1

Basic Ideas of Mathematical Statistics

Elementary statistical computations have been carried out for thousands of years. For example, the arithmetic mean from a number of measures or observation data has been known for a very long time.

First descriptive statistics arose starting with the collection of data, for example, at the national census or in registers of medical cards, and followed by compression of these data in the form of statistics or graphical representations (figures). Mathematical statistics developed on the fundament of probability theory from the end of 19th century on. At the beginning of the 20th century, Karl Pearson and Sir Ronald Aylmer Fisher were notable pioneers of this new discipline. Fisher's book (1925) was a milestone providing experimenters such basic concepts as his well-known maximum likelihood method and analysis of variance as well as notions of sufficiency and efficiency. An important information measure is still called the Fisher information (see Section 1.4).

Concerning historical development we do not want to go into detail. We refer interested readers to Stigler (1986, 1990). Instead we will describe the actual state of the theory. Nevertheless many stimuli come from real applications. Hence, from time to time we will include real examples.

Although the probability calculus is the fundament of mathematical statistics, many practical problems containing statements about random variables cannot be solved with this calculus alone. For example, we often look for statements about parameters of distribution functions although we do not partly or completely know these functions. Mathematical statistics is considered in many introductory textbooks as the theory of analysing experiments or samples; that is, it is assumed that a random sample (corresponding to Section 1.1) is given. Often it is not considered how to get such a random sample in an optimal way. This is treated later in design of experiments. But in concrete applications, the experiment first has to be planned, and after the experiment is finished, the analysis has to be carried out. But in theory it is appropriate to determine firstly the optimal evaluation, for example, the smallest sample size for a variance optimal estimator. Hence we proceed in such a way and start with the optimal evaluation, and after

Mathematical Statistics, First Edition. Dieter Rasch and Dieter Schott.

this we work out the design problems. An exception is made for sequential methods where planning and evaluation are realised together.

Mathematical statistics involves mainly the theory of point estimation, statistical selection theory, the theory of hypothesis testing and the theory of confidence estimation. In these areas theorems are proved, showing which procedures are the best ones under special assumptions.

We wish to make clear that the treatment of mathematical statistics on the one hand and its application to concrete data material on the other hand are totally different concepts. Although the same terms often occur, they need not be confused. Strictly speaking, the notions of the empirical sphere (hence of the real world) are related to corresponding models in theory.

If assumptions for deriving best methods are not fulfilled in practical applications, the question arises how good these best methods still are. Such questions are answered by a part of empirical statistics – by simulations. We often find that the assumption of a normal distribution occurring in many theorems is far from being a good model for many data in applications. In the last years simulation developed into its own branch in mathematics. This shows a series of international workshops on simulation. The first to sixth workshops took place in St. Petersburg (Russia) in 1994, 1996, 1998, 2001, 2005 and 2009. The seventh international workshop on simulation took place in Rimini (Italy) in 2013 and the eighth one in Vienna (Austria) in 2015.

Because the strength of assumptions has consequences mainly in hypothesis testing and confidence estimation, we discuss such problems first in Chapter 3, where we introduce the concept of robustness against the strength of assumptions.

1.1 Statistical Population and Samples

1.1.1 Concrete Samples and Statistical Populations

In the empirical sciences, one character or several characters simultaneously (character vector) are observed in certain objects (or individuals) of a population. The main task is to conclude from the sample of observed values to the whole set of character values of all objects of this population. The problem is that there are objective or economical points of view that do not admit the complete survey of all character values in the population. We give some examples:

- The costs to register all character values were out of all proportion to the value of the statement (for instance, measuring the height of all people worldwide older than 18 years).
- The registration of character values results in destruction of the objects (destructive materials testing such as resistance to tearing of ropes or stockings).

- The set of objects is of hypothetic nature, for example, because they partly do not exist at the moment of investigation (as all products of a machine).

We can neglect the few practical cases where all objects of a population can be observed and no more extensive population is demanded, because for them mathematical statistics is not needed. Therefore we assume that a certain part (subset) is chosen from the population to observe a character (or character vector) from which we want to draw conclusions to the whole population. We call such a part a (concrete) sample (of the objects). The set of character values measured for these objects is said to be a (concrete) sample of the character values. Each object of the population is to possess such a character value (independent of whether we register the value or not). The set of character values of all objects in the population is called the corresponding statistical population.

A concrete population as well as the (sought-after/relevant) character and therefore also the corresponding statistical population need to be determined uniquely. Populations have to be circumscribed in the first line in relation to space and time. In principle it must be clear for an arbitrary real object whether it belongs to the population or not. In the following we consider some examples:

Original population		Statistical population	
A	Heifer of a certain breed in a certain region in a certain year	A_1	Yearly yield of milk of these heifer
		A_2	Body mass of these heifer after 180 days
		A_3	Back height of these heifer
B	Inhabitants of a town at a certain day	B_1	Blood pressure of these inhabitants at 6.00 o'clock
		B_2	Age of these inhabitants

It is clear that applying conclusions from a sample to the whole population can be wrong. For example, if the children of a day nursery are chosen from the population B in the table above, then possibly the blood pressure B_1 but without doubt the age B_2 are not applicable to B. Generally we speak of characters, but if they can have a certain influence to the experimental results, they are also called factors. The (mostly only a few) character values are said to be factor levels, and the combinations of factor levels of several factors factor level combinations.

The sample should be representative with respect to all factors that can influence the character of a statistical population. That means the composition of the population should be mirrored in the sample of objects. But that is impossible for small samples and many factor level combinations. For example, there are already about 200 factor level combinations in population B concerning the factors age and sex, which cannot be representatively found in a sample of 100

inhabitants. Therefore we recommend avoiding the notion of 'representative sample' because it cannot be defined in a correct way.

Samples should not be assessed according to the elements included but according to the way these elements have been selected. This way of selecting a sample is called sampling procedure. It can be applied either to the objects as statistical units or to the population of character values (e.g. in a databank). In the latter case the sample of character values arises immediately. In the first case the character must be first registered at the selected objects. Both procedures (but not necessarily the created samples) are equivalent if the character value is registered for each registered object. This is assumed in this chapter. It is not the case in so-called censored samples where the character values could not be registered in all units of the experiment. For example, if the determination of lifespans of objects (as electronic components) is finished at a certain time, measured values of objects with longer lifespans (as time of determination) are missing.

In the following we do not differ between samples of objects and samples of character values; the definitions hold for both.

Definition 1.1 A sampling procedure is a rule of selecting a proper subset, named sample, from a well-defined finite basic set of objects (population, universe). It is said to be at random if each element of the basic set has the same probability p to come into the sample. A (concrete) sample is the result of a sampling procedure. Samples resulting from a random sampling procedure are said to be (concrete) random samples.

There are a lot of random sampling procedures in the theory of samples (see, e.g. Cochran and Boing, 1972; Kauermann and Küchenhoff, 2011; Quatember, 2014) that can be used in practice. Basic sets of objects are mostly called (statistical) populations or synonymously sometimes (statistical) universes in the following.

1.1.2 Sampling Procedures

Concerning random sampling procedures, we distinguish (among other cases)

- The simple sampling, where each element of the population has the same probability to come into the sample.
- The stratified sampling, where a random sampling is done within the before-defined (disjoint) subclasses (strata) of the population. This kind of sampling is only at random as a whole if the sampling probabilities within the classes are chosen proportionally to the cardinalities of the classes.
- The cluster sampling, where the population is divided again into disjoint subclasses (clusters), but the sampling of objects is done not among the objects of the population itself but among the clusters. In the selected clusters all objects are registered. This kind of selection is often used as area samples. It is only at

random corresponding to Definition 1.1 if the clusters contain the same number of objects.

- The multistage sampling, where at least two stages of sampling are taken. In the latter case the population is firstly decomposed into disjoint subsets (primary units). Then a random sampling is done in all primary units to get secondary units. A multistage sampling is favourable if the population has a hierarchical structure (e.g. country, province, towns in the province). It is at random corresponding to Definition 1.1 if the primary units contain the same number of secondary units.
- The (constantly) sequential sampling, where the sample size is not fixed at the beginning of the sampling procedure. At first a small sample is taken and analysed. Then it is decided whether the obtained information is sufficient, for example, to reject or to accept a given hypothesis (see Chapter 3), or if more information is needed by selecting a further unit.

Both a random sampling (procedure) and an arbitrary sampling (procedure) can result in the same concrete sample. Hence we cannot prove by inspecting the sample itself whether the sample is randomly chosen or not. We have to check the sampling procedure used.

For the pure random sampling, Definition 1.1 is applied directly: each object in the population of size N is drawn with the same probability p. The number of objects in a sample is called sample size, mostly denoted by n.

The most important case of a pure random sampling occurs if the objects drawn from a population are not put back. An example is a lottery, where n numbers are drawn from N given numbers (in Germany the well-known Lotto uses $N = 49$ and $n = 6$). Using an unconditioned sampling of size N, the number $M = \binom{N}{n}$ of all possible subsets have the same probability $p = \frac{1}{M}$ to come into the sample.

As mentioned before, the sample itself is only at random if a random sampling method was used. But persons become at once suspicious if the sample is extreme. If somebody gets the top prize buying one lot of 10.000 possible lots, then this case is possible although rather unlikely. It can happen at random, or in other words, it can be the result of (a correct) random sampling. But if this person gets the top prize buying one lot at three consecutive lotteries of the mentioned kind, and if it turns out additionally that the person is the brother of the lot seller, then doubts are justified that there was something fishy going on. We would refuse to accept such unlikely events and would suppose that something is wrong. In our lottery case, we would assume that the selection was not at random and that cheats were at work. Nevertheless, there is an extremely small possibility that this event is at random, namely, $p = 1/1.000.000.000.000$.

Incidentally, the strategy of statistical tests in Chapter 3 is to refuse models (facts) under which observed events possess a very small probability and instead to accept models where these events have a larger probability.

A pure sampling also occurs if the random sample is obtained by replacing the objects immediately after drawing and observing that each object has the same probability to come into the sample using this procedure. Hence, the population always has the same number of objects before a new object is taken. That is only possible if the observation of objects works without destroying or changing them (examples where that is impossible are tensile breaking tests, medical examinations of killed animals, felling of trees and harvesting of food). The discussed method is called simple random sampling with replacement.

If a population of N objects is given and n objects are selected, then it is $n < N$ for sampling without replacement, while objects that can multiply occur in the sample and $n > N$ is possible for sampling with replacement.

A method that can sometimes be realised more easily is the systematic sampling with random start. It is applicable if the objects of the finite sampling set are numbered from 1 to N and the sequence is not related to the character considered. If the quotient $m = N/n$ is a natural number, a natural number i between 1 and m is chosen at random, and the sample is collected from objects with numbers $i, m + i, 2m + i, \ldots, (n - 1)m + i$. Detailed information about this case and the case where the quotient m is not natural can be found in Rasch et al. (2008) in Method (1/31/1210).

The stratified sampling already mentioned is advantageous if the population of size N is decomposed in a content-relevant manner into s disjoint subpopulations of sizes $N_1, N_2, \ldots, N_s$. Of course, the population can sometimes be divided into such subpopulations following the levels of a supposed interfering factor. The subpopulations are denoted as strata. Drawing a sample of size n is to realise in such a population an unrestricted sampling procedure holds the danger that not all strata are considered in general or at least not in appropriate way. Therefore in this case a stratified random sampling procedure is favourable. Then partial samples of size n_i are collected from the ith stratum ($i = 1, 2, \ldots, s$) where pure random sampling procedures are used in each stratum. This leads to a random sampling procedure for the whole population if the numbers n_i/n are chosen proportional to the numbers N_i/N.

