx} = nE_{n-1}(x). \tag{1.153}$$

The following symbolic relationships connect the Bernoulli and the Euler numbers:

$$\begin{aligned} E^{n-1} &\equiv \frac{(4B-1)^n - (4B-3)^n}{2n}, \\ E^{2n} &\equiv \frac{4^{2n+1}(B - 1/4)^{2n+1}}{2n+1}. \end{aligned} \tag{1.154}$$

If $m + n$ is odd, then

$$\int_0^1 B_m(x) B_n(x)\, dx = 0 = \int_0^1 E_m(x) E_n(x)\, dx. \tag{1.155}$$

Both the polynomials $B_m(x)$, $B_n(x)$ and the polynomials $E_m(x)$, $E_n(x)$ are *orthogonal* over the interval $(0, 1)$ (see Section 1.1.11), with uniform weight function. For a full discussion of Bernoulli and Euler polynomials, we refer the reader to Nörlund (1923) and Milne-Thompson (1933). Abramowitz and Stegun (1965) give an excellent summary.

1.1.10 Integral Transforms

The *exponential Fourier transform* [J. B. J. Fourier, 1768–1830]

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx \tag{1.156}$$

gives the *characteristic function* of a distribution.

The *Laplace transform*

$$L(p) = \int_{0}^{\infty} e^{-px} f(x)\, dx \tag{1.157}$$

(if it exists) yields the moment generating function $M(t)$ of a distribution with probability density function (pdf) $f(x)$ on the nonnegative real line by setting $t = -p$; that is, $M(t) = L(-t)$.

The *Mellin transform* [R. H. Mellin, 1854–1933] and its *inverse* are

$$H(s) = \int_{0}^{\infty} x^{s-1} f(x)\, dx, \tag{1.158}$$

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-s} H(s)\, ds. \tag{1.159}$$

If $f(x)$ is a pdf, then (1.158) gives the $(s-1)$th moment about the origin of a distribution on the nonnegative real line. Springer (1979) has demonstrated the key role of the Mellin transform and its inverse in the derivation of distributions of products, quotients, and other algebraic functions of independent random variables.

For a comprehensive coverage of these and other types of integral transforms, see Erdélyi et al. (1954, Vols. 1 and 2).

1.1.11 Orthogonal Polynomials

If the polynomial $P_r(x)$ of degree r is a member of a family of polynomials $\{P_j(x)\}$, $j = 0, 1, \ldots$, and

$$\int_{-\infty}^{\infty} w(x) P_m(x) P_n(x)\, dx = 0 \tag{1.160}$$

is satisfied whenever $m \neq n$, then the family of polynomials is said to be *orthogonal* with respect to the weight function $w(x)$. In particular cases, $w(x)$ may be zero outside certain intervals.

Two families of *orthogonal polynomials* have special importance in distribution theory. These are the Hermite polynomials and the generalized Laguerre polynomials. The *Hermite polynomials* have the weight function

$$w(x) = e^{-x^2/2}. \tag{1.161}$$

The rth Hermite polynomial is defined by

$$H_r(x) = (-1)^r e^{x^2/2} D^r (e^{-x^2/2}), \qquad r = 0, 1, \ldots, \tag{1.162}$$

It follows that

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= x, \\ H_2(x) &= x^2 - 1, \\ H_3(x) &= x^3 - 3x, \\ H_4(x) &= x^4 - 6x^2 + 3, \\ H_5(x) &= x^5 - 10x^3 + 15x, \end{aligned}$$

and generally

$$\begin{aligned} H_r(x) = x^r - \frac{r(r-1)}{1! \cdot 2} x^{r-2} + \frac{r(r-1)(r-2)(r-3)}{2! \cdot 2^2} x^{r-4} - \cdots \\ + (-1)^j \frac{r!}{(r-2j)! j! \cdot 2^j} x^{r-2j} + \cdots \end{aligned} \tag{1.163}$$

(cf. Section 1.1.7). The series terminates after $j = [r/2]$, where $[r]$ denotes the largest integer less than or equal to r.

The *generalized Laguerre polynomials* have the weight function

$$w(x) = \begin{cases} x^a e^{-x}, & x \geq 0, \qquad a > -1, \\ 0, & x < 0. \end{cases} \tag{1.164}$$

The rth generalized Laguerre polynomial of order a is

$$\begin{aligned} L_r^{(a)}(x) &= \sum_{j=0}^{r} (-1)^j \binom{r+a}{r-j} \frac{x^j}{j!} \\ &= \binom{r+a}{r} {}_1F_1[-r; a+1; x]. \end{aligned} \tag{1.165}$$

The recurrence formula

$$x L_r^{(a+1)}(x) = (x - r) L_r^{(a)}(x) + (a + r) L_{r-1}^{(a)}(x) \tag{1.166}$$

is useful in computation.

The *Jacobi, Chebyshev, Krawtchouk, and Charlier polynomials* are other families of orthogonal polynomials that are occasionally used in statistical theory. The weight function for the Jacobi polynomial $P_n^{(a,b)}(x)$ is

$$w(x) = \begin{cases} (1-x)^a(1+x)^b, & -1 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.167}$$

The other three families have the following weight functions:

Chebyshev polynomial $T_n(x)$:

$$w(x) = \begin{cases} (1-x^2)^{-1/2}, & -1 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.168}$$

Chebyshev polynomial $U_n(x)$:

$$w(x) = \begin{cases} (1-x^2)^{1/2}, & -1 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.169}$$

Krawtchouk polynomials:

$$w(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n, \\ 0, & \text{otherwise.} \end{cases} \tag{1.170}$$

Charlier polynomials:

$$w(x) = \begin{cases} e^{-\theta}\theta^x/x!, & x = 0, 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases} \tag{1.171}$$

Szegö (1939, 1959, 1967) is a standard work on orthogonal polynomials. Their properties are summarized in Abramowitz and Stegun (1965). Stuart and Ord (1987, Chapter 6) demonstrate some of their statistical uses.

1.1.12 Basic Hypergeometric Series

Heine's generalization of the hypergeometric series [H. E. Heine, 1821–1881] is known as a *basic hypergeometric series*, also as a *q-series* and as a

q-hypergeometric series; it is defined as

$$
\begin{aligned}
&{}_2\phi_1(a, b; c; q, z) \\
&\quad = 1 + \frac{(1-a)(1-b)z}{(1-c)(1-q)} + \frac{(1-a)(1-aq)(1-b)(1-bq)z^2}{(1-c)(1-cq)(1-q)(1-q^2)} + \cdots \\
&\quad = \sum_{j=0}^{\infty} \frac{(a;q)_j (b;q)_j z^j}{(c;q)_j (q;q)_j},
\end{aligned} \tag{1.172}
$$

where $|q| < 1$, $|z| < 1$; there are two numerator parameters and one denominator parameter. By $(a; q)_j$ we mean

$$
(a;q)_0 = 1, \qquad (a;q)_j = (1-a)(1-aq)\cdots(1-aq^{j-1}).
$$

Readers are WARNED that there are several differing notations for this expression in the literature; for example:

$(a; q)_j$	Slater (1966), Andrews (1986), Gasper and Rahman (1990),
$(a)_{q,j}$	Bailey (1935)
$[a]_j$	Jackson (1921)
$[a; q, j]$	Exton (1983)

[For a complete list of F. H. Jackson's numerous publications over the period 1904–1954, see Chaundy (1962).]

The Gaussian (basic) binomial coefficient is

$$
\begin{bmatrix} n \\ 0 \end{bmatrix}_q = 1, \qquad \begin{bmatrix} n \\ x \end{bmatrix}_q = \frac{(q;q)_n}{(q;q)_x (q;q)_{n-x}}.
$$

The definition of a general basic hypergeometric series (q-series) that was given in the second edition of this book was the one used by Bailey (1935) and Slater (1966):

$$
\begin{aligned}
{}_A\phi_B(a_1, \ldots, a_A; b_1, \ldots, b_B; q, z) &= {}_A\phi_B \left[\begin{matrix} a_1, \ldots, a_A; \\ b_1, \ldots, b_B \end{matrix} q, z \right] \\
&= \sum_{j=0}^{\infty} \frac{(a_1;q)_j \ldots (a_A;q)_j z^j}{(b_1;q)_j \ldots (b_B;q)_j (q;q)_j}.
\end{aligned}
$$

This is no longer in general use.

The publication of the book by Gasper and Rahman (1990) (G/R) on basic hypergeometric functions has led to the universal adoption of a new notation for

generalized basic hypergeometric series (q-series) in the mathematics and physics literature:

$$
{}_A\phi_B(a_1,\ldots,a_A;b_1,\ldots,b_B;q,z)
= \sum_{j=0}^{\infty} \frac{(a_1;q)_j \ldots (a_A;q)_j z^j}{(b_1;q)_j \ldots (b_B;q)_j (q;q)_j} \left[(-1)^j q^{\binom{j}{2}}\right]^{B-A+1}. \tag{1.173}
$$

The only difference between the two definitions is the additional factor

$$
[(-1)^j q^{j(j-1)/2}]^{B-A+1};
$$

there is no difference when $A = B + 1$. The very considerable advantage conferred by the use of the additional factor when $A \neq B + 1$ is that limiting forms of G/R q-series as parameters tend to zero are themselves q-series.

As $q \to 1$,

$$
\begin{aligned}
\frac{(q^a;q)_j}{(1-q)^j} &= \left(\frac{1-q^a}{1-q}\right)\left(\frac{1-q^{a+1}}{1-q}\right)\cdots\left(\frac{1-q^{a+j-1}}{1-q}\right) \\
&= (1+q+\cdots+q^{a-1})(1+q+\cdots+q^a)\cdots(1+q+\cdots+q^{a+j-2}) \\
&\to a(a+1)\cdots(a+j-1) = (a)_j,
\end{aligned}
$$

where $(a)_j$ is Pochhammer's symbol. It follows that as $q \to 1$ a generalized basic hypergeometric series tends to a generalized hypergeometric series:

$$
\begin{aligned}
&\lim_{q\to 1} {}_A\phi_B(q^{a_1},\ldots,q^{a_A};q^{b_1},\ldots,q^{b_B};q,(1-q)^{B+1-A}z) \\
&\quad = {}_AF_B[a_1,\ldots,a_A;b_1,\ldots,b_B;z].
\end{aligned}
$$

Heine's theorem

$$
{}_1\phi_0(a;-;q,z) = {}_1\phi_0(a;\ ;q,z) = \prod_{j=0}^{\infty} \frac{1-aq^jz}{1-q^jz} \tag{1.174}
$$

follows from the relationship

$$
(1-z)\,{}_1\phi_0(a;\ ;q,z) = (1-az)\,{}_1\phi_0(a;\ ;q,qz).
$$

When $a = q^{-k}$ and k is a positive integer, Heine's theorem gives the following q-series analog of the binomial theorem:

$$
\prod_{j=0}^{k-1}(1-q^{j-k}z) = {}_1\phi_0(q^{-k};\ ;q,z). \tag{1.175}
$$

Another consequence of Heine's theorem is

$$ {}_1\phi_0(a;\ ;q,z)\,{}_1\phi_0(b;\ ;q,az) = {}_1\phi_0(ab;\ ;q,z). \tag{1.176} $$

Letting $a \to 0$ gives

$$ {}_0\phi_0(-;-;q,z) = {}_0\phi_0(\ ;\ ;q,z) = \prod_{j=0}^{\infty}(1-q^j z)^{-1}; \tag{1.177} $$

that is,

$$ \begin{aligned} &1 + \frac{z}{1-q} + \frac{z^2}{(1-q)(1-q^2)} + \cdots + \frac{z^j}{(1-q)\cdots(1-q^j)} + \cdots \\ &\quad = (1-z)^{-1}(1-qz)^{-1}(1-q^2z)^{-1}\cdots. \end{aligned} \tag{1.178} $$

If z is replaced by $-z/a$ and $a \to \infty$, we obtain

$$ \begin{aligned} &1 + \frac{z}{1-q} + \frac{qz^2}{(1-q)(1-q^2)} + \cdots + \frac{q^{j(j-1)/2}z^j}{(1-q)\cdots(1-q^j)} + \cdots \\ &\quad = (1+z)(1+qz)(1+q^2z)\cdots. \end{aligned} \tag{1.179} $$

The general bilateral basic hypergeometric series with base q, $0 < q < 1$, r numerator parameters, and s denominator parameters is

$$ \begin{aligned} {}_r\psi_s(a_1,\ldots,a_r;b_1,\ldots,b_s;q,z) &= {}_r\psi_s\begin{bmatrix} a_1,\ldots,a_r; q, z \\ b_1,\ldots,b_s \end{bmatrix} \\ &= \sum_{j=-\infty}^{\infty} \frac{(a_1;q)_j \ldots (a_r;q)_j z^j}{(b_1;q)_j \ldots (b_s;q)_j}(-1)^{(s-r)j}\, q^{(s-r)j(j-1)/2}\, z^j \\ &= \sum_{j=-\infty}^{\infty} v_j\, z^j, \end{aligned} \tag{1.180} $$

where v_{j+1}/v_j is a rational function of q^j (Gasper and Rahman, 1990). It is assumed that each term in (1.180) is well defined; this is achieved when $q \neq 0$, $z \neq 0$, and $b_j \neq q^n$, where $n \in \mathbb{Z}^+$, $j = 0, 1, \ldots, s$.