While for the stratified random sampling objects are selected from each subset, for the multistage sampling, subsets or objects are selected at random at each stage as described below. Let the population consist of k disjoint subsets of size N_0, the primary units, in the two-stage case. Further, it is supposed that the character values in the single primary units differ only at random, so that objects need not to be selected from all primary units. If the wished sample size is $n = r\, n_0$ with $r < k$, then, in the first step, r of the k given primary units are selected using a pure random sampling procedure. In the second step n_0 objects (secondary units) are chosen from each primary unit again applying a pure random sampling. The number of possible samples is $\binom{k}{r} \cdot \binom{N_0}{n_0}$, and each

Table 1.1 Possible samples using different sampling procedures.

Sampling	Number K of possible samples
Simple random sampling	$K = \binom{1000}{100} > 10^{140}$
Systematic sampling with random start $k = 10$	$K = 10$
Stratified random sampling $k = 20$, $N_i = 50$, $i = 1, \ldots, 20$	$K = \left[\binom{50}{5}\right]^{20} = 2.11876 \cdot 10^{120}$
Stratified random sampling $k = 10$, $N_i = 100$, $i = 1, \ldots, 10$	$K = \left[\binom{100}{10}\right]^{10} = 1.7310309 \cdot 10^{130}$
Stratified random sampling $k = 5$, $N_i = 200$, $i = 1, \ldots, 5$	$K = \left[\binom{200}{20}\right]^{5} = 1.6135878 \cdot 10^{135}$
Stratified random sampling $k = 2$, $N_1 = 400$, $N_2 = 600$	$K = \binom{400}{40} \cdot \binom{600}{60} = 5.4662414 \cdot 10^{138}$
Two-stage sampling $k = 20$, $N_0 = 50$, $r = 4$	$K = \binom{20}{4} \cdot \binom{50}{25} = 6.1245939 \cdot 10^{17}$
Two-stage sampling $k = 20$, $N_0 = 50$, $r = 5$	$K = \binom{20}{5} \cdot \binom{50}{20} = 7.3069131 \cdot 10^{17}$
Two-stage sampling $k = 10$, $N_0 = 100$, $r = 2$	$K = \binom{10}{2} \cdot \binom{100}{50} = 4.5401105 \cdot 10^{30}$
Two-stage sampling $k = 10$, $N_0 = 100$, $r = 4$	$K = \binom{10}{4} \cdot \binom{100}{25} = 5.0929047 \cdot 10^{25}$
Two-stage sampling $k = 5$, $N_0 = 200$, $r = 2$	$K = \binom{5}{2} \cdot \binom{200}{50} = 4.5385838 \cdot 10^{48}$
Two-stage sampling $k = 2$, $N_0 = 500$, $r = 1$	$K = \binom{2}{1} \cdot \binom{500}{100} > 10^{100}$

object of the population has the same probability $p = \frac{r}{k} \cdot \frac{n_0}{N_0}$ to reach the sample corresponding to Definition 1.1.

Example 1.1 A population has $N = 1000$ objects. A sample of size $n = 100$ should be drawn without replacement of objects. Table 1.1 lists the number

of possible samples using the discussed sampling methods. The probability of selection for each object is $p = 0.1$.

1.2 Mathematical Models for Population and Sample

In mathematical statistics notions are defined that are used as models (generalisations) for the corresponding empirical notions. The population, which corresponds to a frequency distribution of the character values, is related to the model of probability distribution. The concrete sample selected by a random procedure is modelled by the realised (theoretical) random sampling. These model concepts are adequate, if the size N of the populations is very large compared with the size n of the sample.

Definition 1.2 An n-dimensional random variable

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T, n \geq 1$$

with components $\boldsymbol{y}_i$ is said to be a random sample, if

- All $\boldsymbol{y}_i$ have the same distribution characterised by the distribution function $F(y_i, \theta) = F(y, \theta)$ with the parameter (vector) $\theta \in \Omega \subseteq R^p$ and
- All $\boldsymbol{y}_i$ are stochastically independent from each other, that is, it holds for the distribution function $F(Y,\theta)$ of $\boldsymbol{Y}$ the factorisation

$$F(Y, \theta) = \prod_{i=1}^{n} F(y_i, \theta), \theta \in \Omega \subseteq R^p.$$

The values $Y = (y_1, y_2, \ldots, y_n)^T$ of a random sample $\boldsymbol{Y}$ are called realisations. The set $\{Y\}$ of all possible realisations of $\boldsymbol{Y}$ is called sample space.

In this book the random variables are printed with bold characters, and the sample space $\{Y\}$ belongs always to an n-dimensional Euclidian space, that is, $\{Y\} \subset R^n$.

The function

$$L(Y,\theta) = \begin{cases} f(Y, \theta) = \dfrac{\partial F(Y, \theta)}{\partial Y}, & \text{for continuous } y \\ p(Y, \theta), & \text{for discrete } y \end{cases}$$

with the probability function $p(Y, \theta)$ and the density function $f(Y, \theta)$ correspondingly is said to be for given Y as function of θ the likelihood function (of the distribution).

Random sample can have two different meanings, namely:

- Random sample as random variable $\boldsymbol{Y}$ corresponding to Definition 1.2
- (Concrete) random sample as subclass of a population, which was selected by a random sample procedure.

The realizations Y of a random sample $\boldsymbol{Y}$ we call a realized random sample.

The random sample $\boldsymbol{Y}$ is the mathematical model of the simple random sample procedure, where concrete random sample and realised random sample correspond to each other also in the symbolism.

We describe in this book the 'classical' philosophy, where $\boldsymbol{Y}$ is distributed by the distribution function $F(Y, \theta)$ with the fixed (not random) parameter $\theta \in \Omega \subseteq R^p$. Besides there is the philosophy of Bayes where a random $\boldsymbol{\theta}$ is supposed, which is distributed a priori with a parameter φ assumed to be known. In the empirical Bayesian method, the a priori distribution is estimated from the data collected.

1.3 Sufficiency and Completeness

A random variable involves certain information about the distribution and their parameters. Mainly for large n (say, $n > 100$), it is useful to condense the objects of a random sample in such a way that fewest possible new random variables contain as much as possible of this information. This vaguely formulated concept is to state more precisely stepwise up to the notion of minimal sufficient statistic. First, we repeat the definition of an exponential family.

The distribution of a random variable $\boldsymbol{y}$ with parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ belongs to a k-parametric exponential family if its likelihood function can be written as

$$f(y, \theta) = h(y)\mathrm{e}^{\sum_{i=1}^{k} \eta_i(\theta) \cdot T_i(y) - B(\theta)},$$

where the following conditions hold:

- η_i and B are real functions of θ and B does not depend on y.
- The function $h(y)$ is non-negative and does not depend on θ.

The exponential family is in canonical form with the so-called natural parameters η_i, if their elements can be written as

$$f(y, \eta) = h(y)\mathrm{e}^{\sum_{i=1}^{k} \eta_i \cdot T_i(y) - A(\eta)} \text{ with } \eta = (\eta_1, \ldots, \eta_k)^T.$$

Let $(P_\theta, \theta \epsilon \Omega)$ be a family of distributions of random variables $\boldsymbol{y}$ with the distribution function $F(y, \theta)$, $\theta \in \Omega$. The realisations $Y = (y_1, \ldots, y_n)^T$ of the random sample

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T,$$

where the components $\boldsymbol{y_i}$ are distributed as $\boldsymbol{y}$ itself lie in the sample space $\{Y\}$. According to Definition 1.2 the distribution function $F(Y, \theta)$ of a random sample $\boldsymbol{Y}$ is just as $F(y, \theta)$ uniquely determined.

Definition 1.3 A measurable mapping $\boldsymbol{M} = M(\boldsymbol{Y}) = [M_1(\boldsymbol{Y}), \ldots, M_r(\boldsymbol{Y})]^T, r \le n$ of $\{Y\}$ on a space $\{M\}$, which does not depend on $\theta \in \boldsymbol{\Omega}$, is called a statistic.

Definition 1.4 A statistic $\boldsymbol{M}$ is said to be sufficient relative to a distribution family $(P_\theta, \theta \epsilon \boldsymbol{\Omega})$ or relative to $\theta \in \boldsymbol{\Omega}$, respectively, if the conditional distribution of a random sample $\boldsymbol{Y}$ is independent of θ for given $\boldsymbol{M} = M(\boldsymbol{Y}) = M(Y)$.

Example 1.2 Let the components of a random sample $\boldsymbol{Y}$ satisfy a two-point distribution with the values 1 and 0. Further, let be $P(\boldsymbol{y}_i = 1) = p$ and $P(\boldsymbol{y}_i = 0) = 1 - p, 0 < p < 1$. Then $\boldsymbol{M} = M(\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{y_i}$ is sufficient relative (corresponding) to $\theta \in (0, 1) = \boldsymbol{\Omega}$. To show this, we have to prove that $P\left(\boldsymbol{Y} = Y \mid \sum_{i=1}^{n} \boldsymbol{y}_i = M\right)$ is independent of p. Now it is

$$P(\boldsymbol{Y} = Y \mid M) = \frac{P(\boldsymbol{Y} = Y, \boldsymbol{M} = M)}{P(\boldsymbol{M} = M)}, M = 0, 1, \ldots, n.$$

We know from probability theory that $\boldsymbol{M} = M(\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{y_i}$ is binomially distributed with the parameters n and p. Hence it follows that

$$P(\boldsymbol{M} = M) = \binom{n}{M} p^M (1-p)^{n-M}, M = 0, 1, \ldots, n.$$

Further we get with $y_i = 0$ or $y_i = 1$ and $A(M) = \{Y \mid M(Y) = M\}$ the result

$$\begin{aligned} P[\boldsymbol{Y} = Y, M(\boldsymbol{Y}) = M] &= P(\boldsymbol{y}_1 = y_1, \ldots, \boldsymbol{y}_n = y_n) I_{A(M)}(Y) \\ &= \prod_{i=1}^{n} \left(\binom{1}{y_i} p^{y_i} (1-p)^{1-y_i} \right) I_{A(M)}(Y) \\ &= p^{\sum_{i=1}^{n} y_i} (1-p)^{n - \sum_{i=1}^{n} y_i} I_{A(M)}(Y) \\ &= p^M (1-p)^{(n-M)}. \end{aligned}$$

Consequently it is $P(\boldsymbol{Y} = Y \mid M) = \dfrac{1}{\binom{n}{M}}$, and this is independent of p.

This way of proving sufficiency is rather laborious, but we can also apply it for continuous distributions as the next example will show.

Example 1.3 Let the components $\boldsymbol{y}_i$ of a random sample $\boldsymbol{Y}$ of size n be distributed as $N(\mu,1)$ with expected value μ and variance $\sigma^2 = 1$. Then $\boldsymbol{M} = \sum \boldsymbol{y}_i$ is sufficient relative to $\mu \in R^1 = \Omega$. To show this we first remark that $\boldsymbol{Y}$ is distributed as $N(\mu e_n, I_n)$. Then we apply the one-to-one transformation

$$\boldsymbol{Z} = A\boldsymbol{Y} = \left(\boldsymbol{z}_1 = \sum \boldsymbol{y}_i, \boldsymbol{y}_2 - \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n - \boldsymbol{y}_1\right) \text{ with } A = \begin{pmatrix} 1 & e_{n-1}^T \\ -e_{n-1} & I_{n-1} \end{pmatrix}$$

where $|A| = n$. We write $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{Z}_2) = (\sum \boldsymbol{y}_i, \boldsymbol{y}_2 - \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n - \boldsymbol{y}_1)$ and recognise that

$$\operatorname{cov}(\boldsymbol{Z}_2, \boldsymbol{z}_1) = \operatorname{cov}(\boldsymbol{Z}_2, \boldsymbol{M}) = \operatorname{cov}\left((-e_{n-1}, I_{n-1})\boldsymbol{Y}, e_n^T \boldsymbol{Y}\right) = 0_{n-1}.$$

Considering the assumption of normal distribution, the variables $\boldsymbol{M}$ and $\boldsymbol{Z}_2$ are stochastically independent. Consequently $\boldsymbol{Z}_2$ but also $\boldsymbol{Z}_2 \,|\boldsymbol{M}$ and $\boldsymbol{Z} \,|\boldsymbol{M}$ are independent of μ. Taking into account that the mapping $\boldsymbol{Z} = A\boldsymbol{Y}$ is biunique, also $\boldsymbol{Y}|\boldsymbol{M}$ is independent of μ. Hence $\boldsymbol{M} = \sum \boldsymbol{y}_i$ is sufficient relative to $\mu \in R^1$. With a sufficient $\boldsymbol{M} = \sum \boldsymbol{y}_i$ and a real number $c \neq 0$, then $c\,\boldsymbol{M}$ is also sufficient, that is, $\frac{1}{n}\sum \boldsymbol{y}_i = \bar{\boldsymbol{y}}$ is sufficient.