1.2 PROBABILITY AND STATISTICAL PRELIMINARIES

1.2.1 Calculus of Probabilities

A *σ-field* is a collection $\mathcal{F}$ of subsets of a set Ω that contains the empty set (ϕ) as a member and is closed under countable unions and complements.

A *probability measure* P on a σ*-field* $\mathcal{F}$ of subsets of Ω is a function from $\mathcal{F}$ to the unit interval $[0, 1]$ such that $P(\Omega) = 1$ and the probability measure of a countable union of disjoint sets $\{E_i\}$ is equal to $\sum P(E_i)$.

A *probability space* is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, $\mathcal{F}$ is a σ*-field*, and P is a *probability measure* on $\mathcal{F}$.

For $\Pr(E)$ to be a probability measure, we require the following *probability axioms* to be satisfied:

1. $0 \leq \Pr(E) \leq 1$.

2. $\Pr(\Omega) = 1$.

3. If the events E_i are *mutually exclusive*, then $\Pr(\bigcup_i E_i) = \sum_i \Pr(E_i)$.

Probabilities defined in this way accord with the intuitive notion that the probability of an event E is the proportion of times that E might be expected to occur in repeated independent observations under specified conditions and that the probability of E therefore takes some value in the (closed) interval $[0, 1]$. The probability of an impossibility is taken to be zero, while the sum of the probabilities of all possibilities is deemed to be unity (the probability of a certainty). Given two events that cannot occur simultaneously, then the probability that one or other of them will occur is equal to the sum of their separate probabilities. If all the outcomes are equally likely, then

$$\Pr(A) = \frac{n(A)}{n(\Omega)} = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega}.$$

The compound event "either E_1 or E_2 or both" is called the *logical sum*, or *union*, of E_1 and E_2 and is written symbolically as $E_1 + E_2$ or $E_1 \bigcup E_2$. (The two names and symbols refer to the same concept.)

The compound event "both E_1 and E_2" is called the *logical product*, or *intersection*, of E_1 and E_2 and is written symbolically as $E_1 E_2$ or $E_1 \bigcap E_2$. (Again the two names and symbols refer to the same concept.)

If $\Pr(E_1 \cap E_2) = 0$, the events E_1 and E_2 are *mutually exclusive*.

These definitions can be extended to combinations of any number of events. Thus $E_1 + E_2 + \cdots + E_k$ or $\bigcup_{j=1}^{k} E_j$ means "at least one of $E_1, E_2, \ldots, E_k$," while $E_1 E_2 \cdots E_k$ or $\bigcap_{j=1}^{k} E_j$ means "every one of $E_1, E_2, \ldots, E_k$." By a natural extension we can form such compound events as $(E_1 \cup E_2) \cap E_3$, meaning "both E_3 and at least one of E_1 and E_2." By a further extension we can form compounds of enumerable infinities of events $\bigcup_{j=1}^{\infty} E_j$ and $\bigcap_{j=1}^{\infty} E_j$.

The following theorems hold:

1. $\Pr(\phi) = 0$ where ϕ is the empty set.

2. The event "negation of E" is called the *complement* of E and is often denoted by $\overline{E}$. We have $\Pr(E \cup \overline{E}) = 1$.

3. For any events E_1 and E_2,

$$\Pr(E_1) = \Pr(E_1 \cap E_2) + \Pr(E_1 \cap \overline{E}_2). \tag{1.181}$$

4. *De Morgan's laws* state that

$$\begin{aligned} \Pr(\overline{E}) &= 1 - \Pr(E), \\ \Pr(\overline{E_1 \cup E_2}) &= \Pr(\overline{E_1} \cap \overline{E_2}), \\ \Pr(\overline{E_1 \cap E_2}) &= \Pr(\overline{E_1} \cup \overline{E_2}). \end{aligned} \tag{1.182}$$

5. If $E \subset A$, then $\Pr(E) \leq \Pr(A)$.

6. For any events $E_1, E_2, \ldots, E_n$,

$$\Pr\left(\bigcup_{j=1}^{n} E_j\right) \leq \sum_{j=1}^{n} \Pr(E_j). \tag{1.183}$$

7. An important formula connecting probabilities of different but related events is

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2). \tag{1.184}$$

8. The following extension of this formula is known as *Boole's formula*:

$$\begin{aligned} \Pr\left(\bigcup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} \Pr(E_j) &- \sum\sum \Pr(E_{j_1} \cap E_{j_2}) \\ &+ \sum\sum\sum \Pr(E_{j_1} \cap E_{j_2} \cap E_{j_3}) - \cdots \\ &+ (-1)^{n-1} \Pr\left(\bigcap_{j=1}^{n} E_j\right), \end{aligned} \tag{1.185}$$

where a summation sign repeated m times means summation over all integers $j_1, j_2, \ldots, j_m$ subject to $1 \leq j_i \leq n$, $j_1 < j_2 < \cdots < j_m$. [The *inclusion–exclusion principle* is closely related to Boole's formula (see Section 10.2); it is important in the derivation of matching and occupancy distributions.]

9. The absolute values of the terms in (1.185) are nonincreasing. Boole's formula therefore enables bounds to be obtained for $\Pr(\bigcup_{j=1}^{n} E_j)$ by stopping at any two consecutive sets of terms. For example,

$$\sum_{j=1}^{n} \Pr(E_j) - \sum\sum \Pr(E_{j_1} \cap E_{j_2}) \leq \Pr\left(\bigcup_{j=1}^{n} E_j\right) \leq \sum_{j=1}^{n} \Pr(E_j). \tag{1.186}$$

10. If every pair of the events $E_1, E_2, \ldots, E_n$ is mutually exclusive, then Boole's formula becomes

$$\Pr\left(\bigcup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} \Pr(E_j). \tag{1.187}$$

[The mutually exclusive events $E_1, E_2, \ldots, E_n$ are said to be *exhaustive* if $\sum_{j=1}^{n} \Pr(E_j) = 1$.]

11. The *conditional probability* of event E_1 given that E_2 has occurred is denoted by $\Pr(E_1|E_2)$ and is given by

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}, \tag{1.188}$$

where $\Pr(E_2) > 0$; therefore

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2|E_1) = \Pr(E_2)\Pr(E_1|E_2). \tag{1.189}$$

12. More generally,

$$\Pr\left(\bigcap_{j=1}^{n} E_j\right) = \Pr(E_1)\Pr(E_2|E_1)\Pr(E_3|E_1 \cap E_2)\cdots\Pr\left(E_n\middle|\bigcap_{j=1}^{n-1} E_j\right). \tag{1.190}$$

13. If $\Pr(E_2|E_1) = \Pr(E_2)$, then E_2 is said to be *independent* of the event E_1 and (1.189) becomes

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2). \tag{1.191}$$

14. We say that n events are *mutually independent* if, for every subset $\{E_{j_1}, E_{j_2}, \ldots, E_{j_k}\}$, $k \leq n$,

$$\Pr(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_k}) = \prod_{i=1}^{k} \Pr(E_{j_i}). \tag{1.192}$$

15. If E_j is independent of $\bigcap_{i=1}^{j-1} E_i$ for all $j \leq n$ (this is certainly true if the n events are mutually independent), then (1.190) simplifies to

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_n) = \prod_{i=1}^{n} \Pr(E_j). \tag{1.193}$$

16. The *theorem of total probability* states that, if $E_1, E_2, \ldots, E_n$ are mutually exclusive and exhaustive, then

$$\Pr(A) = \sum_{j=1}^{n} \Pr(A|E_j) \Pr(E_j). \tag{1.194}$$

17. *Bayes's theorem* is an important consequence of (1.188) and (1.194) and is central to modern Bayesian methods of inference; see the next section.

1.2.2 Bayes's Theorem

Bayesian methods of inference involve the systematic formulation and use of Bayes's theorem. These approaches are distinguished from other statistical approaches in that, prior to obtaining the data, the statistician formulates *degrees of belief* concerning the possible models that may give rise to the data. These degrees of belief are regarded as probabilities.

Suppose that $\{M_1, M_2, \ldots, M_k\}$ is a mutually exclusive and exhaustive set of possible probability models for the experimental situation of interest, and suppose that $\{D_1, D_2, \ldots, D_r\}$ is the set of possible outcomes when the experiment is carried out. Also let

1. $\Pr(M_i), i = 1, \ldots, k$, be the probability that the correct model is M_i prior to learning the outcome of the experiment;
2. $\Pr(D_j), j = 1, \ldots, r$, be the probability that the result of the experiment is the outcome D_j;
3. $\Pr(D_j|M_i)$ be the probability that model M_i will produce the outcome D_j; and
4. $\Pr(M_i|D_j)$ be the probability that the model M_i is the correct model given that the experiment has had the outcome D_j.

Then, by the definition of conditional probability,

$$\Pr(M_i|D_j) \Pr(D_j) = \Pr(M_i \cap D_j) = \Pr(D_j|M_i) \Pr(M_i),$$

and by the theorem of total probability,

$$\Pr(D_j) = \sum_i \Pr(D_j|M_i) \Pr(M_i);$$

together these lead to the *discrete form of Bayes's theorem*

$$\Pr(M_i|D_j) = \frac{\Pr(D_j|M_i) \Pr(M_i)}{\sum_i \Pr(D_j|M_i) \Pr(M_i)}, \tag{1.195}$$

where

$\Pr(M_i)$ is termed the *prior probability* of the model M_i,

$\Pr(D_j|M_i)$ is termed the *likelihood of the outcome* D_j under the model M_i, and

$\Pr(M_i|D_j)$ is termed the *posterior probability* of the model M_i given that the outcome D_j has occurred.

It follows that

$$\frac{\Pr(M_i|D_j)}{\Pr(\overline{M_i}|D_j)} = \frac{\Pr(D_j|M_i)}{\Pr(D_j|\overline{M_i})} \frac{\Pr(M_i)}{\Pr(\overline{M_i})}. \tag{1.196}$$

Because the ratio $\Pr(A)/[1 - \Pr(A)]$ is called the *odds on* A, the discrete form of Bayes's theorem is sometimes rephrased as "posterior odds are equal to the likelihood ratio times the prior odds."

Suppose now that the models do not form an enumerable set but instead are indexed by a parameter θ. Let $p(\theta)$ be the prior probability density of the parameter θ, let $p(x|\theta)$ be the likelihood that the experiment will yield an observed value x for the random variable X given the value of the parameter θ, and let $p(\theta|x)$ be the posterior probability density of θ given that the experiment has yielded the observation x. Then

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{\displaystyle\int_\Theta p(x|\theta)\,p(\theta)\,d\theta}, \tag{1.197}$$

where Θ is the set of all possible values of the parameter θ; that is,

$$\int_\Theta p(\theta)\,d\theta = 1.$$

This is the *continuous form of Bayes's theorem*; it is sometimes summarized as "posterior density is proportional to likelihood times prior density."

The resolution of the problem of assigning a prior distribution to the parameter θ by the use of Bayes's postulate caused controversy. *Bayes's postulate* is, in brief, the assumption that, if there is no information to the contrary, then all prior probabilities are to be regarded as equal. This is known as the adoption of a *vague (diffuse, uninformative) prior*, in contradistinction to an *informative prior* that takes into account positive empirical or theoretical information concerning the distribution of θ.

Smith (1984) and Lindley (1990) have written helpful expository articles on Bayesian inference. Books that are written from a Bayesian standpoint include the seminal work by Box and Tiao (1973) and those by O'Hagan (1994) and Congdon (2003). Much of the literature on Bayesian methods is in edited volumes, for example, the Oxford University Press Series, *Bayesian Statistics,* edited by Bernado et al. (1992, 1996, 1999).

1.2.3 Random Variables

A *random variable* X is a mapping from a sample space into the real numbers, with the property that for every outcome there is an associated probability $\Pr[X \leq x]$ which exists for all real values of x. Random variables (rv's) will be denoted throughout this work by uppercase letters. Realized values of a rv will be denoted by the corresponding lowercase letter.

The *cumulative distribution function* (cdf) of X, often just called the *distribution function* (DF), is defined as $\Pr[X \leq x]$ and regarded as a function of x; it is customarily denoted by $F_X(x)$.

Clearly $F_X(x)$ is a nondecreasing function of x and $0 \leq F_X(x) \leq 1$. If $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$, then the distribution is *proper*. We shall be concerned only with proper distributions.

The study of distributions is essentially a study of cdf's. In all cases in these volumes the cdf belongs to one of two classes, discrete or continuous, or it can be constructed by mixing elements from the two classes.

For *discrete distributions* $F_X(x)$ is a step function with only an enumerable number of steps. If the height of the step at x_j is p_j, then

$$\Pr[X = x_j] = p_j.$$

We call p_j a *probability mass function* (pmf), and we say that its *support* is the set $\{x_j\}$. If the distribution is proper, $\sum_j p_j = 1$. Random variables belonging to this class are called *discrete random variables*. Most of the discrete distributions of interest are defined either on the nonnegative unit lattice $x = 0, 1, \ldots$ or on $1, 2, \ldots$. A discrete distribution is said to be *logconvex* when $p_x p_{x+2} / p_{x+1}^2 > 1$. It is *logconcave* when $p_x p_{x+2} / p_{x+1}^2 < 1$.

For *continuous distributions* $F_X(x)$ is absolutely continuous and can be expressed as an integral,

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx. \tag{1.198}$$

Any function $f_X(x)$ for which (1.198) holds for every x is a probability density function (pdf) of X. Random variables in this class are called *continuous random variables*.

When the subscripts for $f_X(x)$ and $F_X(x)$ are well understood, they are often dropped, provided that this does not cause confusion.

The above concepts can be extended to the *joint distribution* of a finite number of rv's $X_1, X_2, \ldots, X_n$. The *joint cumulative distribution function* is

$$\begin{aligned} \Pr\left[\bigcap_{j=1}^{n} (X_j \leq x_j)\right] &= F_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \\ &= F(x_1, x_2, \ldots, x_n). \end{aligned} \tag{1.199}$$

If $\Pr\left[\bigcap_{j=1}^{n}(X_j = x_j)\right]$ is zero except for an enumerable number of sets of values $\{x_{1_i}, x_{2_i}, \ldots, x_{n_i}\}$ and

$$\sum_i \Pr\left[\bigcap_{j=1}^{n}(X_j = x_{j_i})\right] = 1,$$

then we have a *discrete joint distribution*. For such distributions

$$\sum_i \Pr\left[\bigcap_{j=1}^{n-1}(X_j = x_j) \cap (X_n = x_{n_i})\right] = \Pr\left[\bigcap_{j=1}^{n-1}(X_j = x_j)\right], \tag{1.200}$$

where the summation is over all values of x_{n_i} for which the probability is not zero.

If $F(x_1, x_2, \ldots, x_n)$ is absolutely continuous, then

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}} \cdots \int_{-\infty}^{x_1} f(x_1, x_2, \ldots, x_n)\, dx_1 \cdots dx_n, \tag{1.201}$$

where $f(x_1, x_2, \ldots, x_n)$ [or strictly $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n)$] is the *joint probability density function* of $X_1, X_2, \ldots, X_n$. For a continuous joint distribution

$$\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_n = f(x_1, x_2, \ldots, x_{n-1}). \tag{1.202}$$

By repeated summation or integration, it is possible in principle to obtain the joint distribution of any subset of $X_1, X_2, \ldots, X_n$; in particular the distributions of each separate X_j can be found. These are called *marginal distributions*.

The *conditional joint distribution* of $X_1, X_2, \ldots, X_r$, given X_{r+1}, $X_{r+2}, \ldots, X_n$ [i.e., the joint distribution of the subset of the first r rv's in the case where particular values have been given to the remaining $(n - r)$ variables], is defined as

$$\Pr\left[\bigcap_{j=1}^{r}(X_j = x_j)\,\middle|\,\bigcap_{i=r+1}^{n}(X_i = x_i)\right] = \frac{\Pr\left[\bigcap_{j=1}^{n}(X_j = x_j)\right]}{\Pr\left[\bigcap_{j=r+1}^{n}(X_j = x_j)\right]} \tag{1.203}$$

(provided that $\Pr[\bigcap_{j=r+1}^{n}(X_j = x_j)] > 0$) for discrete distributions and by the pdf

$$f(x_1, x_2, \ldots, x_r | x_{r+1} \ldots, x_n) = \begin{cases} \dfrac{f(x_1, x_2, \ldots, x_n)}{f(x_{r+1} \ldots, x_n)}, & f(x_{r+1}, \ldots, x_n) > 0, \\ 0, & f(x_{r+1}, \ldots, x_n) = 0 \end{cases} \tag{1.204}$$

for continuous distributions. (The subscripts for F and f have been omitted for convenience in three of the above equations.)

Usually a distribution depends on one (or more) *parameters*, say θ. When we want to emphasize the dependence of the distribution on the value of θ, we write

$$F_X(x) = F_X(x;\theta) = F_X(x|\theta) \quad \text{and} \quad f_X(x) = f_X(x;\theta) = f_X(x|\theta). \tag{1.205}$$

1.2.4 Survival Concepts

Here we consider only lifetimes on the nonnegative integers with $\Pr[T = t] = p_t$, $t = 0, 1, \ldots$. We use T, not X, and t, not x, in this section to emphasize that time is assumed to be discrete.

Often $p_0 = 0$; the case $p_0 \neq 0$ corresponds to a nonzero probability of a death at birth (e.g., an egg that fails to hatch) or to a proportion p_0 of dud items.

Six representations characterize such distributions:

1. The pmf is

$$\Pr[T = t] = p_t, \qquad t = 0, 1, 2, \ldots, \tag{1.206}$$

2. The *survival function (survivor function)* is

$$S_0 = 1, \qquad S_t = 1 - \Pr(T < t) = \sum_{j \geq t} p_j, \qquad t = 1, 2, \ldots. \tag{1.207}$$

This is a nonincreasing step function that is left continuous since

$$\lim_{\epsilon \to 0} [S_{t-\epsilon} - S_t] = 0, \qquad \epsilon > 0, \qquad t \geq 0.$$

3. The *hazard function (failure rate, FR)* is

$$h_t = \frac{p_t}{\sum_{j \geq t} p_j} = \frac{S_t - S_{t+1}}{S_t}; \tag{1.208}$$

it is the probability that an item has survived to time t, given that it has survived to at least time t, that is, the amount of risk associated with an item at time t.

4. By analogy with the continuous case, the function

$$\Lambda_t = -\ln S_t \tag{1.209}$$

is called the *cumulative hazard function*.

5. In the discrete case summing the hazard function gives

$$H_t = \sum_{j=0}^{t} h_j. \tag{1.210}$$

This will be called the *accumulated hazard function*; it is a more tractable function for discrete data; see Kemp (2004). In general, $H_t \neq \Lambda_t$.

6. Following Kalbfleisch and Prentice (1980), Lawless (1982), and Leemis (1995), the *mean residual life function* is

$$L_t = E[(T - t)|T \geq t], \qquad t \geq 0. \tag{1.211}$$

Each of these six discrete lifetime functions can be stated uniquely in terms of each of the other functions. For example,

$$L_t = \frac{\sum_{j \geq t} j p_j}{\sum_{j \geq t} p_j} - t = \frac{\sum_{j > t} S_j}{S_t} = \sum_{j \geq t} \prod_{x=t}^{j} (1 - h_x)$$

$$= \sum_{j > t} e^{\Lambda_t - \Lambda_j} \quad = \sum_{j \geq t} \prod_{x=t}^{j} (1 - H_x + H_{x-1})$$

and

$$p_t = \left(1 - \frac{L_t}{1 + L_{t+1}}\right) \prod_{j=0}^{t-1} \left(\frac{L_j}{1 + L_{j+1}}\right), \qquad h_t = 1 - \frac{L_t}{1 + L_{t+1}},$$

$$S_t = \frac{S_{t-1} L_{t-1}}{1 + L_t} = \prod_{j=0}^{t-1} \left(\frac{L_j}{1 + L_{j+1}}\right), \qquad \Lambda_t = -\sum_{j=0}^{t-1} \ln \left(\frac{L_j}{1 + L_{j+1}}\right),$$

$$H_t = t + 1 - \sum_{j=0}^{t} \left(\frac{L_j}{1 + L_{j+1}}\right).$$

The above definitions lead by analogy with the continuous case to the following classes of discrete lifetime distributions:

IFR/DFR A discrete distribution with infinite support has a monotonically nondecreasing failure rate with time (IFR) or a monotonically nonincreasing failure rate with time (DFR) according as

$$\frac{p_{t+1}}{p_t} \gtrless \frac{p_{t+2}}{p_{t+1}} \tag{1.212}$$

(Gupta, Gupta, and Tripathi, 1997).

IFRA/DFRA A discrete lifetime distribution has an increasing or decreasing failure rate on average (IFRA or DFRA) according as

$$\frac{H_t}{t+1} \gtrless \frac{H_{t-1}}{t}, \qquad t \geq 1, \tag{1.213}$$

where H_t is the accumulated hazard function.

NBU/NWU A lifetime distribution is new better than used (NBU) or new worse than used (NWU) according as the conditional survival probability at time x for an item that has survived to time t is less (or greater) than the survival probability at time x for a new item, that is, according as

$$\frac{S_{t+x}}{S_t} \lessgtr S_x. \tag{1.214}$$

NBUE/NWUE A discrete lifetime distribution is new better than used in expectation (NBUE) or new worse than used in expectation (NWUE) according as

$$\frac{\sum_{j=0}^{\infty} S_{t+j}}{S_t} \lessgtr \sum_{j=0}^{\infty} S_j. \tag{1.215}$$

IMRL/DMRL An increasing mean residual life (IMRL) or a decreasing mean residual life (DMRL) is determined by

$$\begin{aligned} L_t - L_{t+1} &= \sum_{j>t} \left(\frac{S_j}{S_t} - \frac{S_{j+1}}{S_{t+1}} \right) = \sum_{j>t} \left[(h_j - h_t) \prod_{x=t+1}^{j-1} (1 - h_x) \right] \\ &\gtrless 0, \end{aligned} \tag{1.216}$$

that is, according as $h_j \gtrless h_t,\ j > t$.

The above definitions enable the interrelationships between these classes of discrete lifetime distributions to be stated as

$$\begin{aligned} \text{IFR/DFR} &\Rightarrow \text{IFRA/DFRA} \\ &\Rightarrow \text{NBU/NWU} \Rightarrow \text{NBUE/NWUE} \\ &\Rightarrow \text{DMRL/IMRL} \end{aligned}$$

See Kemp (2004) and the references therein for examples and proofs.