But sufficiency plays such a crucial part in mathematical statistics, and we need simpler methods for proving sufficiency and mainly for finding sufficient statistics. The following theorem is useful in this direction.

Theorem 1.1 (Decomposition Theorem)
Let a distribution family $(P_\theta, \theta \epsilon \Omega)$ of a random sample $\boldsymbol{Y}$ be given that is dominated by a finite measure ν. The statistic $M(\boldsymbol{Y})$ is sufficient relative to θ, if the Radon–Nikodym density f_θ of P_θ can be written corresponding to ν as

$$f_\theta(Y) = g_\theta[M(Y)]h(Y) \tag{1.1}$$

ν – almost everywhere. Then the following holds:

- The ν – integrable function g_θ is non-negative and measurable.
- h is non-negative and $h(Y) = 0$ is fulfilled only for a set of P_θ – measure 0.

The general proof came from Halmos and Savage (1949); it can be also found, for example, in Bahadur (1955) or Lehmann and Romano (2008).

In the present book we work only with discrete and continuous probability distributions satisfying the assumptions of Theorem 1.1. A proof of the theorem for such distributions is given in Rasch (1995). We do not want to repeat it here.

For discrete distributions Theorem 1.1 means that the probability function is of the form

$$p(Y, \theta) = g[M(Y), \theta]h(Y). \tag{1.2}$$

For continuous distributions the density function has the form

$$f(Y,\theta) = g[M(Y),\theta]h(Y). \tag{1.3}$$

Corollary 1.1 If the distribution family $(P^*(\theta), \theta \epsilon \Omega)$ of the random variable $\boldsymbol{y}$ is a k-parametric exponential family with the natural parameter η and the likelihood function

$$L^*(y,\eta) = h^*(y)e^{\sum_{i=1}^{k} M_j^*(y) - A(\eta)}, \tag{1.4}$$

then, denoting the random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$,

$$M(\boldsymbol{Y}) = \left(\sum_{i=1}^{n} \boldsymbol{M}_1^*(\boldsymbol{y_i}), \ldots, \sum_{i=1}^{n} \boldsymbol{M}_k^*(\boldsymbol{y_i}) \right)^T \tag{1.5}$$

is sufficient relative to θ.

Proof: It is

$$L(y,\eta) = \prod_{i=1}^{n} h^*(y_i) \mathrm{e}^{\sum_{j=1}^{k} \eta_j \sum_{i=1}^{n} M_j^*(y_i) - nA(\eta)}, \tag{1.6}$$

which is of the form (1.2) and (1.3), respectively, where $h(Y) = \prod_{i=1}^{n} h^*(y_i)$ and $\theta = \eta$.

Definition 1.5 Two likelihood functions, $L_1(Y_1, \theta)$ and $L_2(Y_2, \theta)$, are said to be equivalent, denoted by $L_1 \sim L_2$, if

$$L_1(Y_1,\theta) = a(Y_1,Y_2)\, L_2(Y_2,\theta) \tag{1.7}$$

with a function $a(Y_1, Y_2)$ that is independent of θ.

Then it follows from Theorem 1.1

Corollary 1.2 $M(\boldsymbol{Y})$ is sufficient relative to θ if and only if (*iff*) the likelihood function $L_M(M, \theta)$ of $\boldsymbol{M} = M(\boldsymbol{Y})$ is equivalent to the likelihood function of a random sample $\boldsymbol{Y}$.

Proof: If $M(\boldsymbol{Y})$ is sufficient relative to θ, then because of

$$L_M(M,\theta) = a(Y)\, L(Y,\theta), a(Y) > 0, \tag{1.8}$$

$L_M(M, \theta)$ has together with $L(Y, \eta)$ the form (1.1). Reversely, (1.8) implies that the conditional distribution of a random sample $\boldsymbol{Y}$ is for given $M(\boldsymbol{Y}) = M$ independent of θ.

Example 1.4 Let the components y_i of a random sample $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$ be distributed as $N(\mu, 1)$. Then it is

$$L(Y,\mu) = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2}(Y-\mu e_n)^T (Y-\mu e_N)} = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2}\sum_{i=1}^n (y_i-\bar{y})^2} e^{-\frac{n}{2}(\bar{y}-\mu)^2}. \tag{1.9}$$

Since $M(\boldsymbol{Y}) = \bar{\boldsymbol{y}}$ is distributed as $N\left(\mu, \frac{1}{n}\right)$, we get

$$L_M(\bar{y},\mu) = \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{n}{2}(\bar{y}-\mu)^2}. \tag{1.10}$$

Hence $L_M(\bar{y},\mu) \sim L(Y,\mu)$ holds, and $\bar{\boldsymbol{y}}$ is sufficient relative to μ.

Generally we immediately obtain from Definition 1.4 the

Corollary 1.3 If $c > 0$ is a real number chosen independently of θ and $M(\boldsymbol{Y})$ is sufficient relative to θ, then $c\,M(\boldsymbol{Y})$ is also sufficient relative to θ.

Hence, for example, by putting $\boldsymbol{M} = \sum \boldsymbol{y}_i$ and $c = \frac{1}{n}$, also $\frac{1}{n}\sum \boldsymbol{y}_i = \bar{\boldsymbol{y}}$ is sufficient.

The problem arises whether there exist under the statistics sufficient relative to the distribution family $(P^*(\theta), \theta \in \Omega)$ such, which are minimal in a certain sense, containing as few as possible components. The following example shows that this problem is no pure invention.

Example 1.5 Let $(P^*(\theta), \theta \in \Omega)$ be the family of $N(\mu,\sigma^2)$-normal distributions $(\sigma > 0)$. We consider the statistics

$$\boldsymbol{M}_1(\boldsymbol{Y}) = \boldsymbol{Y}$$

$$\boldsymbol{M}_2(\boldsymbol{Y}) = \left(\boldsymbol{y}_1^2, \ldots, \boldsymbol{y}_n^2\right)^T$$

$$\boldsymbol{M}_3(\boldsymbol{Y}) = \left(\sum_{i=1}^r \boldsymbol{y}_i^2, \sum_{i=r+1}^n \boldsymbol{y}_i^2\right)^T, \quad r = 1, \ldots, n-1$$

$$\boldsymbol{M}_4(\boldsymbol{Y}) = \left(\sum_{i=1}^n \boldsymbol{y}_i^2\right)$$

of a random sample $\boldsymbol{Y}$ of size n, which are all sufficient relative to σ^2. This can easily be shown using Corollary 1.1 of Theorem 1.1 (decomposition theorem). The likelihood functions of $\boldsymbol{M}_1(\boldsymbol{Y})$ and $\boldsymbol{Y}$ are identical (and therefore equivalent). Since both the $\boldsymbol{y}_i$ and the $\boldsymbol{y}_i^2$ are independent and $\frac{\boldsymbol{y}_i^2}{\sigma^2} = \boldsymbol{\chi}_i^2$ are distributed as $CS(1)$ (χ^2-distribution with 1 degree of freedom; see Appendix A: Symbolism), it follows after the transformation $\boldsymbol{y}_i^2 = \sigma^2 \boldsymbol{\chi}_i^2$:

$$L_M\left(\boldsymbol{M}_2(\boldsymbol{Y}),\sigma^2\right) \sim L\left(Y,\sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2}. \tag{1.11}$$

Analogously we proceed with $\boldsymbol{M}_3(\boldsymbol{Y})$ and $\boldsymbol{M}_4(\boldsymbol{Y})$.

Obviously, $\boldsymbol{M}_4(\boldsymbol{Y})$ is the most extensive compression of the components of a random sample $\boldsymbol{Y}$, and therefore it is preferable compared with other statistics.

Definition 1.6 A statistic $M^*(\boldsymbol{Y})$ sufficient relative to θ is said to be minimal sufficient relative to θ if it can be represented as a function of each other sufficient statistic $M(\boldsymbol{Y})$.

If we consider Example 1.5, there is

$$\boldsymbol{M}_4(\boldsymbol{Y}) = \boldsymbol{M}_1^T(\boldsymbol{Y})\boldsymbol{M}_1(\boldsymbol{Y}) = e_n^T\boldsymbol{M}_2 = (1\ 1)\,\boldsymbol{M}_3, \quad r = 1,\ldots,n-1.$$

Hence, $\boldsymbol{M}_4(\boldsymbol{Y})$ can be written as a function of each sufficient statistic of this example. This is not true for $\boldsymbol{M}_1(\boldsymbol{Y})$, $\boldsymbol{M}_2(\boldsymbol{Y})$ and $\boldsymbol{M}_3(\boldsymbol{Y})$; they are not functions of $\boldsymbol{M}_4(\boldsymbol{Y})$. $\boldsymbol{M}_4(\boldsymbol{Y})$ is the only statistic of Example 1.5 that could be minimal sufficient relative to σ^2. We will see that it has indeed this property. But, how can we show minimal sufficiency? We recognise that the sample space can be decomposed with the help of the statistic $M(\boldsymbol{Y})$ in such a way into disjoint subsets that all Y for which $M(Y)$ supplies the same value M belong to the same subset. Vice versa, a given decomposition defines a statistic. Now we present a decomposition that is shown to generate a minimal sufficient statistic.

Definition 1.7 Let $Y_0 \in \{Y\}$ be a fixed point in the sample space (a certain value of a realised random sample), which contains the realisations of a random sample $\boldsymbol{Y}$ with components from a family $(P^*(\theta), \theta\epsilon\Omega)$ of probability distributions. The likelihood function $L(Y,\theta)$ generates by

$$M(Y_0) = \{Y : L(Y,\theta) \sim L(Y_0,\theta)\} \tag{1.12}$$

a subset in $\{Y\}$. If Y_0 runs through the whole sample space $\{Y\}$, then a certain decomposition is generated. This decomposition is called likelihood decomposition, and the corresponding statistic $\boldsymbol{M}_L(\boldsymbol{Y})$ satisfying $M_L(Y)$ = const. for all $Y \in M(Y_0)$ and each Y_0 is called likelihood statistic.

Before we construct minimal sufficient statistics for some examples by this method, we state

Theorem 1.2 The likelihood statistic $\boldsymbol{M}_L(\boldsymbol{Y})$ is minimal sufficient relative to θ.