1.2.5 Expected Values

The *expected value* of a mathematical function $g(X_1, X_2, \ldots, X_n)$ of $X_1, X_2, \ldots, X_n$ is defined as

$$E[g(X_1, X_2, \ldots, X_n)] = \sum_i g(x_{1_i}, x_{2_i}, \ldots, x_{n_i}) \Pr \left[\bigcap_{j=1}^{n} (X_j = x_{j_i}) \right] \tag{1.217}$$

for discrete distributions and as

$$E[g(X_1, X_2, \ldots, X_n)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1 \ldots dx_n \quad (1.218)$$

for continuous distributions.

In particular, when $n = 1$,

$$E[g(X)] = \sum_x g(x) \Pr[X = x] \qquad \text{or} \qquad \int_{-\infty}^{\infty} g(x) f(x)\, dx. \quad (1.219)$$

If K is constant, then

$$E[K] = K,$$
$$E[Kg(X)] = KE[g(X)].$$

Also

$$E[g_1(X_1) + g_2(X_2)] = E[g_1(X_1)] + E[g_2(X_2)]. \quad (1.220)$$

More generally,

$$E\left[\sum_{j=1}^{M} K_j g_j(X_1, X_2, \ldots, X_n)\right] = \sum_{j=1}^{M} K_j E[g_j(X_1, X_2, \ldots, X_n)]. \quad (1.221)$$

These results apply to both discrete and continuous rv's. Conditional expected values are defined similarly, and formulas like (1.217) and (1.218) are valid for them.

The continuous rv's X_1, X_2 are said to be *independent*, if, for all real x_1, x_2, the events $(X_1 \le x_1)$, $(X_2 \le x_2)$ are independent.

The set $\{X_1, X_2, \ldots, X_n\}$ is a *mutually independent* set of discrete rv's if, for any combination of values $x_1, x_2, \ldots, x_n$ assumed by the rv's $X_1, X_2, \ldots, X_n$, $\Pr[X_1 = x_1, \ldots, X_n = x_n] = \Pr[X_1 = x_1] \ldots \Pr[X_n = x_n]$. In this case

$$E\left[\prod_{j=1}^{k} g_j(X_j)\right] = \prod_{j=1}^{k} E[g_j(X_j)]. \quad (1.222)$$

The *Shannon entropy* for a discrete rv is

$$H(X) = E\left[\log_2\left(\frac{1}{p_x}\right)\right] = -\sum_{\forall x} p_x \log_2(p_x). \quad (1.223)$$

For a continuous rv it is

$$H(X) = -\int_{\forall x} f(x) \log_2[f(x)]\, dx. \tag{1.224}$$

A stochastic process $\{X_n | n \geq 1\}$ with $E[|X_n|] < \infty$ for all n is a *martingale* process if

$$E[X_{n+1} | X_1, X_2, \ldots, X_n] = X_n. \tag{1.225}$$

It is a submartingale if $E[X_{n+1} | X_1, X_2, \ldots, X_n] \geq X_n$; it is a supermartingale if $E[X_{n+1} | X_1, X_2, \ldots, X_n] \leq X_n$.

1.2.6 Inequalities

1. *Cauchy–Schwartz Inequality* If X and Y are rv's such that $E[X^2]$ and $E[Y^2]$ exist, then

$$(E[XY])^2 \leq E[X^2]E[Y^2].$$

2. *Jensen's Inequality* If $E[X]$ exists and if $f(x)$ is a convex function, then

$$E[f(X)] \geq f(E[X]).$$

3. *Chebyshev Inequality* If $c > 0$ is a real number and if X is a rv such that $E[(X - c)^2]$ is finite, then

$$\Pr[|X - c|] \geq \epsilon] \leq \frac{1}{\epsilon^2} E[(X - c)^2]$$

for every $\epsilon > 0$.

4. *Bienaymé–Chebyshev Inequality* If $a > 0$ is a real number and if $E[|X|^r]$ is finite, then

$$\Pr[|X|] \geq a] \leq \frac{E[|X|^r]}{a^r}.$$

5. *Markov's Inequality* If X is a rv that takes only nonnegative values, then for all $a > 0$

$$\Pr[X \geq a] \leq \frac{E[X]}{a}.$$

Patel, Kapardia, and Owen (1976) have catalogued further inequalities, with references. Their Sections 2.1 and 2.2 give moment and Chebyshev-type inequalities; their Section 10.14 contains combinatorial inequalities.

1.2.7 Moments and Moment Generating Functions

Uncorrected Moments The expected value of X^r for r any real number is termed the rth *uncorrected (crude) moment* (alternatively the rth *moment about zero*):

$$\mu'_r(X) = \mu'_r = E[X^r]. \tag{1.226}$$

Unless otherwise stated, we will restrict consideration to integer values of r. The *(uncorrected) moment generating function* (mgf), if it exists (i.e., is finite), is

$$M_X(t) = E[e^{tX}] = 1 + \sum_{r\geq 1} \frac{\mu'_r t^r}{r!} \tag{1.227}$$

[when $M_X(t)$ exists for some interval $|t| < T$, where $T > 0$, then μ'_r is the coefficient of $t^r/r!$ in the Taylor expansion of $M_X(t)$]. If $\varphi(t)$ is the characteristic function of X (see Section 1.2.10), then $M(t) = \varphi(-it)$.

The first uncorrected moment μ'_1 is called the *mean* and is often written as μ.

The uncorrected moments can also be obtained from the cdf $F_X(x)$. If X is a continuous rv, then

$$E[X^r] = \int_0^\infty r x^{r-1}[1 - F_X(x) + (-1)^r F_X(-x)]\, dx. \tag{1.228}$$

If X is discrete, taking values $0, 1, \ldots, n$, where n is finite or infinite, then

$$E[X^r] = \sum_{x=0}^{n-1}[(x+1)^r - x^r][1 - F(x)]. \tag{1.229}$$

From the definition of the mgf,

$$M_{X+c}(t) = e^{ct} M_X(t). \tag{1.230}$$

Moreover, if X_1 and X_2 are independent rv's, then

$$M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t) \qquad \text{and} \qquad M_{X_1-X_2}(t) = M_{X_1}(t) M_{X_2}(-t). \tag{1.231}$$

If $X_1, X_2, \ldots, X_n$ are mutually independent rv's, then

$$M_{X_1+X_2+\cdots+X_n}(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t). \tag{1.232}$$

Moments about the Mean The rth moment about a constant a is $E[(X-a)^r]$. When $a = \mu$, we have the rth *moment about the mean* (also called the rth *central*

moment or the rth *corrected moment*),

$$\mu_r(X) = \mu_r = E[(X-\mu)^r] = E[(X - E[X])^r]. \tag{1.233}$$

The *central moment generating function*, if it exists, is

$$E[e^{(X-\mu)t}] = e^{-\mu t} M_X(t) = 1 + \sum_{r\geq 1} \frac{\mu_r t^r}{r!}. \tag{1.234}$$

The first central moment μ_1 is always zero. The second central moment μ_2 is called the *variance* of X [written as $\mathrm{Var}(X)$]. The positive square root of this quantity is called the *standard deviation* and is often denoted by the symbols $\sigma(X)$, or σ when no confusion is likely to arise. We have $\mu_2 \equiv \sigma^2$. The *coefficient of variation* (sometimes abbreviated to C. of V. or CV) is σ/μ.

Median The *median* of a distribution is the value of the variate that divides the total frequency into equal halves. For a continuous distribution this is unique. For a discrete distribution with $2N+1$ elements, the median is the value of the $(N+1)$th element; when there are $2N$ elements, there is ambiguity, and it is usual to define the median as the average of the Nth and $(N+1)$th elements.

Mode If $f_X(x)$ is a continuous and twice-differentiable pdf, then x is a *mode* if $df_X(x)/dx = 0$ and $d^2 f_X(x)/dx^2 < 0$. A discrete distribution has a mode at $X = x$ if

$$\Pr[X = x - c_1 - 1] < \Pr[X = x - c_1] \leq \cdots \leq \Pr[X = x]$$

and

$$\Pr[X = x] \geq \cdots \geq \Pr[X = x + c_2] > \Pr[X = x + c_2 + 1],$$

where $0 \leq c_1$, $0 \leq c_2$.

A distribution with only one mode is said to be *unimodal*; otherwise it is *multimodal*. A distribution with support $x \geq 0$ and a peak in frequency at $X = 0$ is sometimes said to have a *half mode* at $X = 0$ and to be *sesquimodal*. Abouammoh and Mashour (1981) have given necessary and sufficient conditions for a discrete distribution to be unimodal. Olshen and Savage (1970) have introduced the concept of α-unimodality for continuous distributions. For *discrete α-unimodality* (for discrete distributions), see Abouammoh (1987) and Steutel (1988).

Shape Commonly used indices of the shape of a distribution are the *moment ratios*. The most important are

$$\alpha_3(X) = \sqrt{\beta_1(X)} = \mu_3(\mu_2)^{-3/2} \quad \text{(an index of } \textit{skewness}\text{)}, \tag{1.235}$$

$$\alpha_4(X) = \beta_2(X) \quad = \mu_4(\mu_2)^{-2} \quad \text{(an index of } \textit{kurtosis}\text{)}, \tag{1.236}$$

and more generally

$$\alpha_r(X) = \mu_r(\mu_2)^{-r/2}. \tag{1.237}$$

The α and β notations are both in use. Note that these moment ratios have the same value for any linear function $A + BX$ with $B > 0$. When $B < 0$, the absolute values are not altered, but ratios of odd order have their signs reversed.

Moments about the Mean from Uncorrected Moments It is often convenient to calculate the central moments μ_r from the uncorrected moments and, less often, vice versa. Formulas for this involve the binomial coefficients:

$$\mu_r = E[(X - E[X])^r] = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu'_{r-j} \mu^j; \tag{1.238}$$

we refer the reader to Stuart and Ord (1987, p. 73) for further relevant formulas. In particular

$$\begin{aligned} \mu_2 &= \mu'_2 - \mu^2, \\ \mu_3 &= \mu'_3 - 3\mu'_2\mu + 2\mu^3, \\ \mu_4 &= \mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4. \end{aligned} \tag{1.239}$$

For the inverse calculation

$$\begin{aligned} \mu'_2 &= \mu_2 + \mu^2, \\ \mu'_3 &= \mu_3 + 3\mu_2\mu + \mu^3, \\ \mu'_4 &= \mu_4 + 4\mu_3\mu + 6\mu_2\mu^2 + \mu^4. \end{aligned} \tag{1.240}$$

The characterization of a distribution via its moment properties has been studied by several authors; see Johnson and Kotz (1990a,b) for a discussion of the methods that have been used and for new results.

Absolute Moments Besides the uncorrected and the central moments there are *absolute moments*, defined as the expected values of the absolute values (moduli) of various functions of X. Thus the rth *absolute moment about zero* of X is

$$\nu'_r(X) = E[|X|^r], \tag{1.241}$$

while the rth *absolute central moment* is

$$\nu_r(X) = E[|X - E[X]|^r]. \tag{1.242}$$

If r is even, $\nu'_r = \mu'_r$ and $\nu_r = \mu_r$, but not if r is odd. Whereas $\mu_1 = 0$, in general $\nu_1 > 0$. We call ν_1 the *mean deviation* of X.

Factorial Moments When studying discrete distributions, it is often advantageous to use the *factorial moments*. Those most commonly used are the descending

factorial moments. The rth *descending factorial moment* of X is the expected value of $X!/(X-r)!$:

$$\mu'_{[r]} = E\left[\frac{X!}{(X-r)!}\right]. \tag{1.243}$$

Readers are WARNED that there are other notations in use for the descending factorial moments. In the first edition of this book $\mu_{(r)}$ was used. Patel, Kapardia, and Owen (1976) use $\mu'_{(r)}$. The rth *ascending factorial moment* of X is $E[(X+r-1)!/(X-1)!]$.

Since $X!/(X-r)! = \sum_{j=0}^{r} s(r,j)X^j$, where $s(r,j)$ is the Stirling number of the first kind (see Section 1.1.3), we find that

$$\mu'_{[r]} = \sum_{j=0}^{r} s(r,j)\mu'_j. \tag{1.244}$$

Thus

$$\begin{aligned}
\mu'_{[1]} &= \mu, \\
\mu'_{[2]} &= \mu'_2 - \mu, \\
\mu'_{[3]} &= \mu'_3 - 3\mu'_2 + 2\mu, \\
\mu'_{[4]} &= \mu'_4 - 6\mu'_3 + 11\mu'_2 - 6\mu, \\
\mu'_{[5]} &= \mu'_5 - 10\mu'_4 + 35\mu'_3 - 50\mu'_2 + 24\mu, \\
\mu'_{[6]} &= \mu'_6 - 15\mu'_5 + 85\mu'_4 - 225\mu'_3 + 274\mu'_2 - 120\mu.
\end{aligned} \tag{1.245}$$

Similarly

$$X^r = \sum_{j=0}^{r} \frac{S(r,j)X!}{(X-r)!},$$

where $S(r,j)$ are the Stirling numbers of the second kind (see Section 1.1.3), and so

$$\mu'_r = \sum_{j=0}^{r} S(r,j)\mu'_{[j]}. \tag{1.246}$$

Hence

$$\begin{aligned}
\mu &= \mu'_{[1]}, \\
\mu'_2 &= \mu'_{[2]} + \mu, \\
\mu'_3 &= \mu'_{[3]} + 3\mu'_{[2]} + \mu, \\
\mu'_4 &= \mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu, \\
\mu'_5 &= \mu'_{[5]} + 10\mu'_{[4]} + 25\mu'_{[3]} + 15\mu'_{[2]} + \mu, \\
\mu'_6 &= \mu'_{[6]} + 15\mu'_{[5]} + 65\mu'_{[4]} + 90\mu'_{[3]} + 31\mu'_{[2]} + \mu.
\end{aligned} \tag{1.247}$$

The (descending) *factorial moment generating function* (fmgf), if it exists, is

$$E[(1+t)^X] = 1 + \sum_{r\geq 1} \frac{\mu'_{[r]} t^r}{r!}. \tag{1.248}$$

The relationship between the fmgf and the probability generating function enables the probabilities of a discrete distribution to be expressed in terms of its factorial moments; see Section 1.2.11.

Finite difference methods for obtaining the moments of a discrete distribution were discussed in a series of papers and letters in *The American Statistician* in the early 1980s; see, in particular, Johnson and Kotz (1981), Chan (1982), and Khatri (1983).

1.2.8 Cumulants and Cumulant Generating Functions

Cumulants The logarithm of the uncorrected moment generating function of X is the *cumulant generating function* (cgf) of X. If the mgf exists, then so does the cgf. The coefficient of $t^r/r!$ in the Taylor expansion of the cgf is the rth *cumulant* of X and is denoted by the symbol $\kappa_r(X)$ or, when no confusion is likely to arise, by κ_r:

$$K_X(t) = \ln M_X(t) = \sum_{r\geq 1} \frac{\kappa_r t^r}{r!} \tag{1.249}$$

(note that there is no term in t^0 in this equation).

We have

$$K_{X+a}(t) = at + K_X(t). \tag{1.250}$$

Hence for $r \geq 2$ the coefficients of $t^r/r!$ in $K_{X+a}(t)$ and $K_X(t)$ are the same; that is, the cumulants for $r \geq 2$ are not affected by the addition of a constant to X. For this reason the cumulants have also been called *seminvariants* or *half invariants*. Putting $a = -\mu$ shows that, for $r \geq 2$, the cumulants κ_r are functions of the central moments. In fact,

$$\begin{aligned} &\kappa_1 = \mu, \qquad \kappa_2 = \mu_2, \qquad \kappa_3 = \mu_3, \\ &\kappa_4 = \mu_4 - 3\mu_2^2, \qquad \kappa_5 = \mu_5 - 10\mu_3\mu_2. \end{aligned} \tag{1.251}$$

Note that the first three moments are equal to the first three cumulants.

Smith (1995) gave the following formulas connecting the uncorrected moments μ'_r to the cumulants for discrete distributions:

$$\mu'_r = \sum_{i=0}^{r-1} \binom{r-1}{i} \kappa_{r-i} \mu'_i \tag{1.252}$$

$$\kappa_r = \mu'_r - \sum_{i=1}^{r-1} \binom{r-1}{i} \kappa_{r-i} \mu'_i. \tag{1.253}$$

He gave analogous formulas for multivariate discrete distributions; see also Balakrishnan, Johnson, and Kotz (1998).

Let $X_1, X_2, \ldots, X_n$ be independent rv's and let $X = \sum_1^n X_j$; then, if the relevant functions exist,

$$K_X(t) = \sum_{j=1}^{n} K_{X_j}(t). \tag{1.254}$$

It follows from this equation that

$$\kappa_r \left(\sum_{j=1}^{n} X_j \right) = \sum_{j=1}^{n} \kappa_r(X_j) \quad \text{for all } r; \tag{1.255}$$

that is, the cumulant of a sum equals the sum of the cumulants, which makes the name "cumulant" appropriate.

Factorial Cumulants The logarithm of the (descending) fmgf is called the *factorial cumulant generating function* (fcgf). The coefficient of $t^r/r!$ in the Taylor expansion of this function is the rth *factorial cumulant* $\kappa_{[r]}$:

$$\ln G(1+t) = \sum_{r \geq 1} \frac{\kappa_{[r]} t^r}{r!}. \tag{1.256}$$

Formulas connecting $\{\kappa_r\}$ and $\{\kappa_{[r]}\}$ parallel those connecting $\{\mu'_r\}$ and $\{\mu'_{[r]}\}$:

$$\begin{aligned} \kappa_1 &= \kappa_{[1]} = \mu, \\ \kappa_2 &= \kappa_{[2]} + \mu, \\ \kappa_3 &= \kappa_{[3]} + 3\kappa_{[2]} + \mu, \\ \kappa_4 &= \kappa_{[4]} + 6\kappa_{[3]} + 7\kappa_{[2]} + \mu, \\ &\vdots \end{aligned} \tag{1.257}$$

Douglas (1980) has given a very full account of the relationships between the various types of moments and cumulants; see also Stuart and Ord (1987).

A sampling distribution arises as the distribution of some function of observations taken over all possible samples from a particular distribution according to a specified sampling scheme. The moment properties of a sampling distribution can be expressed in terms of symmetric functions of the observations, known as k-statistics; these were introduced by Fisher (1929). The expected value of the univariate k-statistic of order r is the rth cumulant; see Stuart and Ord (1987, Chapter 12).

1.2.9 Joint Moments and Cumulants

Moments of joint distributions, that is, quantities like $E\left[\prod_{j=1}^{n} X_j^{a_j}\right]$, are called *product moments (about zero)* and are denoted by $\mu'_{a_1 a_2 \ldots a_n}$. Quantities like

$$E\left[\prod_{j=1}^{n}(X_j - E[X_j])^{a_j}\right] = \mu_{a_1 a_2 \ldots a_n} \tag{1.258}$$

are called *central product moments* (sometimes *central mixed moments*).

The central product moment

$$\mu_{11} = E[(X_j - E[X_j])(X_{j'} - E[X_{j'}])] \tag{1.259}$$

is called the *covariance* of X_j and $X_{j'}$ and is denoted by $\mathrm{Cov}(X_j, X_{j'})$. The *correlation* between X_j and $X_{j'}$ is defined as

$$\rho(X_j X_{j'}) = \rho_{jj'} = \frac{\mathrm{Cov}(X_j, X_{j'})}{[\mathrm{Var}(X_j)\mathrm{Var}(X_{j'})]^{1/2}}. \tag{1.260}$$

[This is also sometimes written as $\mathrm{Corr}(X_j X_{j'})$.] It can be shown that $-1 \leq \rho_{jj'} \leq 1$. If X_j and $X_{j'}$ are mutually independent, then $\mathrm{Cov}(X_j X_{j'}) = 0 = \rho_{jj'}$; the converse is not necessarily true.

The *joint moment generating function* of $X_1, X_2, \ldots, X_n$ is defined as a function of n generating variables $t_1, t_2, \ldots, t_n$:

$$M_{X_1,\ldots,X_n}(t_1, t_2, \ldots, t_n) = M(t_1, t_2, \ldots, t_n) = E\left[\exp\sum_{j=1}^{n} t_j X_j\right]. \tag{1.261}$$

The *joint central moment generating function* is

$$E\left[\exp\left(\sum_{j=1}^{n} t_j(X_j - E[X_j])\right)\right] = \exp\left[-\sum_{j=1}^{n} t_j E[X_j]\right] M(t_1, t_2, \ldots, t_n). \tag{1.262}$$

The *joint cumulant generating function* is $\ln M_{X_1,\ldots,X_n}(t_1, t_2, \ldots, t_n)$. Use of these generating functions is similar to that for the single-variable functions.

The *regression function* of a rv X on m other random variables $X_1, X_2, \ldots, X_m$ is defined as

$$E[X|X_1, X_2, \ldots, X_m]; \tag{1.263}$$

it is an important tool for the prediction of X from $X_1, X_2, \ldots, X_m$. If (1.263) is a linear function of $X_1, X_2, \ldots, X_m$, then the regression is called *linear*

(or *multiple linear*). The variance of the conditional distribution of X given $X_1, X_2, \ldots, X_m$ is called the *scedasticity*. If $\mathrm{Var}(X|X_1, X_2, \ldots, X_m)$ does not depend on $X_1, X_2, \ldots, X_m$, then the conditional distribution is said to be *homoscedastic*.

Given that X_1 and X_2 are random variables, their joint distribution is determined by the distribution of X_1 together with the conditional distribution of X_2 given X_1. There has been much research during the past three decades on characterizations based on regression properties; see, for instance, Korwar (1975) and Papageorgiou (1985). Kotz and Johnson (1990) have provided a good review concerning characterizations for discrete distributions; see also Prasaka Rao (1992).

1.2.10 Characteristic Functions

The *characteristic function* (cf) of a continuous distribution is defined as

$$\varphi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx}, \, dF(x), \tag{1.264}$$

where $i = \sqrt{-1}$ and t is real. It is a complex-valued function. For a discrete distribution on the nonnegative integers, it is defined as

$$\varphi(t) = E[e^{itX}] = \sum_{j=0}^{\infty} e^{ijt} \Pr[X = j]. \tag{1.265}$$

The cf has great theoretical importance, particularly for continuous distributions. It is uniquely determined by the cdf and exists for all distributions. It satisfies (1) $\varphi(0) = 1$, (2) $|\varphi(t)| \leq 1$, and (3) $\varphi(-t) = \overline{\varphi(t)}$, where the overline denotes the complex conjugate.

If the distribution with cdf $F(x)$ has finite moments μ_r' up to order n, then

$$\mu_r' = i^r \varphi^{(r)}(0), \qquad 1 \leq r \leq n, \tag{1.266}$$

where $\varphi^{(r)}(0)$ is the rth derivative of $\varphi(t)$ evaluated at $t = 0$ and $i^2 = -1$.

The cf uniquely determines the pdf of a continuous distribution; we have

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) \, dt. \tag{1.267}$$

Gauss (1900) called this *Ein Schönes Theorem der Wahrscheinlichkeitsrechnung*.

The corresponding inversion formula for discrete distributions on the nonnegative integers is

$$\Pr[X = x] = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} \varphi(t) \, dt. \tag{1.268}$$

Lukacs (1970) gave further inversion formulas for continuous and for discrete distributions.

If X_1 and X_2 are independent rv's with cf's $\varphi_1(t)$ and $\varphi_2(t)$, respectively, then the cf of their sum $X_1 + X_2$ is the product of their cf's $\varphi_1(t)\varphi_2(t)$. Moreover the cf of their difference $X_1 - X_2$ is $\varphi_1(t)\varphi_2(-t)$.

Under very general conditions the cf for a limiting distribution is the limiting cf. If $\lim_{j\to\infty} \varphi_{X_j}(t) = \varphi_X(t)$, where $\varphi_X(t)$ is the cf of a rv with cdf F_X, then $\lim_{j\to\infty} F_{X_j}(x) = F_X(x)$.

A cf $\varphi(t)$ is said to be *infinitely divisible* if

$$\varphi(t) = [\varphi_n(t)]^n$$

for all positive integer n, where $\varphi_n(t)$ is itself a cf.