Proof: Considering the likelihood statistic $\boldsymbol{M}_L(\boldsymbol{Y})$, it holds

$$\boldsymbol{M}_L(\boldsymbol{Y}_1) = \boldsymbol{M}_L(\boldsymbol{Y}_2)$$

for $Y_1, Y_2 \in \{Y\}$ iff $L(Y_1, \theta) \sim L(Y_2, \theta)$ is fulfilled. Hence, $L(Y, \theta)$ is a function of $M_L(\boldsymbol{Y})$ having the form

$$L(Y, \theta) = a(Y) g^* \left(M_L(Y), \theta \right). \tag{1.13}$$

Therefore $M_L(\boldsymbol{Y})$ is sufficient relative to θ taking Theorem 1.1 (decomposition theorem) into account. If $M(\boldsymbol{Y})$ is any other statistic that is sufficient relative to θ, if further for two points $Y_1, Y_2 \in \{Y\}$ the relation $M(\boldsymbol{Y}_1) = M(\boldsymbol{Y}_2)$ is satisfied, and finally, if $L(Y_i, \theta) > 0$ holds for $i = 1, 2$, then again Theorem 1.1 supplies

$$L(Y_1, \theta) = h(Y_1) g(M(Y_1), \theta) = h(Y_2) g(M(Y_2), \theta)$$

because of $M(\boldsymbol{Y}_1) = M(\boldsymbol{Y}_2)$ and

$$L(Y_2, \theta) = h(Y_2) g(M(Y_2), \theta) \text{ or equivalently } g(M(Y_2), \theta) = \frac{L(Y_2, \theta)}{h(Y_2)}.$$

Hence, we obtain

$$L(Y_1, \theta) = \frac{h(Y_1)}{h(Y_2)} L(Y_2, \theta), h(Y_2) > 0,$$

which means $L(Y_1, \theta) \sim L(Y_2, \theta)$. But this is just the condition for $M(\boldsymbol{Y}_1) = M(\boldsymbol{Y}_2)$. Consequently $M_L(\boldsymbol{Y})$ is a function of $M(\boldsymbol{Y})$, independent of how $M(\boldsymbol{Y})$ is chosen, that is, it is minimal sufficient.

We demonstrate the method giving two examples.

Example 1.6 Let the components $\boldsymbol{y}_i$ of a random sample $\boldsymbol{Y}$ fulfil a binomial distribution $B(N, p)$, N fixed and $0 < p < 1$. We look for a statistic that is minimal sufficient relative to p. The likelihood function is

$$L(Y, p) = \prod_{i=1}^{n} \binom{N}{y_i} p^{y_i} (1-p)^{N-y_i}, y_i = 0, 1, \ldots, N.$$

For all $Y_0 = (y_{01}, \ldots, y_{0N})^T \in \{Y\}$ with $L(Y_0, p) > 0$, it is

$$\frac{L(Y, p)}{L(Y_0, p)} = \frac{\prod_{i=1}^{n} \binom{N}{y_i}}{\prod_{i=1}^{n} \binom{N}{y_{0i}}} \left(\frac{p}{1-p} \right)^{\sum_{i=1}^{n} (y_i - y_{0i})}.$$

Therefore $M(Y_0)$ is also defined by $M(Y_0) = \left\{ Y : \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_{0i} \right\}$, since just there $L(Y, p) \sim L(Y_0, p)$ holds. Hence $M(\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{y}_i$ is a minimal sufficient statistic.

Example 1.7 Let the components $\boldsymbol{y}_i$ of a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ be gamma distributed. Then for $y_i > 0$ we get the likelihood function

$$L(Y, a, k) = \frac{a^{nk}}{[\Gamma(k)]^n} \mathrm{e}^{-a\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} y_i^{k-1}.$$

For all $Y_0 = (y_{01}, \ldots, y_{0N})^T \epsilon \{Y\}$ with $L(Y_0, a, k) > 0$, it is

$$\frac{L(Y, a, k)}{L(Y_0, a, k)} = \mathrm{e}^{-a\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_{0i}\right)} \frac{\prod_{i=1}^{n} y_i^{k-1}}{\prod_{i=1}^{n} y_{0i}^{k-1}}.$$

For given a the product $\prod_{i=1}^{n} \boldsymbol{y}_i$ is minimal sufficient relative to k. If k is known, then $\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{y}_i$ is minimal sufficient relative to a. If both a and k are unknown parameters, then $(\prod_{i=1}^{n} \boldsymbol{y}_i, \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{y}_i)$ is minimal sufficient relative to (a, k).

More generally the following statement holds:

Theorem 1.3 If $(P^*(\theta), \theta \epsilon \Omega)$ is a k-parametric exponential family with likelihood function in canonical form

$$L(y, \theta) = \mathrm{e}^{\sum_{i=0}^{k} \eta_i M_i(y) - A(\eta)} h(y),$$

where the dimension of the parameter space is k (i.e. the $\eta_1, \ldots, \eta_k$ are linearly independent), then

$$M(\boldsymbol{Y}) = \left(\sum_{i=1}^{n} M_1(\boldsymbol{y}_i), \ldots, \sum_{i=1}^{n} M_k(\boldsymbol{y}_i) \right)^T$$

is minimal sufficient relative to $(P^*(\theta), \theta \, \epsilon \, \Omega)$.

Proof: The sufficiency of $M(\boldsymbol{Y})$ follows from Corollary 1.1 of the decomposition theorem (Theorem 1.1), and the minimal sufficiency follows from the fact that $M(\boldsymbol{Y})$ is the likelihood statistic, because it is $L(Y, \theta) \sim L(Y_0, \theta)$ if

$$\sum_{j=1}^{k} \eta_j \sum_{i=1}^{n} \left[M_j(\boldsymbol{y}_i) - M_j(\boldsymbol{y}_{0i}) \right] = 0.$$

Regarding the linear independence of η_i, it is only the case if $M(\boldsymbol{Y}) = M(\boldsymbol{Y}_0)$ is fulfilled.

Example 1.8 Let $(P^*(\theta), \theta \, \epsilon \, \Omega)$ the family of a two-dimensional normal distributions with the random variable $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$, the expectation $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and the

covariance matrix $\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$. This is a four-parametric exponential family with the natural parameters

$$\eta_1 = \frac{\mu_x}{\sigma_x^2}, \eta_2 = \frac{\mu_y}{\sigma_y^2}, \eta_3 = -\frac{1}{2\,\sigma_x^2}, \eta_4 = -\frac{1}{2\,\sigma_y^2}$$

and the factors

$$M_1\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{x},\ M_2\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{y},$$

$$M_3\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{x}^2,\ M_4\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{y}^2,$$

$$A(\eta) = \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right).$$

If $\dim(\Omega) = 4$, then

$$\boldsymbol{M} = \left(\sum_{i=1}^{n} \boldsymbol{M}_{1i}, \sum_{i=1}^{n} \boldsymbol{M}_{2i}, \sum_{i=1}^{n} \boldsymbol{M}_{3i}, \sum_{i=1}^{n} \boldsymbol{M}_{4i}\right)^T$$

is minimal sufficient relative to $(P^*(\theta), \theta \epsilon \Omega)$. Assuming that $\left(\check{P}^{**}(\theta), \theta \epsilon \Omega\right) \subseteq (P^*(\theta), \theta \epsilon \Omega)$ is the subfamily of $(P^*(\theta), \theta \epsilon \Omega)$ with $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then $\dim(\Omega) = 3$ follows, and $\boldsymbol{M}$ is not minimal sufficient relative to $(\check{P}^{**}(\theta), \theta \epsilon \Omega)$.

The natural parameters of $(\check{P}^{**}(\theta), \theta \epsilon \Omega)$ are

$$\eta_1 = \frac{\mu_x}{\sigma^2}, \eta_2 = \frac{\mu_y}{\sigma^2}, \eta_3 = -\frac{1}{2\sigma^2}.$$

Further we have $A(\eta) = \frac{1}{2\sigma^2}\left(\mu_x^2 + \mu_y^2\right)$, and the factors of the η_i are

$$\check{M}_1\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{x}, \check{M}_2\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{y},\ \check{M}_3\left[\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right] = \boldsymbol{x}^2 + \boldsymbol{y}^2.$$

Relative to $\left(\check{P}^{**}(\theta), \theta \epsilon \Omega\right)$,

$$\check{\boldsymbol{M}} = \left(\sum_{i=1}^{n} \check{\boldsymbol{M}}_{1i}, \sum_{i=1}^{n} \check{\boldsymbol{M}}_{2i}, \sum_{i=1}^{n} \check{\boldsymbol{M}}_{3i}\right)^T$$

is minimal sufficient.

As it will be shown in Chapter 6 for model II of analysis of variance, the result of Theorem 1.3 is suitable also in more sophisticated models to find minimal sufficient statistics.

Completeness and bounded completeness are further important properties for the theory of estimation. We want to introduce both together by the following definition.

Definition 1.8 A distribution family $P = (P_\theta, \theta \epsilon \Omega)$ with distribution function $F(y, \theta)$, $\theta \epsilon \Omega$ is said to be complete, if for each P-integrable function $h(\boldsymbol{y})$ of the random variable $\boldsymbol{y}$ the condition

$$E[h(\boldsymbol{y})] = \int h(y)dF(y) = 0 \text{ for all } \theta \epsilon \Omega \tag{1.14}$$

implies the relation

$$P_\theta[h(\boldsymbol{y}) = 0] = 1 \text{ for all } \theta \epsilon \Omega. \tag{1.15}$$

If this is true only for bounded functions $h(y)$, then $P = (P_\theta, \theta \epsilon \Omega)$ is called bounded complete.

We want to consider an example for a complete distribution family.

Example 1.9 Let P be the family $\{P_p\}$, $p \in (0,1)$ of binomial distributions with the probability function

$$p(y,p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} \nu^y (1-p)^n, \quad 0 < p < 1,$$

$$y = 0, 1, \ldots, n, \ \nu = \frac{p}{1-p}.$$

Integrability of $h(y)$ means finiteness of $(1-p)^n \sum_{y=0}^{n} h(y) \binom{n}{y} \nu^y$, and (1.14) implies

$$\sum_{y=0}^{n} h(y) \binom{n}{y} \nu^y = 0 \text{ for all } p \in (0,1).$$

The left-hand side of the equation is a polynomial of nth degree in ν, which has at most n real zeros. To fulfil this equation for all $\nu \in R^+$, the factor $\binom{n}{y} h(y)$ must vanish for $y = 0, 1, \ldots, n$, and because of $\binom{n}{y} > 0$, it follows

$$P_\theta[h(\boldsymbol{y}) = 0)] = 1 \text{ for all } p \in (0,1).$$

Theorem 1.4 A k-parametric exponential family of the distribution of a sufficient statistic is complete under the assumptions of Theorem 1.3 ($\dim(\Omega) = k$).

The proof can be found in Lehmann and Romano (2008).

Definition 1.9 Let a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ be given whose components satisfy a distribution from the family

$$P^* = (P_\theta, \theta \in \Omega).$$

A statistic $M(\boldsymbol{Y})$, whose distribution is independent of θ, is called an ancillary statistic. If P is the family of distributions induced by the statistic $M(\boldsymbol{Y})$ in P^* and if P is complete and $M(\boldsymbol{Y})$ is sufficient relative to P^*, then $M(\boldsymbol{Y})$ is said to be complete sufficient.

Example 1.10 Let P^* be the family of normal distributions $N(\mu, 1)$ with expectation $\mu = \theta$ and variance 1, that is, it holds $\Omega = R^1$. This is a one-parametric exponential family with $\dim(\Omega) = 1$, which is complete by Theorem 1.4. If $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ is a random sample with components from P^*, then $M_1(\boldsymbol{Y}) = \bar{\boldsymbol{y}}$ is distributed as $N(\mu, \frac{1}{n})$. Consequently the family of distributions P^* is also complete. Because of Theorem 1.3, $\bar{\boldsymbol{y}}$ is minimal sufficient and therefore complete sufficient. The distribution family of $CS(n-1)$-distributions (χ^2-distributions with $n - 1$ degrees of freedom) induced by $(n-1)M_2(\boldsymbol{Y}) = \sum \boldsymbol{y}_i^2 - n\bar{\boldsymbol{y}}^2$ is independent of μ. Hence $\boldsymbol{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2$ is an ancillary statistic relative to $\mu = \theta$.

We close this section with the following statement:

Theorem 1.5 Let $\boldsymbol{Y}$ be a random sample with components from $P = (P_\theta, \theta \in \Omega)$ and let $M_1(\boldsymbol{Y})$ be bounded complete sufficient relative to P. Further, if $M_2(\boldsymbol{Y})$ is a statistic with a distribution independent of θ, then $M_1(\boldsymbol{Y})$ and $M_2(\boldsymbol{Y})$ are (stochastically) independent.