A cf $\varphi(t)$ is said to be *decomposable* if there are two nondegenerate cf's $\varphi_1(t)$ and $\varphi_2(t)$ such that

$$\varphi(t) = \varphi_1(t)\varphi_2(t).$$

A cf is said to be *stable* if $\varphi(a_1 t)e^{itb_1}\varphi(a_2 t)e^{itb_2} = \varphi(a_3 t)e^{itb_3}$, where $a_i > 0$ for $i = 1, 2, 3$.

Seminal references concerning cf's are Lukacs (1970, 1983) and Laha (1982).

1.2.11 Probability Generating Functions

Consider a nonnegative discrete rv X with nonzero probabilities only at nonnegative integer values. Let

$$p_j = \Pr[X = j], \qquad j = 0, 1, \ldots. \tag{1.269}$$

If the distribution is proper, then $\sum_{j=0} p_j = 1$, and hence $\sum_{j=0} p_j z^j$ converges for $|z| \le 1$. (This is also true when the distribution is not proper since then $0 < \sum_{j=0} p_j < 1$. However, we will be concerned only with proper distributions.)

The *probability generating function* (pgf) of the distribution with probability mass function (1.269) (or equivalently of the rv X) is defined as

$$G(z) = \sum_{j=0}^{\infty} p_j z^j = E[z^X]. \tag{1.270}$$

Although it would be logical to use the notation $G_X(z)$ for the pgf of X, we will in general suppress the suffix when it is clearly understood.

Probability generating functions have many properties:

1. The pgf is closely related to the cf; we have

$$\varphi_X(t) = E[e^{itX}] = G(e^{it}). \tag{1.271}$$

2. The pgf is defined by the probabilities; the uniqueness of a power series expansion implies that the pgf in turn defines the probabilities. We find that

$$p_j = \left[\frac{1}{j!}\frac{d^j G(z)}{dz^j}\right]_{z=0}, \qquad j = 0, 1, \ldots. \tag{1.272}$$

3. The rth moment, if it exists, is

$$\mu'_r = \sum_{j=0}^{\infty} j^r p_j = \left[\frac{d^r G(e^t)}{dt^r}\right]_{t=0}, \qquad r = 1, 2, \ldots. \tag{1.273}$$

4. The factorial moment generating function (fmgf) (if it exists) is given by

$$E[(1+t)^X] = G(1+t) = 1 + \sum_{r\geq 1} \frac{\mu'_{[r]} t^r}{r!}. \tag{1.274}$$

When the pgf is known, therefore, successive differentiation of the pgf enables the (descending) factorial moments to be obtained in a straightforward manner:

$$\mu'_{[r]} = \sum_{j=r}^{\infty} \frac{j!}{(j-r)!} p_j = \left[\frac{d^r G(z)}{dz^r}\right]_{z=1} = \left[\frac{d^r G(1+t)}{dt^r}\right]_{t=0}. \tag{1.275}$$

(The moments can be derived from the factorial moments as

$$\begin{aligned} \mu &= \mu'_{[1]}, \\ \mu_2 &= \mu'_{[2]} + \mu - \mu^2, \\ &\vdots \end{aligned}$$

see Section 1.2.7.)

5. The following relationships hold between the probabilities and the factorial moments in the case of a discrete distribution:

$$\Pr[X = x] = \sum_{j\geq x} (-1)^{x+j} \binom{j}{x} \frac{\mu'_{[j]}}{j!} = \sum_{r\geq 0} (-1)^r \frac{\mu'_{[x+r]}}{x!r!} \tag{1.276}$$

(Fréchet, 1940, 1943) and

$$\sum_{i\geq x} \Pr[X = i] = \sum_{j\geq x} (-1)^{x+j} \binom{j-1}{x-1} \frac{\mu'_{[j]}}{j!} \tag{1.277}$$

(Laurent, 1965).

6. If X_1 and X_2 are two independent rv's with pgf's $G_1(z)$ and $G_2(z)$, then the distribution of their sum $X = X_1 + X_2$ has the pgf

$$G(z) = E[z^X] = E[z^{X_1}]E[z^{X_2}] = G_1(z)G_2(z). \tag{1.278}$$

This is called the *convolution* of the two distributions. Let A, B, and C be the names of the distributions of X_1, X_2, and X; then we write $C \sim A^*B$.

7. More generally, let $X_1, X_2, \ldots, X_n$ be mutually independent rv's with pgf's $G_1(z), G_2(z), \ldots, G_n(z)$, respectively. Then the pgf of $X = \sum_{i=1}^{n} X_j$ is

$$G_X(z) = \prod_{j=1}^{n} G_j(z). \tag{1.279}$$

8. The pgf for the difference of two independent discrete rv's with pgf's $G_1(z)$ and $G_2(z)$ is

$$G_{X_i - X_j} = G_i(z)G_j(z^{-1}), \tag{1.280}$$

where the definition of a pgf is extended to encompass negative values of the variable.

9. The *joint probability generating function* of n discrete variables $X_1, X_2, \ldots, X_n$ is

$$G(z_1, z_2, \ldots, z_n) = E\left[\prod_{j=1}^{n} z_j^{X_j}\right], \tag{1.281}$$

where

$$\Pr\left[\bigcap_{j=1}^{n}(X_j = a_j)\right] = P_{a_1,a_2,\ldots,a_n}, \qquad a_j = 0, 1, 2, \ldots.$$

Let $r = \sum_{j=1}^{n} r_j$. Then the factorial moments of the distribution are given by

$$\mu'_{[r_1,r_2,\ldots,r_n]} = \left[\frac{\partial^r G(z_1, z_2, \ldots, z_n)}{\partial z_1^{r_1}\, \partial z_2^{r_2} \cdots \partial z_n^{r_n}}\right]_{z_1=z_2=\cdots=z_n=1}. \tag{1.282}$$

10. Relationships between probability generating functions and other generating functions are as follows:

Table 1.1 Relationships between Generating Functions

Probability generating function	$G(z)$
Characteristic function	$\varphi(t) = G(e^{it})$
Moment generating function	$M(t) = G(e^t)$
Central moment generating function	$e^{-\mu t}M(t) = e^{-\mu t}G(e^t)$
Factorial moment generating function	$G(1+t)$
Cumulant generating function	$K(t) = \ln G(e^t)$
Factorial cumulant generating function	$\ln G(1+t)$

An historical account of the use of pgf's in discrete distribution theory has been given by Seal (1949b).

1.2.12 Order Statistics

Let $X_1, X_2, \ldots, X_n$ be independent continuous rv's; then the jth *order statistic* $X_{j:n} \equiv X_{(j)}$, $j = 1, 2, \ldots, n$, is defined to be equal to the jth smallest of these. Here, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are called the *order statistics* corresponding to $X_1, X_2, \ldots, X_n$. Evidently

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

In particular,

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\} \qquad \text{and} \qquad X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}.$$

For a continuous distribution the probability of a *tie* (two equal values) is zero, and therefore the definition is unambiguous. Ties may, however, occur given a discrete distribution; unambiguity can be achieved here also, provided that we interpret "jth smallest value" to mean "not more than $j-1$ smaller values *and* not more than $n-j$ larger values."

The difference, $w = X_{(n)} - X_{(1)}$, is called the *range*. If n is odd, the "middle" value $X_{(n+1)/2}$ is called the *median*; see Section 1.2.7. When n is even, the median is not uniquely defined; often the jth order statistic observations are grouped, and reference is made to the *median class*. The closely related concepts of *hinges* and *fences* play a central role in exploratory data analysis (EDA); we refer the reader to Tukey (1977) and Emerson and Hoaglin (1983).

In the continuous case the pdf for the jth order statistic is

$$\begin{aligned} f_{X_{(j)}}(x) &= n\binom{n-1}{j-1}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x) \\ &= \frac{n!}{(j-1)!(n-j)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x); \end{aligned} \tag{1.283}$$

the cdf for the jth order statistic is

$$F_{X_{(j)}}(x) = \Pr[X_{(j)} \le x] = \Pr[j \text{ or more observations } \le x]$$
$$= \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}. \tag{1.284}$$

For the first order statistic, the pdf and the cdf are, respectively,

$$f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} f(x) \qquad \text{and} \qquad F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n.$$

For the nth order statistic they are

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1} f(x) \qquad \text{and} \qquad F_{X_{(n)}}(x) = [F(x)]^n.$$

In the discrete case, equation (1.284) holds and the pmf is

$$f_{X_{(j)}}(x) = F_{X_{(j)}}(x) - F_{X_{(j)}}(x - 1).$$

There has been much work on characterizations of distributions using properties of their order statistics; see, for instance, Arnold and Meeden (1975), Shah and Kabe (1981), Hwang and Lin (1984), Khan and Ali (1987), and Lin (1987). David's (1981) book provided, at the time, an encyclopedic coverage of properties, statistical techniques, characterizations, and applications relating to order statistics from both continuous and discrete distributions. Harter's (1988) paper gave definitions and examined the history and the importance of order statistics. Balakrishnan (1986) extended previous work on recurrence relations for single and product moments of order statistics in both the continuous and the discrete case.

Arnold, Balakrishnan, and Nagaraja (1992) have written a good introduction to the subject, including a chapter on discrete order statistics. Nagaraja's (1990) book also gives a good account of work on order statistics for discrete distributions, particularly concerning characterizations of the geometric distribution, and extreme order statistics. Estimation methods based on order statistics are discussed by Balakrishnan and Cohen (1991).

Nagaraja's (1992) survey of results on order statistics for random samples from discrete distributions includes discussion and rejoinder. It reviews finite sample theory, characterization results, and asymptotic results and discusses applications to testing and selection. The emphasis is on order statistics from the discrete uniform, geometric, binomial, and multinomial distributions.

1.2.13 Truncation and Censoring

If values of the rv's $X_1, X_2, \ldots, X_n$ in a given region $\overline{R}$ are excluded, then the joint cdf of the variables is

$$F(x_1, x_2, \ldots, x_n|R) = \Pr\left[\bigcap_{j=1}^{n}(X_j \le x_j)|(X_1, \ldots, X_n) \subset R\right]$$
$$= \frac{\Pr\left[\bigcap_{j=1}^{n}(X_j \le x_j) \cap (X_1, \ldots, X_n) \subset R\right]}{\Pr[(X_1, X_2, \ldots, X_n) \subset R]}, \qquad (1.285)$$

where R is the complement of $\overline{R}$ and comprises all the points that are not truncated. The distribution given by (1.285) is called a *truncated distribution*. All the quantities on the right-hand side of the equation can be calculated from the (unconditional) joint cdf $F(x_1, x_2, \ldots, x_n)$.

We shall usually be concerned with truncated distributions of single variables, for which R is a finite or infinite interval. If R is a finite interval with end points a and b inside the range of values taken by X, the distribution is *doubly truncated* (or *left-and-right truncated*, or *truncated below and above*); a and b are the *truncation points*.

If R consists of all values greater than a, then the distribution is said to be *truncated from below*, or *left truncated*; if R consists of all values less than b, then the distribution is said to be *truncated from above*, or *right truncated*. (The same terms are also used when R includes values equal to a or to b, as the case may be.)

If X' is a rv having a distribution formed by doubly truncating the distribution of a continuous random variable X, then the pdf of X', in terms of the pdf and cdf of X, is

$$f_{X'}(x') = \frac{f_X(x')}{F_X(b) - F_X(a)}, \qquad a \le x' \le b. \qquad (1.286)$$

A distinction needs to be made between truncation of a distribution and truncation of a sample. Truncation of a distribution occurs when a range of possible variate values either is ignored or is impossible to observe.

Truncation of a sample is commonly called *censoring*. Sometimes censoring is with respect to a fixed variate value; for instance, in a survival study it may be impossible within a limited time span to ascertain the length of survival of all the patients. When the existence of observations outside a certain range is known but their exact value is unknown, the form of censoring is known as *type I censoring*.