Proof: Let $\{Y_0\} \subset \{Y\}$ be a subset of the sample space $\{Y\}$. Then $M_2(\boldsymbol{Y})$ maps $\{Y\}$ onto $\{M\}$ and $\{Y_0\}$ onto $\{M_0\}$. Since the distribution of $M_2(\boldsymbol{Y})$ is independent of θ, $P[M_2(\boldsymbol{Y}) \in \{M_0\}]$ is independent of θ. Moreover, observing the sufficiency of $M_1(\boldsymbol{Y})$ relative to θ, also $P[M_2(\boldsymbol{Y}) \in \{M_0\} | M_1(\boldsymbol{Y})]$ is independent of θ. We consider the statistic

$$h(M_1(\boldsymbol{Y})) = P[M_2(\boldsymbol{Y}) \in \{M_0\} | M_1(\boldsymbol{Y})] - P[M_2(\boldsymbol{Y}) \in \{M_0\}]$$

depending on $M_1(\boldsymbol{Y})$, such that analogously to (1.14)

$$E_\theta[h(M_1(\boldsymbol{Y}))] = E_\theta[P[M_2(\boldsymbol{Y}) \epsilon \{M_0\}]M_1(\boldsymbol{Y}) - P[M_2(\boldsymbol{Y}) \epsilon \{M_0\}]] = 0$$

follows for all $\theta \epsilon \Omega$. Since $M_1(\boldsymbol{Y})$ is bounded complete,

$$P[M_2(\boldsymbol{Y})\epsilon\{M_0\}|M_1(\boldsymbol{Y})] - P[M_2(\boldsymbol{Y})\epsilon\{M_0\}] = 0$$

holds for all $\theta \epsilon \Omega$ with probability 1, analogously to (1.15). But this means that $M_1(\boldsymbol{Y})$ and $M_2(\boldsymbol{Y})$ are independent.

1.4 The Notion of Information in Statistics

Concerning the heuristic introduction of sufficient statistics in Section 1.2, we emphasised that a statistic should exhaust the information of a sample to a large extent. Now we turn to the question what the information of a sample really means. The notion of information was introduced by R. A. Fisher in the field of statistics, and his definition is still today of great importance. We speak of the Fisher information in this connection. A further notion of information originates from Kullback and Leibler (1951), but we do not present this definition here. We restrict ourselves in this section at first to distribution families

$$P = (P_\theta, \theta \epsilon \Omega), \Omega \subset R^1$$

with real parameters θ. We denote the likelihood function ($\boldsymbol{Y} = \boldsymbol{y}$) of P by $L(y, \theta)$.

Definition 1.10 Let $\boldsymbol{y}$ be distributed as

$$P = (P_\theta, \theta \epsilon \Omega), \Omega \subset R^1.$$

Further let the following assumption V1 be fulfilled:

1) Ω is an open interval.
2) For each $y \in \{Y\}$ and for each $\theta \epsilon \Omega$, the derivative $\frac{\partial}{\partial\theta}L(y,\theta)$ exists and is finite. The set of points satisfying $L(y, \theta) = 0$ does not depend on θ.
3) For each $\theta \epsilon \Omega$ there exist an $\varepsilon > 0$ and a positive P_θ-integrable function $k(y,\theta)$ such that for all θ_0 in an ε-neighbourhood of θ the inequality

$$\left|\frac{L(y,\theta) - L(y,\theta_0)}{\theta - \theta_0}\right| \leqslant k(y,\theta_0)$$

holds.
4) The derivative $\frac{\partial}{\partial\theta}L(y,\theta)$ is quadratic P_θ-integrable, and it holds for all $\theta \epsilon \Omega$

$$0 < E\left\{\left[\frac{\partial}{\partial\theta}\ln L(\boldsymbol{y},\theta)\right]^2\right\}.$$

Then the expectation

$$I(\theta) = E\left\{ \left[\frac{\partial}{\partial \theta} \ln L(\boldsymbol{y}, \theta) \right]^2 \right\} \tag{1.16}$$

is said to be the Fisher information of the distribution P_θ and of the variable $\boldsymbol{y}$, respectively.

It follows from the third condition of V1 that P_θ-integration and differentiation by θ can be exchanged for $L(\mathrm{y}, \theta)$, and because of

$$\frac{\partial}{\partial \theta} \ln L(y, \theta) = \frac{\frac{\partial}{\partial \theta} L(y, \theta)}{L(\mathrm{y}, \theta)},$$

we obtain

$$E_\theta \left[\frac{\partial}{\partial \theta} \ln L(\boldsymbol{y}, \theta) \right] = \int_Y \frac{\partial}{\partial \theta} \ln L(y, \theta) L(y, \theta) dy = \frac{\partial}{\partial \theta} \int_Y L(y, \theta) dy = \frac{\partial}{\partial \theta} 1 = 0$$

for all $\theta \in \Omega$. Hence, we have

$$I(\theta) = \operatorname{var}\left\{ \frac{\partial}{\partial \theta} \ln L(\boldsymbol{y}, \theta) \right\}. \tag{1.17}$$

Now let the second derivative of $\ln L(y, \theta)$ with respect to θ for all y and θ exist, and let $\int_Y L(\mathrm{y}, \theta) dP_\theta$ be differentiable twice, where integration and double differentiation can be commuted. Then by considering

$$\frac{\partial^2}{\partial \theta^2} \ln L(y, \theta) = \frac{L(y, \theta) \frac{\partial^2}{\partial \theta^2} L(y, \theta) - \left(\frac{\partial}{\partial \theta} L(\boldsymbol{y}, \theta) \right)^2}{(L(y, \theta))^2} = \frac{\frac{\partial^2}{\partial \theta^2} L(y, \theta)}{L(y, \theta)} - \left[\frac{\frac{\partial}{\partial \theta} L(\boldsymbol{y}, \theta)}{L(y, \theta)} \right]^2$$

and

$$0 = \frac{\partial^2}{\partial \theta^2} \int \ln L(y, \theta) dP_\theta = \int \frac{\partial^2}{\partial \theta^2} \ln L(y, \theta) dP_\theta,$$

the relation

$$E_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln L(\boldsymbol{y}, \theta) \right] = -E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \ln L(\boldsymbol{y}, \theta) \right]^2 \right\} = -I(\theta)$$

follows and therefore also

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln L(\boldsymbol{y}, \theta) \right]. \tag{1.18}$$

We present an example determining the Fisher information for both a discrete and a continuous distribution.

Example 1.11 Let P be the family of binomial distributions for given n and $\Omega = (0,1)$. The likelihood function is

$$L(y,p) = \binom{n}{y} p^y (1-p)^{n-y}.$$

The assumption V1 is satisfied, namely, the square of

$$\frac{\partial}{\partial p} \ln L(y,p) = \frac{y}{p} - \frac{n-y}{1-p}$$

after replacing y by random $\boldsymbol{y}$ has the finite expectation

$$I(p) = E_p \left\{ \left[\frac{\partial}{\partial p} \ln L(\boldsymbol{y},p) \right]^2 \right\} = \sum\nolimits_{y=0}^{n} \left(\frac{y}{p} - \frac{n-y}{1-p} \right)^2 \binom{n}{y} p^y (1-p)^{n-y},$$

and this means

$$I(p) = \frac{n}{p(1-p)}, 0 < p < 1.$$

Example 1.12 Let P be the family of normal distributions $N(\mu, \sigma^2)$ with known σ^2. It is $\Omega = R^1$, and the likelihood function has the form

$$L(y,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

For these distributions assumption V1 is fulfilled, too. We obtain

$$\frac{\partial}{\partial \mu} \ln L(y,\mu) = \frac{1}{\sigma^2}(y-\mu)$$

and

$$I(\mu) = \frac{1}{\sigma^4} E(\boldsymbol{y}-\mu)^2 = \frac{1}{\sigma^4} \operatorname{var}(\boldsymbol{y}) = \frac{1}{\sigma^2}.$$

Now we show the additivity of the Fisher information.

Theorem 1.6 If the Fisher information $I(\theta) = I_1(\theta)$ exists for a family P of probability distributions with $\Omega = R^1$ and if $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ is a random sample with components $\boldsymbol{y}_i$ (i = l, ..., n) all distributed as $P_i \in P$, then the Fisher information $I_n(\theta)$ of the distribution corresponding to $\boldsymbol{Y}$ is given by

$$I_n(\theta) = nI_1(\theta). \tag{1.19}$$

Proof: It follows from Definition 1.2 that the likelihood function $L_n(Y,\theta)$ of a random sample $\boldsymbol{Y}$ is

$$L_n(Y,\theta) = \prod_{i=1}^{n} L(y_i,\theta).$$

Consequently we get

$$\ln L_n(Y,\theta) = \sum_{i=1}^{n} \ln L(y_i,\theta)$$

and

$$\frac{\partial}{\partial\theta}\ln L_n(Y,\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\ln L(y_i,\theta).$$

Observing (1.17) we finally arrive at

$$I_n(\theta) = \operatorname{var}\left\{\frac{\partial}{\partial\theta}\ln L_n(\boldsymbol{Y},\theta)\right\} = \sum_{i=1}^{n} \operatorname{var}\left\{\frac{\partial}{\partial\theta}\ln L(\boldsymbol{y_i},\theta)\right\} = nI_1(\theta).$$

Theorem 1.7 Let $M(\boldsymbol{Y})$ be a sufficient statistic with respect to the distribution $P_\theta \in P, \boldsymbol{\Omega} \subseteq R^1$ of the components of the random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$. Let the distribution P_θ fulfil the condition V1 of Definition 1.10. Then the Fisher information

$$I_M(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln L_M(\boldsymbol{M},\theta)\right]^2\right\} \tag{1.20}$$

of $\boldsymbol{M} = M(\boldsymbol{Y})$ exists where $L_M(M,\theta)$ is the likelihood function of $\boldsymbol{M}$ and

$$I_n(\theta) = I_M(\theta). \tag{1.21}$$

Proof: Considering (1.2) and (1.3), respectively, we have

$$L(Y,\theta) = h(Y)g(M(Y),\theta)$$

and therefore

$$\frac{\partial}{\partial\theta}\ln L(Y,\theta) = \frac{\partial}{\partial\theta}\ln g(M(Y),\theta)$$

since h (Y) is by assumption independent of θ. Taking Corollary 1.1 of Theorem 1.1 into account, the likelihood function $L_M(M,\theta)$ of $\boldsymbol{M}$ satisfies also condition V1 of Definition 1.10, and therefore $I_M(\theta)$ in (1.20) exists. Observing the equivalence of $L_M(M,\theta)$ and $L(Y,\theta)$, the assertion follows because of $\frac{\partial}{\partial\theta}\ln L_M(M,\theta) = \frac{\partial}{\partial\theta}\ln g(M,\theta)$.

Consequently the Fisher information of a sufficient statistic is the Fisher information of the corresponding random sample.

Now we consider parameters $\theta \in \Omega \subseteq R^p$.

Definition 1.11 Let $\boldsymbol{y}$ be distributed as $P_\theta \in P, \Omega \subseteq R^p, \theta = (\theta_1, \ldots, \theta_p)^T$. Let the conditions 2, 3 and 4 of V1 in Definition 1.10 be fulfilled for each component θ_i ($i = 1, \ldots, p$). Let Ω be an open interval in R^p. Further, assume that the expectation of $\frac{\partial}{\partial\theta_i}\ln L(\boldsymbol{y},\theta)\frac{\partial}{\partial\theta_j}\ln L(\boldsymbol{y},\theta)$ exists for all θ and all $i, j = 1, \ldots, p$. Then the quadratic matrix

$$I(\theta) = (I_{i,j}(\theta)), \quad i,j = 1,\ldots,p$$

of order p given by

$$I(\theta) = E\left\{\frac{\partial}{\partial\theta_i}\ln L(\boldsymbol{y},\theta)\frac{\partial}{\partial\theta_j}\ln L(\boldsymbol{y},\theta)\right\}$$

is said to be the (Fisher) information matrix with respect to P_θ.