When a predetermined number of order statistics are omitted from a sample, the form of censoring is known as *type II*. If the ℓ smallest values $X_{(1)}, \ldots, X_{(\ell)}$ are omitted, it is *censoring from below*, or *left censoring*; when the m largest values are omitted, it is *censoring from above*, or *right censoring*. If both sets of order statistics are omitted, we have *double censoring*.

The term "truncation" is used in a different sense in sequential analysis, where it refers to the imposition of a cutoff point leading to cessation of the sequential sampling process before a decision has been reached.

1.2.14 Mixture Distributions

A *mixture of distributions* is a superimposition of distributions with different functional forms or different parameters, in specified proportions.

Suppose that

$$\{F_j(x_1, x_2, \ldots, x_n)\}, \qquad j = 0, 1, 2, \ldots, m,$$

represents a set of different (proper) cdf's, where m is finite or infinite. Suppose also that $a_j \geq 0$, $\sum_{j=0}^{m} a_j = 1$. Then

$$F(x_1, x_2, \ldots, x_n) = \sum_{j=0}^{m} a_j F_j(x_1, x_2, \ldots, x_n) \tag{1.287}$$

is a proper cdf. This mixture of the distributions $\{F_j\}$ is *finite* or *infinite* according as m is finite or infinite.

For many of the mixture distributions in this book, the distributions to be mixed all have cdf's of the same functional form but are dependent on some parameter Θ. If Θ itself has a discrete distribution with pmf $\Pr[\Theta = \theta_j] = p_j$, $j = 0, 1, \ldots$, then the resultant mixture has the cdf

$$\sum_{j \geq 0} p_j F(x_1, x_2, \ldots, x_n | \theta_j).$$

If Θ has a continuous distribution with cdf $H_\Theta(\theta)$, then the resultant mixture has cdf

$$\int F(x_1, \ldots, x_n | \theta)\, dH_\Theta(\theta),$$

where integration is over all values of θ. In either case (discrete or continuous distribution of Θ) we have

$$F(x_1, x_2, \ldots, x_n) = E_\Theta[F_j(x_1, x_2, \ldots, x_n | \theta)], \tag{1.288}$$

where the expectation is with respect to Θ. We call $F(x_1, x_2, \ldots, x_n)$ the *mixture distribution* and say that the distribution of Θ is the *mixing distribution*.

More generally the distribution of $X_1, X_2, \ldots, X_n$ may depend on several parameters $\Theta_1, \ldots, \Theta_k, \Theta_{k+1}, \ldots, \Theta_m$, where $\Theta_1, \ldots, \Theta_k$ vary and $\Theta_{k+1}, \ldots, \Theta_m$ are constant. The mixture distribution then has the cdf

$$F(x_1, x_2, \ldots, x_n | \theta_{k+1}, \ldots, \theta_m) = E_{\Theta_1, \ldots, \Theta_k}[F(x_1, x_2, \ldots, x_n | \theta_1, \ldots, \theta_m)]. \tag{1.289}$$

Note that the parameters $\theta_1, \ldots, \theta_k$ do not appear in the mixture distribution because they have been summed out (for a discrete mixture) or integrated out (for a continuous distribution). The parameters $\Theta_{k+1}, \ldots, \Theta_m$ have not been eliminated in this way.

Mixtures of discrete distributions are dealt with in depth in Chapter 8.

1.2.15 Variance of a Function

Given the moments of a rv X, suppose that we wish to obtain the moments of a mathematical function of X, that is, of $Y = h(X)$.

If exact expressions can be obtained and are convenient to use, this should of course be done. However, in some cases it may be necessary to use approximate methods. One approximate method is to expand $h(X)$ as a Taylor series about $E[X]$:

$$Y = h(E[X]) + (X - E[X])h'(E[X]) + (X - E[X])^2 \frac{h''(E[X])}{2!} + \cdots. \tag{1.290}$$

Then, taking expected values of both sides of (1.290),

$$E[Y] = h(E[X]) + \mathrm{Var}(X)\frac{h''(E[X])}{2} + R \tag{1.291}$$

$$\approx h(E|X]) + \mathrm{Var}(X)\frac{h''(E[X])}{2}. \tag{1.292}$$

Also

$$\{Y - h(E[X])\}^2 \approx \{(X - E[X])h'(E[X])\}^2,$$

whence

$$\mathrm{Var}(Y) \approx \{h'(E[X])\}^2 \mathrm{Var}(X). \tag{1.293}$$

This method of approximation has been used widely under a number of different names, for example, the *delta method*, the *method of statistical differentials*, and the *propagation of error*. The method assumes that the expected value of the remainder term R in (1.291) is small and that the higher order central moments do not become large; otherwise the outcome may be very unreliable. The method will usually be more reliable for small values of $\mathrm{Var}(X)$.

Equation (1.293) can be made the basis for an approximate *variance-stabilizing transformation*. If $\mathrm{Var}(X)$ is a function $g(E[X])$ of $E[X]$, then $\mathrm{Var}(Y)$ might be expected to be more nearly constant if $[h'(E[X])]^2 g(E[X])$ is a constant.

This will be so if

$$h(X) \propto \int^X \frac{dt}{[g(t)]^{1/2}}. \tag{1.294}$$

This suggests the use of $Y = h(X)$, with $h(X)$ satisfying (1.294), as a variance-stabilizing transformation. Often such transformations are also effective as *normalizing transformations* in that the distribution of Y is nearer to normality than that of X; see Johnson et al. (1994, Chapter 12).

When $Y = X^a$, the method gives

$$\begin{aligned} E[X^a] &\approx \mu^a \left[1 + \frac{a(a-1)\sigma^2}{2\mu^2}\right], \\ \mathrm{Var}(X^a) &\approx \mu^{2a-2} a^2 \sigma^2, \end{aligned} \tag{1.295}$$

where $\mu = E[X]$ and $\sigma^2 = \mathrm{Var}(X)$. The coefficient of variation of X^a is therefore very approximately $|a|(\sigma/\mu)$.

There are exact methods for obtaining the moments of a product of two rv's. For the quotient of two nonnegative rv's the delta method gives

$$\begin{aligned} E\left[\frac{X_1}{X_2}\right] &\approx \frac{\xi_1}{\xi_2}\left[1 + \frac{\mathrm{Var}(X_2)}{\xi_2^2} - \frac{\mathrm{Cov}(X_1, X_2)}{\xi_1 \xi_2}\right], \\ \mathrm{Var}\left(\frac{X_1}{X_2}\right) &\approx \frac{\xi_1^2}{\xi_2^2}\left[\frac{\mathrm{Var}(X_1)}{\xi_1^2} - \frac{2\,\mathrm{Cov}(X_1, X_2)}{\xi_1 \xi_2} + \frac{\mathrm{Var}(X_2)}{\xi_2^2}\right], \end{aligned} \tag{1.296}$$

where ξ_1 and ξ_2 are the expected values of X_1 and X_2, respectively.

For further discussion and use of the delta method see Stuart and Ord (1987, Chapter 10).

1.2.16 Estimation

Since the publication of the first edition of this book an immense amount of research has been devoted to statistical estimation, both to theoretical developments and to practical aspects of inferential procedures. Computer-intensive methods are now widely used. The books by Cox and Hinkley (1974) and Barnett (1999) are valuable for their lucid discussions of the many approaches to inference. Desmond and Godambe (1998) and Doksum (1998) give good introductory accounts with references.

Readers of this book will occasionally meet references to results from goodness of fit, hypothesis testing, and decision theory. Kocherlakota and Kocherlakota (1986) concentrated on goodness of fit tests for discrete distributions. The many well-written books on these topics also include D'Agostino and Stephens (1986), Rayner and Best (1989) (*goodness of fit*), DeGroot (1970), Cox and Hinkley (1974), Lehmann (1986) (*hypothesis testing*), and Berger (1985) (*Bayesian decision theory*).

Bayesian methods of inference continue to receive special attention and are widely applied. Many aspects of Bayesian statistics are discussed in the volumes edited by Bernado et al. (1988, 1992, 1996, 1999). Maritz and Lwin (1989) were concerned with empirical Bayes methods. O'Hagan (1994) and Congdon (2003) have useful bibliographies.

In this preliminary chapter we sketch only briefly some of the basic concepts and methods in classical estimation theory.

Parameters and Statistics A cdf that depends on the values of a finite number of quantities $\theta_1, \theta_2, \ldots, \theta_m$ (called *parameters*) is written

$$F(X_1, X_2, \ldots, X_n | \theta_1, \theta_2, \ldots, \theta_m).$$

Often we want to estimate the values of these parameters. This is done using functions of the random variables $T_j \equiv T_j(X_1, X_2, \ldots, X_n)$ called *statistics*. When a statistic T_j is used to estimate a parameter θ_j, it is called an *estimator* of θ_j. An *estimate* is a realized value of an estimator for a particular sample of data.

Properties of Estimators

1. A statistic T_j is said to be an *unbiased estimator* of the parameter θ_j if $E[T_j] = \theta_j$. If $E[T_j] \neq \theta_j$, the estimator is *biased*. The bias is $B(T_j) = E[T_j - \theta_j]$.
2. All distributions with finite means and variances possess unbiased estimators of their means and variances, namely,

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \text{and} \qquad s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}, \tag{1.297}$$

   respectively, where n is the sample size.
3. If T_j and T_j^* are both unbiased estimators of the same parameter θ_j, then any weighted average $wT_j + (1-w)T_j^*$ is also an unbiased estimator of θ_j. An estimator is said to be *asymptotically unbiased* if $\lim_{n\to\infty} E[T_j] = \theta_j$.
4. The *relative efficiency* of two unbiased estimators is measured by the inverse ratio of their variances, that is, $\mathrm{Var}(T_j^*)/\mathrm{Var}(T_j)$ measures the efficiency of T_j relative to T_j^*. Comparisons of the efficiencies of biased estimators are often made on the basis of their mean-squared errors. The *mean-squared error* is defined to be

$$E[(T_j - \theta_j)^2] = \mathrm{Var}(T_j) + (E[T_j] - \theta_j)^2. \tag{1.298}$$

5. If a measure of overall efficiency is required when several parameters $\theta_1, \theta_2, \ldots, \theta_m$ are being estimated by the unbiased estimators $T_1, T_2, \ldots, T_m$, respectively, then the *generalized variance* may be used. This is a determinant in which the element in the jth row and jth column is

$$\mathrm{Cov}(T_j, T_{j'} | \theta_1, \theta_2, \ldots, \theta_m).$$

   (Comparisons of the generalized variances of biased estimators are only meaningful if the biases are small enough to be neglected.)
6. A *consistent estimator* is one for which

$$\lim_{n\to\infty} \Pr\left[|T_j - \theta_j| \geq c\right] = 0 \tag{1.299}$$

for all positive c. If T_j is unbiased, then it will also be consistent provided that

$$\lim_{n\to\infty} \mathrm{Var}(T_j) = 0.$$

Consistency is an asymptotic property.

7. A *minimum-variance unbiased estimator* (MVUE) T_j of θ_j is an unbiased estimator of θ_j with a variance that is not greater than that of any other unbiased estimator of θ_j. If T_j is an unbiased estimator of θ_j, then the Cramér–Rao theorem states that the variance of T_j satisfies the *Cramér–Rao inequality*

$$\mathrm{Var}(T_j) \geq \frac{1}{nE[\{\partial \ln f(x)/\partial\theta_j\}^2]} \tag{1.300}$$

An MVUE may or may not, however, attain the Cramér–Rao lower bound.

8. The *efficiency of an unbiased estimator* is the ratio of its variance to the Cramér–Rao lower bound. An estimator is called an *efficient estimator* if this ratio is unity; it is said to be an *asymptotically efficient estimator* if this ratio tends to unity as the sample size becomes large.

9. A *sufficient estimator* is one that summarizes from the sample of observations all possible information concerning the parameter; that is, no other statistic formed from the observations provides any more information. Such a statistic will exist if and only if the likelihood (see below) can be factorized into two parts, one depending only on the statistic and the parameters and the other depending only on the sample observations. If an unbiased estimator has a variance equal to the Cramér–Rao lower bound, then it must be a sufficient estimator.

10. A family of distributions dependent on a vector of parameters Θ is said to be *complete* if $E_\Theta[h(T)] = 0$ for all values of the parameters implies that $\Pr[h(T) = 0] = 1$ for all Θ, where $h(T)$ is a function of the observations and $E_\Theta[\cdot]$ denotes expectation with respect to the distribution with parameters Θ.

Stuart and Ord (1987) have given a careful and very full account of estimation principles as well as details concerning the major types of estimation procedures.

Estimation Methods

1. The method of *maximum likelihood* is widely advocated. If observed values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$, then their likelihood is

$$L(x_1, x_2, \ldots, x_n) = \Pr\left[\bigcap_{j=1}^{n} (X_j = x_j | \theta_1, \theta_2, \ldots, \theta_m)\right] \tag{1.301}$$

for discrete distributions and

$$L(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_m) \tag{1.302}$$

for continuous distributions. In either case the values $\hat{\theta}_1 = T_1$, $\hat{\theta}_2 = T_2, \ldots,$ $\hat{\theta}_m = T_m$ that maximize the likelihood are called *maximum-likelihood estimators* (MLEs). (Note that the $\hat{\theta}_j$'s are random variables.) If $X_1, X_2, \ldots, X_n$ are mutually independent and have identical distributions, then under rather general conditions

(i) $\lim_{n\to\infty} E[\hat{\theta}_j|\theta_1, \theta_2, \ldots, \theta_m] = \theta_j$, $j = 1, 2, \ldots, m$, and

(ii) asymptotic estimates of the variances and covariances of the $\hat{\theta}_j$'s are given by the corresponding elements in the inverse of the information matrix evaluated at the maximum-likelihood values.

A MLE may or may not be unique, may or may not be unbiased, and need not be consistent. Nevertheless, MLEs possess certain attractive properties. Under certain mild regularity conditions they are asymptotically MVUEs and are also asymptotically normally distributed. The maximum-likelihood estimation method yields sufficient estimators whenever they exist. Also, if $\hat{\theta}$ is a MLE of θ and if $h(\cdot)$ is a function with a single-valued inverse, then the MLE of $h(\theta)$ is $h(\hat{\theta})$.

Maximizing the likelihood can usually be achieved by solving the equations

$$\frac{\partial L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_m)}{\partial \theta_i} = 0, \qquad i = 1, 2, \ldots, m; \tag{1.303}$$

these equations are called the *maximum-likelihood equations*. They are often intractable and require iteration (e.g., by the Newton–Raphson method) for their solution. The almost universal accessibility to cheap computing power has led to the development of a number of computer routines for maximum-likelihood estimation. Also the log-likelihood is negative, and so maximizing the likelihood is equivalent to minimizing the absolute value of the log-likelihood; this can be achieved by means of a computer optimization routine. The leading computer packages supply maximum-likelihood and suitable function optimization routines.

Reparameterization, where feasible, so that the new parameters are orthogonal, has been advocated by a number of authors; see, for example, Cox and Reid (1987), Ross (1990), and Willmot (1988b).

2. The *method of moments* usually requires less onerous calculations than maximum-likelihood estimation, although the method cannot be guaranteed to give explicit estimators. The method is based on equating the first k *uncorrected sample moments* about zero, $m'_r = n^{-1}\sum_{j=1}^{n} x_j^r$, $r = 1, \ldots, k$, to the corresponding theoretical expressions for μ'_r, where k is the number of unknown parameters. If preferred, the first k *central sample moments* $m_r = n^{-1}\sum_{j=1}^{n}(x_j - \overline{x})^r$, $r = 1, \ldots, k$, can be equated to the corresponding expressions for the central

moments μ_r, or the first k factorial sample moments $m'_{[r]}$, $r = 1, \ldots, k$, can be set equal to $\mu'_{[r]}$; the three procedures give identical estimators. The equation obtained by equating the *sample mean* to the theoretical mean is called the *first-moment equation*.

The higher sample moments generally have large variances, however. This has led to the use of methods based on the quantiles (for continuous distributions) and on the mean and lowest observed frequencies (for discrete distributions). An alternative approach is to solve equations that are approximations to the maximum-likelihood equations. This approach has been discussed by A. W. Kemp (1986) for discrete distributions.

When the method of moments or a similar method leads to explicit estimators, they can be used to provide initial estimates for maximum-likelihood estimation.

The desirable properties of MLEs have led to the development of a number of variants of maximum-likelihood estimation, especially for situations where a family of distributions is to be fitted to data, and hence there are a large number of parameters to be estimated:

Generalized MLEs exist and have near-optimal properties in cases where MLEs do not exist. In most other cases, though, the two methods give identical results (see, e.g., Weiss (1983).

Modified maximum-likelihood estimation is used particularly for censored data. In many situations where maximum-likelihood estimation requires iteration, modified maximum-likelihood estimation gives explicit estimators (see, e.g., Tiku (1989).

Penalized maximum-likelihood estimation is used in curve estimation. It involves a sacrifice of efficiency in order to achieve smooth fits (see, e.g., Silverman (1985).

Partial maximum-likelihood estimation was introduced by Cox (1975) for analyzing regression models involving explanatory variables as a way to reduce the number of nuisance parameters (see, e.g., Kay (1985).

Conditional, marginal, and *profile* likelihood procedures are methods somewhat similar to partial maximum-likelihood estimation; they are collectively described as *pseudolikelihood* methods. They have probabilistic interpretations [see, e.g., Kalbfleisch (1986) and Barndorff-Neilsen (1991)].

Quasi-likelihood estimation is a nonlinear weighted least-squares method for generalized linear models. It is based on families of linear-exponential distributions. Only second-moment assumptions are used, and hence linear exponentiality does not necessarily hold. The method yields equations similar to (1.303) and can be useful for situations with overdispersion (see, e.g., McCullagh (1991).

Interval Estimates Often interval estimates of a parameter θ are wanted. These have the property

$$\Pr(\theta \in \{a \leq \theta \leq b\}) = 1 - \alpha,$$

where a and b are functions of a statistic (or statistics) derived from the sample data. The confidence probability (confidence level) is the probability that the (random) interval covers the true (fixed) parameter θ.

Confidence intervals are often symmetric but need not be so. There are various ways of determining a and b; inverting a significance test may be a convenient method. Intervals based on minimal sufficient statistics, where they exist, are generally preferred.

For discrete distributions, when a particular value of α is specified, it is often difficult (or maybe impossible) to construct an interval where the confidence probability is exactly $1 - \alpha$; This is due to the discrete nature of the statistic(s) on which the interval is based (see, e.g., Agresti and Coull, 1998). References to ingenious methods for obtaining "exact confidence intervals" for discrete distributions are given in later chapters of this book.

Robinson's (1982) introductory article gives useful references. Hahn and Meeker (1991) discuss various kinds of confidence intervals, prediction intervals, and tolerance intervals as well as their applications. They devote a chapter each to the binomial and Poisson distributions.

1.2.17 General Comments on the Computer Generation of Discrete Random Variables

The methods of generating rv's that are discussed in this section depend on an infinite sequence of random numbers $\{U_i\}$ uniformly distributed on [0, 1]. A suitable sequence is customarily generated by a computer, using a *pseudo-random-number algorithm*, that is, an algorithm that generates a deterministic stream of numbers that appears to have the same relevant statistical properties as a sequence of truly random numbers. Linear congruential generators (particularly multiplicative generators), shift registers, generalized feedback shift register generators, lagged-Fibonacci, inversive congruential generators, and nonlinear congruential generators are described in detail by Gentle (1998); see also Lewis and Orav (1989) and L'Ecuyer (1990). The *period of a generator* is the length of the sequence of numbers that it produces before it starts to repeat itself. Research is partly driven by the need in parallel processing for generators with longer periods than many currently in use.

A number of very fast general methods for generating discrete random variables have been developed. Such methods are distribution nonspecific, in the sense that they require tables relating to the actual values of the probability mass function, for example, a table of the cdfs, rather than knowledge of the structural properties of the distribution. They can, in principle, be applied to any univariate discrete distribution and are usually the method of choice when large numbers of rv's are required from a particular distribution with constant parameters. The following methods are particularly suitable for discrete distributions:

(i) The *inversion method* can be used for the generation of both continuous and discrete distributions. A uniform [0, 1] variate is generated and is transformed

into a variate from the *target distribution* (the distribution of interest) by the use of a monotone transformation of the uniform cdf to the target cdf; this procedure is called *inversion* of the (target) cdf.

(ii) The *table look-up method* is an adaptation of the inversion procedure that is particularly suitable for a discrete distribution and is very widely used. A set-up routine is required in which the cumulative probabilities for the target distribution are calculated correctly and are stored in computer memory. The cdf for a discrete distribution is of course a step function, with step jumps occurring at successive variate values and step heights equal to successive probabilities. A uniform [0, 1] variate is generated, and the appropriate step height interval within which it lies is sought using a search procedure. The variate value corresponding to this step jump is then "returned" (i.e., made available) as a variate from the target distribution. The use of one of the many sophisticated search procedures that are now available can make this a very fast method.

(iii) Walker's (1974, 1977) *alias method* is based on the following theorem:

> Every discrete distribution with probabilities $p_0, p_1, \cdots, p_{K-1}$ can be expressed as an equiprobable mixture of K two-point distributions.

First the probabilities for the target distribution must be calculated. Next a set-up procedure for constructing the K equiprobable mixtures is required; the information concerning these mixtures can be put into two arrays of size K by an ingenious method described in the books referenced below. One uniform variate on [0, 1] is then used to choose a component in the equiprobable mixture, while a second uniform variate on [0, 1] decides which of the two points for that component should be returned as a target variable. Once the set-up algorithm has been implemented (this may take a nontrivial amount of computer time), the generation of large numbers of variates from the target distribution is very rapid. For implementation details for these methods see, for example, Chen and Asau (1974) (for the indexed table look-up method) and Kronmal and Peterson (1979) (for the alias method).

(iv) If the order in which the variates are generated is immaterial, then C. D. Kemp and A. W. Kemp's (1987) *frequency table method* provides an even faster approach in the fixed-parameter situation. The method generates a sample of values in the form of a frequency table and is useful, for example, for studying the properties of estimators; the method does not attempt to provide a sequence of uncorrelated variate values.

Distributionally nonspecific methods are not, however, suitable when the parameters of a distribution change from call to call to the computer generator. Consequently many different *distribution-specific generation methods* have been devised. Some of these are mentioned in the appropriate chapter later in this book. Special attention has been received by the binomial and Poisson distributions because of their central role in discrete distribution theory. It should be emphasized that the practical implementation of computer generation algorithms

is not straightforward; the use of thoroughly tested standard packages (e.g., NAG Libraries [Numerical Algorithms Group], IMSL Libraries [Visual Numerics]) is recommended.

Useful references concerning the computer generation of rv's are the books by Morgan (1984), Ripley (1987), Bratley, Fox, and Schrage (1987), and Dagpunar (1988); Boswell, Gore, Patil, and Taillie (1993) provided a helpful general survey article. Devroye (1986) gives an encyclopedic coverage of the mathematical methodology of nonuniform random-variate generation. Gentle (1998, 2002) covers the recent literature very thoroughly.

1.2.18 Computer Software

The advent of cheap and powerful computing facilities has transformed the statistical analysis of data. The computer languages Algol, APL, and Pascal have largely given way to Fortran 90, C, C++, and S-Plus and the open source variant R.

Many more computer packages are now available. The most flexible of these allow interaction between the user and the package and the incorporation of modules for specific tasks. It is not possible in a limited space to make recommendations. We list some of the major packages with their websites so that the reader can obtain further information.

GenStat	VSN International	www.vsn-intl.com
GLIM	NAG	www.nag.co.uk
IMSL	Visual Numerics	www.vni.com
Minitab	Minitab	www.minitab.com
MLwiN	Centre for Multilevel Modelling	multilevel.ioe.ac.uk
R	R Foundation	www.r-project.org
SAS	SAS Institute	www.sas.com
S-PLUS	Insightful	www.insightful.com
SPSS	SPSS	www.spss.com
Stata	Stata	www.stata.com
StatXact and LogXact	Cytel Software	www.cytel.com