Example 1.13 Let the random variable $\boldsymbol{y}$ be distributed as $N(\mu, \sigma^2)$ where $\theta = (\mu,\sigma^2)^{\mathrm{T}} \in R^1 \times R^+ = \Omega$. Then

$$\ln L(y,\theta) = -\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y-\mu)^2$$

holds, and the assumption of Definition 1.11 is fulfilled with $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Further we have

$$\frac{\partial}{\partial\mu}\ln L(y,\theta) = \frac{y-\mu}{\sigma^2} \text{ and } \frac{\partial}{\partial\sigma^2}\ln L(y,\theta) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y-\mu)^2.$$

Because $E[(\boldsymbol{y}-\mu)^2] = \mathrm{var}(\boldsymbol{y}) = \sigma^2$, it follows $I_{11}(\theta) = \frac{1}{\sigma^2}$. Since the skewness $\gamma_1 = 0$ and since $E(\boldsymbol{y}-\mu) = 0$, we get $I_{12}(\theta) = I_{21}(\theta) = 0$. Further

$$\left[\frac{\partial}{\partial\sigma^2}\ln L(y,\theta)\right]^2 = \frac{1}{4\sigma^4} - \frac{2}{4\sigma^6}(y-\mu)^2 + \frac{1}{4\sigma^8}(y-\mu)^4.$$

Moreover, considering $\gamma_2 = 0$ and $E[(\boldsymbol{y}-\mu)^4] = 3\sigma^4$, it follows

$$I_{22} = E\left\{\left[\frac{\partial}{\partial\sigma^2}\ln L(\boldsymbol{y},\theta)\right]^2\right\} = \frac{1}{\sigma^4}\left[\frac{1}{4} - \frac{1}{2} + \frac{3}{4}\right] = \frac{1}{2\sigma^4},$$

and we obtain

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

On the other hand, if we put $\theta_1 = \mu$ and $\theta_2 = \sigma$, then we have

$$\frac{\partial}{\partial\sigma} \ln L(y,\theta) = -\frac{1}{\sigma} + \frac{1}{\sigma^2}(y-\mu)^2.$$

While I_{11}, I_{12} and I_{21} remain unchanged, we find now

$$I_{22} = E\left\{ \left[\frac{\partial}{\partial\sigma} \ln L(\boldsymbol{y},\theta) \right]^2 \right\} = \frac{1}{\sigma^2}[1-2+3] = \frac{2}{\sigma^2}$$

and therefore

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

This example shows that the Fisher information is not invariant with respect to parameter transformations. Using the chain rule of differential calculus, the following general statement arises.

Theorem 1.8 Let $\psi = h(\theta)$ be a monotone in $\Omega \subseteq R^1$, and with respect to θ differentiable function, let h map Ω onto Π. Then with respect to ψ, the differentiable inverse function $\theta = g(\psi)$ exists. Under the assumptions of Definition 1.10, let $I(\theta)$ be the Fisher information of the distribution $P_o \in \Pi$. Then the Fisher information $I^*(\psi)$ of the distribution P_ψ (i.e. P_θ written with the transformed parameter) is

$$I^*(\psi) = I(\theta)\left(\frac{d}{d\psi} g(\psi)\right)^2. \tag{1.22}$$

Considering Example 1.13 we set (for fixed μ) $\theta = \sigma^2$, $\psi = \sqrt{\theta} = \sigma$ and $\frac{d\theta}{d\psi} = 2\psi = 2\sigma$. By Theorem 1.8 we get with $I(\sigma^2) = \frac{1}{\sigma^2}$ the information

$$I^*(\psi) = I\left(\sigma^2\right) 4\sigma^2 = \frac{2}{\sigma^2}.$$

In Chapter 2 we need the following statement.

Theorem 1.9 (Inequality of Rao and Cramér)
Let assumption V1 of Definition 1.10 hold for all components of the random sample $\boldsymbol{Y}$ possessing the likelihood function $L(Y, \theta)$. Let the set $\{Y_0\} = \{Y \epsilon \{Y\} : L(Y, \theta) = 0\}$ of the points in the sample space satisfying $L(Y, \theta) = 0$ do not depend on θ. Let $P_\theta \in P = (P_\theta, \theta \in \Omega), \Omega \subseteq R^1$ be the distribution of the components, and let $M(\boldsymbol{Y})$ be a statistic with expectation $E[M(\boldsymbol{Y})]$ and variance $\mathrm{var}[M(\boldsymbol{Y})]$ mapping the sample space $\{Y\}$ into Ω. Then the inequality of Rao and Cramér

$$\mathrm{var}[M(\boldsymbol{Y})] \geq \frac{\left(\frac{dE[M(\boldsymbol{Y})]}{d\theta}\right)^2}{nI(\theta)} \tag{1.23}$$

is fulfilled.

Proof: With the notation $M(\boldsymbol{Y}) = \boldsymbol{M}$, we get $E = E[\boldsymbol{M} - E(\boldsymbol{M})] = 0$. Hence

$$\frac{dE}{d\theta} = -\int_{\{Y\}} \frac{dE[M(\boldsymbol{Y})]}{d\theta} dP_\theta + \int_{\{Y\}} (M - E(M)) \frac{d}{d\theta} L(Y, \theta) dP_\theta = 0$$

and

$$\frac{dE}{d\theta} = -\frac{dE[M(\boldsymbol{Y})]}{d\theta} \int_{\{Y\}} dP_\theta + \int_{\{Y\}} (M - E(M)) \frac{d}{d\theta} \ln L(Y, \theta) dY = 0$$

hold, respectively. Then

$$E\left\{(\boldsymbol{M} - E(\boldsymbol{M})) \frac{d}{d\theta} \ln L(Y, \theta)\right\} = \frac{dE(\boldsymbol{M})}{d\theta}$$

follows. Taking Schwarz's inequality into account, we arrive at

$$\begin{aligned}\left\{\frac{dE(\boldsymbol{M})}{d\theta}\right\}^2 &= \left\{E[(\boldsymbol{M} - E(\boldsymbol{M}))] \frac{d}{d\theta} \ln L(Y, \theta)\right\}^2 \\ &\leq E\left[(\boldsymbol{M} - E(\boldsymbol{M}))^2 E\left\{\left[\frac{d}{d\theta} \ln L(\boldsymbol{Y}, \theta)\right]^2\right\}\right].\end{aligned}$$

Considering (1.16) completes the proof.

When choosing $E(\boldsymbol{M}) = \theta$, the inequality of Rao and Cramér takes the form

$$\mathrm{var}[M(\boldsymbol{Y})] \geq \frac{1}{nI(\theta)}. \tag{1.24}$$

Theorem 1.10 If $\boldsymbol{y}$ is distributed as a one-parametric exponential family and if $g(\theta) = \eta = E(\boldsymbol{M})$, then

$$I^*(\eta) = \frac{1}{\mathrm{var}(\boldsymbol{M})}. \tag{1.25}$$

Proof: Since the assumption V1 of Definition 1.10 is fulfilled, $I(\eta)$ exists. Observing

$$\frac{d}{d\eta} \ln L(Y,\eta) = M(Y) - \frac{d}{d\eta} A(\eta)$$

and Theorem 1.8, we obtain

$$\mathrm{var}(\boldsymbol{M}) = I^*(\eta) = I(\theta)[\mathrm{var}(\boldsymbol{M})]^2.$$

Hence, the assertion is true.

Considering Schwarz's inequality for the second moments of statistics $M(\boldsymbol{Y})$ with finite second moment and an arbitrary function $h(\boldsymbol{Y},\theta)$ with existing second moment, then the inequality

$$\mathrm{var}(\boldsymbol{M}) \geq \frac{\mathrm{cov}^2[\boldsymbol{M}, h(\boldsymbol{Y},\theta)]}{\mathrm{var}[h(\boldsymbol{Y},\theta)]}$$

follows.

Theorem 1.11 Let $M(\boldsymbol{Y})$ be a statistic with expectation $g(\theta)$ and existing second moment, and let $\boldsymbol{h_j} = h_j(\boldsymbol{Y}, \theta)$, $j = 1, \ldots, r$ be functions with existing second moments. Then with the notations

$$c_j = \mathrm{cov}\left(M(\boldsymbol{Y}), \boldsymbol{h_j}\right), \sigma_{ij} = \mathrm{cov}\left(\boldsymbol{h_i}, \boldsymbol{h_j}\right), c^T = (c_1, \ldots, c_r)$$

and $\Sigma = (\sigma_{ij})$, $(|\Sigma| \neq 0)$, the inequality

$$\mathrm{var}(\mathbf{M}) \geq c^T \Sigma^{-1} c \tag{1.26}$$

is fulfilled.

Proof: The assertion follows from $\dfrac{c^T \Sigma^{-1} c}{\mathrm{var}(\boldsymbol{y})} \leq 1$.

With the help of (1.26), the inequality (1.24) of Rao and Cramér can be generalised to the p-dimensional case.

Theorem 1.12 Let the components of a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ be distributed as

$$P_\theta \in P = (P_\theta, \theta \in \Omega), \Omega \subseteq R^p, \theta^T = \left(\theta_1, \ldots, \theta_p\right), p > 1.$$

Let $L(Y,\theta)$ be the likelihood function of $\boldsymbol{Y}$. Further, let the assumptions of Definition 1.10 be fulfilled. Additionally, let the set of points in $\{Y\}$ satisfying $L(Y,\theta) = 0$ do not depend on θ. Finally, let $M(\boldsymbol{Y})$ be a statistic, whose expectation $\mathrm{E}[M(\boldsymbol{Y})] = w(\theta)$ exists and is differentiable with respect to all θ_i. Then the inequality

$$\mathrm{var}[M(\boldsymbol{Y})] \geq a^T I^{-1} a$$

holds, where I^{-1} is the inverse of $I(\theta)$ and a is the vector of the derivatives of $w(\theta)$ with respect to the θ_i.

Proof: As $I(\theta)$ is positive definite and therefore I^{-1} exists, the assertion follows with $h_j = \dfrac{d}{d\theta_j} L(Y,\theta)$ from (1.26) considering Definition 1.11.

1.5 Statistical Decision Theory

First, we formulate the general statistical decision problem. Let us start from the assumption that there is a set of random variables $\{\boldsymbol{y}_t\}$, $t \epsilon R^1$ whose distribution $P_\theta \in P = (P_\theta, \theta \in \Omega)$, $\dim\{\Omega\} = p$ is at least partly unknown.

Here we restrict ourselves to the case that only statements about $\psi = g(\theta)$ are demanded, where Ω is mapped by g onto Z and $\dim(Z) = s$. This set Z is called state space.

The statistician has for statements about ψ a set $\{E\}$ of decisions at his disposal. $\{E\}$ is called decision space. Let $Y_{t_i} = \left(\boldsymbol{y}_{t_i 1},\ldots,\boldsymbol{y}_{t_i n_i}\right)^T$ for each fixed t_i be a random sample of size n_i.

Let the set of results of an experiment for which a decision has to be made (i.e. to select from $\{E\}$) be with

$$N = \sum_{i=1}^{k} n_i,\ A_k = (Y_{t_1},\ldots,Y_{t_k}) \in \prod_{i=1}^{k} \{Y_{t_i}\} = \{Y_{k,N}\},$$

the realisation of a random variable $\boldsymbol{A}_k = (\boldsymbol{Y}_{t_1},\ldots,\boldsymbol{Y}_{t_k})$. Now let $d \in D$ be a measurable mapping from $\{Y_{k,N}\}$ onto E, which relates each A_k to a decision $d(A_k)$. Then d is called a decision function, and D is the set of admissible decision functions. A_k will depend on the distribution of $\boldsymbol{A}_k$, the support $\mathfrak{S}_k = (t_1,\ldots,t_k)$ of the experiment and its allocation vector $\mathfrak{N}_k = (n_1,\ldots,n_k)$. We denote the concrete design by

$$\begin{pmatrix} \mathfrak{S}_k \\ \mathfrak{N}_k \end{pmatrix} = \begin{pmatrix} t_1,\ldots,t_k \\ n_1,\ldots,n_k \end{pmatrix} \in V_N$$

belonging to a set V of admissible designs. Additionally, let a loss function L be given as measurable mapping from $E \times Z \times R^1$ into R^m (its definition and therefore that of m is a problem outside of mathematics), which means

$$L[d(A_k),\psi,f(M)], d(A_k) \in E, \psi \in Z \tag{1.27}$$

with a non-negative real function $f(M)$ where $M = (d, \mathfrak{S}_k, \mathfrak{N}_k)$.

The function Z registers the loss occurring if $d(A_k)$ is chosen, and ψ is the value in the transformed parameter space, while $f(M)$ corresponds to the costs for realising M. The task of statistics consists in providing methods for selection of triples $M = (d, \mathfrak{S}_k, \mathfrak{N}_k)$, which minimise a functional $R(d, \mathfrak{S}_k, \mathfrak{N}_k, \psi, f(M))$ of random loss called risk function R. We will denote by d either a decision function (for fixed n) or a sequence of decision functions of the same structure, whose elements differ only with respect to the sample size n. We assume that

$$R(d, \mathfrak{S}_k, \mathfrak{N}_k, \psi, f(M)) = F(d, \mathfrak{S}_k, \mathfrak{N}_k, \psi) + f(\mathfrak{S}_k, \mathfrak{N}_k), \tag{1.28}$$

where f does not depend on d and d^* is the decision function (sequence of decision functions) for which

$$\min_{d \in D} (d, \mathfrak{S}_k, \mathfrak{N}_k, \psi) = F(d^*, \mathfrak{S}_k, \mathfrak{N}_k, \psi) \tag{1.29}$$

is satisfied. Then the risk R can be minimised in two steps. First, d^* is determined in such a way that (1.29) is fulfilled, and second, $(\mathfrak{S}_k{}^*, \mathfrak{N}_k{}^*)$ is chosen to fulfil

$$R(d^*, \mathfrak{S}_k{}^*, \mathfrak{N}_k{}^*, \psi, f(\mathfrak{S}_k{}^*, \mathfrak{N}_k{}^*)) = \min_{\binom{\mathfrak{S}_k}{\mathfrak{N}_k} \in V} R(d^*, \mathfrak{S}_k, \mathfrak{N}_k, \psi, f(\mathfrak{S}_k, \mathfrak{N}_k)).$$

Definition 1.12 A triple $T^* \in V \times D$ is said to be locally R-optimal at the point $\psi_0 \in Z$ relative to $V \times D$ if for all $T \in V \times D$ the inequality

$$R[T^*, \psi_0, f(M^*)] \leq R[T, \psi_0, f(M)]$$

holds. If M^* is for all $\psi_0 \in Z$ locally R-optimal, then M^* is called global R-optimal.

Example 1.14 Let $k = 1$ and $\boldsymbol{y}_{t_1} = \boldsymbol{y}$ be distributed as $N(\mu, \sigma^2)$. Then $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \Omega = R^1 \times R^+$ and $\boldsymbol{A}_1 = \boldsymbol{Y}$. Further, let $\psi = g(\theta) = \mu \in R^1$ and $d(\boldsymbol{Y}) = \hat{\boldsymbol{\mu}}$ be a statistic with realisations in $R^1 = \{E\}$. Finally, let D be the class of statistics with finite second moment and realisations in R^1. Now we choose the loss function

$$L[\hat{\mu}, \mu, f(T)] = c_1(\hat{\mu} - \mu)^2 + c_2 nK, c_1, c_2, K > 0,$$

where K represents the costs of an experiment (a measurement). Besides we define as risk R the expected random loss

$$R(\hat{\boldsymbol{\mu}}, n, \mu, Kn) = E\left[c_1(\mu - \hat{\boldsymbol{\mu}})^2 + c_2 nK\right] = c_2 nK + c_1\left[\mathrm{var}(\hat{\boldsymbol{\mu}}) + B(\hat{\boldsymbol{\mu}})^2\right],$$

where $B(\hat{\boldsymbol{\mu}}) = E(\hat{\boldsymbol{\mu}}) - \mu$. In the class D the choice $\hat{\mu} = \psi_0$ is together with $n = 0$ locally R-optimal for the decision, and for (ψ_0, n) the risk R is equal to 0. The class D can be restricted to exclude this unsatisfactory trivial case. We denote by $D_E \subseteq D$ the subset in D with $B(\hat{\boldsymbol{\mu}}) = 0$. Then we obtain

$$R(\hat{\boldsymbol{\mu}}, n, \mu, Kn) = c_2 nK + c_1 \operatorname{var}(\hat{\boldsymbol{\mu}}), \hat{\boldsymbol{\mu}} \in D_E$$

in the form (1.28). We will see in Chapter 2 that $\operatorname{var}(\hat{\boldsymbol{\mu}})$ becomes minimal for $\hat{\boldsymbol{\mu}}^* = \bar{\boldsymbol{y}}$.

Since for a random sample of n elements we have $\operatorname{var}(\bar{\boldsymbol{y}}) = \dfrac{\sigma^2}{n}$, the first step of minimising R leads to

$$\min_{d \in D_E} c_1 \operatorname{var}(\hat{\boldsymbol{\mu}}) = \frac{c_1}{n}\sigma^2$$

and

$$R(\hat{\boldsymbol{\mu}}, n, \mu, Kn) = c_2 Kn + \frac{c_1}{n}\sigma^2.$$

If we derive the right-hand side of the equation with respect to n and put the derivative equal to 0, then we get $n^* = \sigma\sqrt{\dfrac{c_1}{Kc_2}}$, and this as well as $\bar{\boldsymbol{y}}$ does not depend on $\psi = \mu$. The convexity of the considered function shows that we have indeed found a (global) minimum. Hence, the R-optimal solution of the decision problem in $E \times Z$ (which is in Z global, but in Ω only local because of the dependence on σ) is given by

$$M^* = \left(\bar{\boldsymbol{y}}, n^* = \sigma\sqrt{\frac{c_1}{Kc_2}}\right).$$

If we choose $\psi = g(\theta) = \sigma^2 > 0$, then we obtain $E = R^+$, $k = 1$, $\boldsymbol{A}_1 = \boldsymbol{Y}$ and $N = n$. Let the loss function be

$$L\left[d(\boldsymbol{Y}), \sigma^2, f(M)\right] = c_1\left(\sigma^2 - d\right)^2 + c_2 nK, c_i > 0, K > 0.$$

If we take again

$$R\left(d(\boldsymbol{Y}), n, \sigma^2, Kn\right) = \mathrm{R} = \mathrm{E}(\boldsymbol{L}) = c_1 E\left\{\left(\sigma^2 - d(\boldsymbol{Y})\right)^2\right\} + c_2 nK$$

as risk function, it is of the form (1.28). If we restrict ourselves to $d \in D_E$ by analogous causes as in the previous case such that $E[d(\boldsymbol{Y})] = \sigma^2$ holds, then the first summand of R is minimal for

$$d(\boldsymbol{Y}) = \boldsymbol{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2,$$

which will be seen in Chapter 2.

Since $\frac{\boldsymbol{s}^2}{\sigma^2}(n-1)$ is distributed as CS $(n-1)$ and has therefore the variance $2(n-1)$, we find

$$\operatorname{var}(\boldsymbol{s}^2) = \frac{2\sigma^4}{n-1}.$$

The first step of optimisation supplies

$$R(\boldsymbol{s}^2, n, \sigma^2, Kn) = c_1 \frac{2\sigma^4}{n-1} + c_2 nK.$$

The R-optimal n is given by

$$n^* = 1 + \sigma^2 \sqrt{\frac{2c_1}{Kc_2}}.$$

The locally R-optimal solution of the decision problem is

$$M^* = \left(\boldsymbol{s}^2, n^* = 1 + \sigma^2 \sqrt{\frac{2c_1}{Kc_2}}\right).$$

We consider more detailed theory and further applications in the next chapters where the selection of minimal sample sizes is discussed. We want to assume that d has to be chosen for fixed $\mathfrak{S}_k$ and $\mathfrak{N}_k$ R-optimal relative to a certain risk function. Concerning the optimal choice of $\begin{pmatrix} \mathfrak{S}_k \\ \mathfrak{N}_k \end{pmatrix}$, we refer to Chapters 8 and Chapter 9 treating regression analysis. We write therefore with τ from Definition 1.13

$$R(\boldsymbol{d}, \psi) = E\{L[d(\boldsymbol{Y}), \psi]\} = r(d, \tau). \tag{1.30}$$

In Example 1.14 a restriction to a subset $D_E \subseteq D$ was carried out to avoid trivial locally R-optimal decision functions d. Now two other general procedures are introduced to overcome such problems.

Definition 1.13 Let $\boldsymbol{\theta}$ be a random variable with realisations $\theta \in \Omega$ and the probability distribution $P_\tau, \tau \epsilon \mathfrak{T}$. Let the expectation

$$\int_{\Omega} R(d, \psi) d\Pi_\tau = r(d, \tau) \tag{1.31}$$

of (1.30) exist relative to P_τ, which is called Bayesian risk relative to the a priori distribution P_τ.

A decision function $d_0(\boldsymbol{Y})$ that fulfils

$$r(d_0, \tau) = \inf_{d \epsilon D} [r(d, \tau)],$$

is called Bayesian decision function relative to the a priori distribution P_τ.

Definition 1.14 A decision function $d_0 \in D$ is said to be minimax decision function if

$$\max_{\theta \in \Omega} R(d_0, \psi) = \min_{d \in D} \max_{\theta \in \Omega} R(d, \psi). \tag{1.32}$$

Definition 1.15 Let $d_1, d_2 \in D$ be two decision functions for a certain decision problem with the risk function $R(d, \psi)$ where $\psi = g(\theta), \theta \in \Omega$. Then d_1 is said not to be worse than d_2, if $R(d_1, \psi) \leq R(d_2, \psi)$ holds for all $\theta \in \Omega$. The function d_1 is said to be better than the function d_2 if apart from $R(d_1, \psi) \leq R(d_2, \psi)$ for all $\theta \in \Omega$ at least for one $\theta^* \in \Omega$ the strong inequality $R(d_1, \psi^*) < R(d_2, \psi^*)$ holds where $\psi^* = g(\theta^*)$. A decision function d is called admissible in D if there is no decision function in D better than d. If a decision function is not admissible, it is called inadmissible.

In this chapter it is not necessary to develop the decision theory further. In Chapter 2 we will consider the theory of point estimation, where $d(\boldsymbol{Y}) = S(\boldsymbol{Y})$ is the decision function. Regarding the theory of testing in Chapter 3, the probability for the rejection of a null hypothesis and in the confidence interval estimation a domain in Ω covering the value θ of the distribution P_θ with a given probability is the decision function $d(\boldsymbol{Y})$. Selection rules and multiple comparison methods are further special cases of decision functions.

1.6 Exercises

1.1 For estimating the average income of the inhabitants of a city, the income of owners of each 20th private line in a telephone directory is determined. Is this sample a random sample with respect to the whole city population?

1.2 A set with elements 1, 2, 3 is considered. Selecting elements with replacement, there are $3^4 = 81$ different samples of size $n = 4$. Write down all possible samples, calculate $\bar{y}$ and s^2 and present the frequency distribution of $\bar{y}$ and s^2 graphically as a bar chart. (You may use a program package.)

1.3 Prove that the statistic $M(\boldsymbol{Y})$ is sufficient relative to θ, where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$, $n \geq 1$ is a random sample from a population with distribution $P_\theta, \theta \in \Omega$, by determining the conditional distribution of $\boldsymbol{Y}$ for given $M(\boldsymbol{Y})$.

a) $M(\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{y}_i$ and P_θ is the Poisson distribution with the parameter $\theta \in \Omega \subset R^+$.

b) $M(\boldsymbol{Y}) = (\boldsymbol{y}_{(1)}, \boldsymbol{y}_{(n)})^T$ and P_θ is the uniform distribution in the interval $(\theta, \theta + 1)$ with $\theta \in \Omega \subset R^1$.

c) $M(\boldsymbol{Y}) = \boldsymbol{y}_{(n)}$ and P_θ is the uniform distribution in the interval $(0, \theta)$ with $\theta \in \Omega = R^+$.

d) $M(\boldsymbol{Y}) = \sum_{i=1}^{n} \boldsymbol{y}_i$ and P_θ is the exponential distribution with the parameter $\theta \in \Omega = R^+$.

1.4 Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$, $n \geq 1$ be a random sample from a population with the distribution P_θ, $\theta \in \Omega$. Determine a sufficient statistic with respect to θ using Corollary 1.1 of the decomposition theorem if P_θ, $\theta \in \Omega$ is the density function

a) $(y, \theta) = \theta y^{\theta - 1}$, $0 < y < 1$; $\theta \in \Omega = R^+$,
b) Of the Weibull distribution

$$f(y,\theta) = \theta a(\theta y)^{a-1} e^{-(\theta y)^a}, y \geq 0, \theta \in \Omega = R^+, a > 0 \text{ known}$$

c) Of the Pareto distribution

$$f(y,\theta) = \frac{\theta a^\theta}{y^{\theta+1}}, y > a > 0, \theta \in \Omega = R^+ \text{ known}$$

1.5 Determine a minimal sufficient statistic $M(\boldsymbol{Y})$ for the parameter θ, if $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$, $n \geq 1$ is a random sample from a population with the following distribution P_θ:

a) Geometric distribution with the probability function

$$p(y,p) = p(1-p)^{y-1}, \; y = 1,2,\ldots, 0 < p < 1$$

b) Hypergeometric distribution with the probability function

$$p(y,M,N,n) = \frac{\binom{M}{y}\binom{N-M}{n-y}}{\binom{N}{n}}, n \in \{1,\ldots,N\}, y \in \{0,\ldots,N\}; M \leq N \text{ integer},$$

c) Negative binomial distribution with the probability function

$$\mathrm{p}(y,p,r) = \binom{y-1}{r-1} p^r (1-p)^{y-r}, 0 < p < 1, y \geq r \text{ integer}, r \in \{0,1,\ldots\}$$

and
i) θ = p and r known;
ii) $\theta^T = (p, r)$,

d) Beta distribution with the density function

$$f(y,\theta) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}, 0 < y < 1, 0 < a, b < \infty$$

and
i) $\theta = a$ and b known;
ii) $\theta = b$ and a known

1.6 Prove that the following distribution families $\{P_\theta, \theta \in \Omega\}$ are complete:

a) P_θ is the Poisson distribution with the parameter $\theta \in \Omega = R^+$.
b) P_θ is the uniform distribution in the interval $(0, \theta)$, $\theta \in \Omega = R^+$.

1.7 Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$, $n \geq 1$ be a random sample, whose components are uniformly distributed in the interval $(0, \theta)$, $\Omega = R^+$. Show that $M(\boldsymbol{Y}) = \boldsymbol{y}_{(n)}$ is complete sufficient.

1.8 Let the variable $\boldsymbol{y}$ have the discrete distribution P_θ with the probability function

$$p(y,\theta) = P(\boldsymbol{y} = y) = \begin{cases} \theta & \text{for } y = -1 \\ (1-\theta)^2\theta^y & \text{for } y = 0,1,2,\ldots \end{cases}.$$

Show that the corresponding distribution family with $\theta \in (0, 1)$ is bounded complete, but not complete.

1.9 Let a one-parametric exponential family with the density or probability function

$$f(y,\theta) = h(y)\, e^{\eta(\theta)M(y) - B(\theta)}, \theta \in \Omega$$

be given.

a) Express the Fisher information of this distribution by using the functions $\eta(\theta)$ and $B(\theta)$.
b) Use the result of a) to calculate $I(\theta)$ for the
   i) Binomial distribution with the parameter $\theta = p$
   ii) Poisson distribution with the parameter $\theta = \lambda$
   iii) Exponential distribution with the parameter θ
   iv) Normal distribution $N(\mu, \sigma^2)$ with $\theta = \sigma$ and μ fixed

1.10 Let the assumptions of Definition 1.11 be fulfilled. Besides, the second partial derivatives $\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(y,\theta)$ are to exist for all $i, j = 1, \ldots, p$ and $y \in \{Y\}$ as well as their expectations for random $\boldsymbol{y}$. Moreover, let $\int_{\{Y\}} L(y, \theta)\, dy$ be twice differentiable, where integration and differentiation are commutative.

Prove that in this case the elements of the information matrix in Definition 1.11 have the form

a) $I_{i,j}(\theta) = \operatorname{cov}\left[\frac{\partial}{\partial\theta_i}\ln L(\boldsymbol{y},\theta), \frac{\partial}{\partial\theta_j}\ln L(\boldsymbol{y},\theta)\right]$,

b) $I_{i,j}(\theta) = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln L(\boldsymbol{y},\theta)\right]$,

respectively.

1.11 Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$ be a random sample from a population with the distribution $P_\theta, \theta \in \Omega$ and $M(\boldsymbol{Y})$ a given statistic.

Calculate $E[M(\boldsymbol{Y})]$, var$[M(\boldsymbol{Y})]$, the Fisher information $I(\theta)$ of the distribution and the Rao–Cramér bound for var$[M(\boldsymbol{Y})]$.

Does equality hold in the inequality of Rao and Cramér under the following assumptions?

a) P_θ is the Poisson distribution with the parameter $\theta \in R^+$ and

$$\mathrm{M}(\mathbf{Y}) = \begin{cases} 1 & \text{for } y = 0 \\ 0 & \text{else} \end{cases},$$

(here we have $n = 1$, i.e. $\boldsymbol{y} = \boldsymbol{Y}$).

b) P_θ is the Poisson distribution with the parameter $\theta \in R^+$ and $M(\boldsymbol{Y}) = \left(1 - \frac{1}{n}\right)^{n\bar{y}}$, (generalisation of a) for the case $n > 1$).

c) P_θ is the distribution with the density function

$$f(y, \theta) = \theta y^{\theta - 1}, 0 < y < 1, \theta \in R^+$$

and $M(\boldsymbol{Y}) = -\frac{1}{n}\sum_{i=1}^{n} \ln \boldsymbol{y}_i$.

1.12 In a certain region it is intended to drill for oil. The owner of drilling rights has to decide between strategies from $\{E_1, E_2, E_3\}$.

The following notations are introduced:

E_1 – The drilling is carried out under its own direction.
E_2 – The drilling rights are sold.
E_3 – A part of drilling rights are alienated.

It is not known so far if there really is an oil deposit in the region. Further let be $\Omega = \{\theta_1, \theta_2\}$ with the following meanings:

$\theta = \theta_1$ – Oil occurs in the region.
$\theta = \theta_2$ – Oil does not occur in the region.

The loss function $L(d, \theta)$ has for the decisions $d = E_i, i = 1, 2, 3$ and $\theta = \theta_j, j = 1, 2$ the form

	E_1	E_2	E_3
θ_1	0	10	5
θ_2	12	1	6

The decision is made considering expert's reports related to the geological situation in the region. We denote the result of the reports by $y \in \{0, 1\}$.

Let $p_\theta(y)$ be the probability function of the random variable $\boldsymbol{y}$ – in dependence on θ – with values

	y = 0	y = 1
θ_1	0.3	0.7
θ_2	0.6	0.4

Therefore the variable $\boldsymbol{y}$ states the information obtained by the 'random experiment' of geological reports about existing ($y = 1$) or missing ($y = 0$) deposits of oil in the region. Let the set D of decision functions $d(\boldsymbol{y})$ contain all possible 3^2 discrete functions:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$d_i(0)$	E_1	E_1	E_1	E_2	E_2	E_2	E_3	E_3	E_3
$d_i(1)$	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3

a) Determine the risk $R(d(\boldsymbol{y}), \theta) = E_\theta[L\{d(\boldsymbol{y})\}, \theta]$ for all 18 above given cases.
b) Determine the minimax decision function.
c) Following the opinion of experts in the field of drilling technology, the probability of finding oil after drilling in this region is approximately 0.2. Then $\boldsymbol{\theta}$ can be considered as random variable with the probability function

$\boldsymbol{\theta}$	θ_1	θ_2
$\pi(\theta)$	0.2	0.8

Determine for each decision function the Bayesian risk $r(d_i, \pi)$ and then the Bayesian decision function.

1.13 The strategies of treatment using two different drugs M_1 and M_2 are to be assessed. Three strategies are at the disposal of doctors:

E_1 – Treatment with the drug M_1 increasing blood pressure
E_2 – Treatment without using drugs
E_3 – Treatment with the drug M_2 decreasing blood pressure

Let the variable θ characterise the (suitable transformed) blood pressure of a patient such that $\theta < 0$ indicates too low blood pressure, $\theta = 0$ normal

blood pressure and $\theta > 0$ too high blood pressure. The loss function is defined as follows:

	E_1	E_2	E_3
$\theta < 0$	0	c	b + c
$\theta = 0$	b	0	b
$\theta > 0$	b + c	c	0

The blood pressure of a patient is measured. Let the measurement $\boldsymbol{y}$ be distributed as $N(\theta, 1)$ and let it happen n-times independently from each other: $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)^T$. Based on this sample the decision function

$$d_{r,s} = \begin{cases} E_1, \text{ if } \bar{y} < r \\ E_2, \text{ if } r \le \bar{y} \le s \\ E_3, \text{ if } \bar{y} > s \end{cases}$$

is defined.

a) Determine the risk $R(d_{r,s}(\bar{\boldsymbol{y}}), \theta) = E\{L[d_{r.s}(\bar{\boldsymbol{y}}), \theta]\}$.
b) Sketch the risk function in the case $b = c = 1$, $n = 1$ for
   i) $r = -s = -1$;
   ii) $r = -\frac{1}{2}s = -1$.

For which values of θ the decision function $d_{-1,1}(y)$ should be preferred to the function $d_{-1,2}(y)$?

References

Bahadur, R. R. (1955) Statistics and subfields. *Ann. Math. Stat.*, **26**, 490–497.

Cochran, W. G. and Boing, W. (1972) *Stichprobenverfahren*, De Gruyter, Berlin, New York.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

Halmos, P. R. and Savage, L. J. (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225–241.

Kauermann, G. and Küchenhoff, H. (2011) *Stichproben und praktische Umsetzung mit R*, Springer, Heidelberg.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Lehmann, E. L. and Romano J. P. (2008) *Testing Statistical Hypothesis*, Springer, Heidelberg.

Quatember, A. (2014) *Datenqualität in Stichprobenerhebungen*, Springer, Berlin.
Rasch, D. (1995) *Mathematische Statistik*, Johann Ambrosius Barth, Berlin, Heidelberg.
Rasch, D., Herrendörfer, G., Bock, J., Victor, N. and Guiard, V. (Eds.) (2008) *Verfahrensbibliothek Versuchsplanung und - auswertung*, 2. verbesserte Auflage in einem Band mit CD, R. Oldenbourg Verlag, München, Wien.
Stigler, S. M. (1986, 1990) *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge.