Translations of

MATHEMATICAL MONOGRAPHS

Volume 229

Lectures in Mathematical Statistics

Parts 1 and 2

Yu. N. Lin′kov

American Mathematical Society

Lectures in Mathematical Statistics

Parts 1 and 2

Translations of

MATHEMATICAL MONOGRAPHS

Volume 229

Lectures in Mathematical Statistics

Parts 1 and 2

Yu. N. Lin'kov

Translated by Oleg Klesov and Vladimir Zayats

American Mathematical Society
Providence, Rhode Island

Ю. Н. Линьков

ЛЕКЦИИ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

"ИСТОКИ", ДОНЕЦК, 2001

This work was originally published in Russian by Istoki, Donetsk under the title "Лекции по математической статистике, части 1, 2" © Ю. Н. Линьков, 1999. The present translation was created under license for the American Mathematical Society and is published by permission.

Translated from the Russian by Oleg Klesov and Vladimir Zayats.

2000 *Mathematics Subject Classification.* Primary 62-01.

For additional information and updates on this book, visit
www.ams.org/bookpages/mmono-229

Library of Congress Cataloging-in-Publication Data

Lin'kov, IU. N.

[Lektsii po matematicheskoi statistike. English]

Lectures in mathematical statistics : parts 1 and 2 / Yu. N. Lin'kov ; translated by Oleg Klesov and Vladimir Zayats.

p. cm. – (Translations of mathematical monographs, ISSN 0065-9282 ; v. 229)

Includes bibliographical references and index.

ISBN 0-8218-3732-X (alk. paper)

1. Mathematical statistics. I. Title II. Series

QA276.16.L5513 2005

519.5–dc22 2005052661

10 9 8 7 6 5 4 3 2 1 10 09 08 07 06 05

Contents

Foreword to the English Translation

Parts 1 and 2 of "Lectures in Mathematical Statistics" by Yu. N. Lin'kov were originally published in Russian as two separate books. For the English translation, the two parts are combined into one book. Each part has its own preface and list of references, with chapters, sections, theorems, etc., numbered independently in each part.

Part 1

Preface to Part 1

The author's idea was that this textbook should be aimed at students of mathematics having a background in general university courses in probability theory and mathematical statistics. The textbook was written based on the courses given by the author for students of mathematical departments at Volgograd University, Volgograd, Russia, and Donetsk University, Donetsk, Ukraine.

Among the books which may be used as a first reading in mathematical statistics, we mention the books by Cramér [9] and van der Waerden [34], which have already become the cornerstones in statistics. These books are still an authority, and many generations of experts have been brought up with these books. Elements of mathematical statistics are an essential ingredient of other general courses on probability theory. Let us mention the textbooks by Gnedenko [12], Gikhman, Skorokhod, and Yadrenko [11], Rozanov [27], Sevast'yanov [29], Tutubalin [33], and Shiryaev [30]. The textbooks by Shmetterer [31], Ivchenko and Medvedev [14], and Kozlov and Prokhorov [19] can be regarded as thoroughly developed introductions into mathematical statistics. The books on mathematical statistics by Borovkov [5], [6] take a special rank among textbooks for undergraduate and postgraduate students.

In writing this book, the author has used Russian and foreign literature on mathematical statistics, as well as the experience and traditions of teaching probability at Volgograd University and Donetsk University. Let us mention here the books by Rao [26], Cox and Hinkley [8], van der Waerden [34], and Bickel and Doksum [4] that thoroughly work out, each in its own way, problems for teaching mathematical statistics.

Part 1 of this book begins with a presentation of sampling using one-dimensional samples (Chapter 1) and multidimensional samples (Chapter 2) as an example. The basic sample characteristics are introduced and their asymptotic and nonasymptotic properties are studied. Main distributions related to the multidimensional Gaussian distribution are defined.

Chapter 3 deals with the estimation of parameters of distributions. In this chapter, measures of quality of statistical estimators are introduced and some optimality criteria are given. Optimal estimation of a scale parameter and a location parameter is studied. For regular families of distributions, approaches leading to effective estimators based on the Cramér–Rao inequality are given.

Chapter 4 deals with the theory of sufficient statistics and its applications to the construction of optimal estimators of unknown parameters and parametric functions.

In Chapter 5, general methods for constructing statistical estimators of parameters of distributions are considered and the main properties of the corresponding estimators are established.

The limited size of the book did not allow us to include some important statistical procedures or to consider other topics in the theory of parametric estimation. Part 2 of the textbook will deal with problems related to testing statistical hypotheses. The author hopes that this textbook will enable the reader to work independently, using other sources, on the topics we only touch upon here. We would recommend the books by Wilks [**35**] and Lehmann [**21**] and the three-volume monograph by Kendall and Stuart [**16**]–[**18**]. Our textbook can be used in preparation for general courses on mathematical statistics as well as specialized courses on the subject.

The list of references at the end of Part 1 includes only references available for students in Russia and Ukraine and is by no means complete.

In the textbook, we use the common notational conventions: P and P_θ for probabilities; E and E_θ for mathematical expectations; D and D_θ for variances, etc. We use triple notation for theorems, lemmas, formulas, etc. Therefore, for example, Theorem 4.1.2 refers to Theorem 2 in Section 1 of Chapter 4. Sections are enumerated by double numbers: Section 1.4 stands for Section 4 in Chapter 1. The sign $\square$ marks the end of a proof.

Samples from One-Dimensional Distributions

1.1. Empirical distribution function and its asymptotic behavior

Empirical distribution function. Order statistics. Let ξ be a real-valued random variable with the distribution function

$$F(x) = \mathsf{P}\{\xi < x\}, \qquad x \in \mathbf{R} = (-\infty, \infty).$$

Let $\xi_1, \xi_2, \ldots, \xi_n$ be n independent observations of the random variable ξ. Therefore $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables whose distribution function coincides with that of the random variable ξ, that is,

$$\mathsf{P}\{\xi_i < x\} = F(x)$$

for all $i = 1, 2, \ldots, n$. Denote by $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ the vector of observations (also called a *sample*).

Given $x \in \mathbf{R}$, introduce the random variable

$$\nu_n(x) = \sum_{i=1}^{n} I_{(-\infty,x)}(\xi_i)$$

where $I_A(x)$ is the indicator of a set A. The function

$$F_n(x) = \nu_n(x)/n, \qquad x \in \mathbf{R}, \tag{1.1.1}$$

is called the *empirical distribution function.*

We rearrange the observations $\xi_1, \xi_2, \ldots, \xi_n$ in ascending order and denote the resulting random variables by

$$\zeta_{n,1} \le \zeta_{n,2} \le \cdots \le \zeta_{n,n}. \tag{1.1.2}$$

The terms of this sequence are called *order statistics.*

Note that the empirical distribution function F_n possesses all the properties of regular distribution functions, namely it

(1) assumes values in the interval $[0, 1]$,

(2) does not decrease, and

(3) is left-continuous.

Note also that F_n is a step function whose jumps are at the points $\zeta_{n,1}, \ldots, \zeta_{n,n}$. If all the observations of a sample $\xi^{(n)}$ are different (in which case all the inequalities in (1.1.2) are strict), then $F_n(x)$ has n jumps whose heights are $1/n$. In the general case, equalities may appear in (1.1.2) (in which case the function $F_n(x)$ may have less than n jumps; however the jumps are proportional to $1/n$).

Glivenko's theorem. Definition (1.1.1) of the empirical distribution function implies that $F_n(x)$ is the frequency of the random event $\{\xi < x\}$ in n independent observations. Given x the probability of $\{\xi < x\}$ is constant and equals $F(x)$. By the Bernoulli theorem (the law of large numbers for Bernoulli trials) the empirical distribution function $F_n(x)$ tends in probability to $F(x)$ as $n \to \infty$, that is,

$$\lim_{n\to\infty} \mathsf{P}\{|F_n(x) - F(x)| > \varepsilon\} = 0 \quad \text{for all } \varepsilon > 0. \tag{1.1.3}$$

Moreover, by the Borel theorem (the strong law of large numbers for Bernoulli trials) $F_n(x)$ tends with probability 1 to $F(x)$ as $n \to \infty$, that is,

$$\mathsf{P}\left\{\lim_{n\to\infty} F_n(x) = F(x)\right\} = 1. \tag{1.1.4}$$

The convergence in relations (1.1.3) and (1.1.4) holds for every fixed $x \in \mathbf{R}$. However the following (stronger) Glivenko (1933) result claims that the convergence of $F_n(x)$ to $F(x)$ is, in fact, uniform with respect to x.

THEOREM 1.1.1 (Glivenko).

$$\mathsf{P}\left\{\lim_{n\to\infty} \sup_{x\in\mathbf{R}} |F_n(x) - F(x)| = 0\right\} = 1. \tag{1.1.5}$$

PROOF. Let $x_{r,k}$ be the minimal number x for which

$$F(x) \le \frac{k}{r} \le F(x+0) \tag{1.1.6}$$

where $r = 1, 2, \ldots$ and $k = 0, 1, 2, \ldots, r$. If the system of inequalities (1.1.6) does not have solutions for $k = 0$, then we put $x_{r,0} = -\infty$. Similarly, if (1.1.6) does not have solutions for $k = r$, then we put $x_{r,r} = \infty$. Consider random events

$$E_k^r = \left\{\lim_{n\to\infty} |F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)| = 0\right\}$$

where $a \vee b$ stands for the maximum of two numbers a and b. We also put

$$E^r = \bigcap_{k=0}^{r} E_k^r.$$

It is clear that

$$E^r = \left\{\lim_{n\to\infty} \max_{0\le x\le r} \left(|F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)|\right) = 0\right\}.$$

We have $\mathsf{P}(E_k^r) = 1$ for all $k = 0, 1, \ldots, r$ by the Borel theorem. Thus $\mathsf{P}(E^r) = 1$. Let $E = \bigcap_{r=1}^{\infty} E^r$. Since $\mathsf{P}(E^r) = 1$ for all $r \ge 1$, we get $\mathsf{P}(E) = 1$.

Now let k be such that $x_{r,k} < x_{r,k+1}$ and $x \in (x_{r,k}, x_{r,k+1}]$. Then

$$F_n(x_{r,k}+0) \le F_n(x) \le F_n(x_{r,k+1}), \tag{1.1.7}$$

$$F(x_{r,k}+0) \le F(x) \le F(x_{r,k+1}). \tag{1.1.8}$$

It is clear that

$$F(x_{r,k+1}) - F(x_{r,k}+0) \le \frac{1}{r}. \tag{1.1.9}$$

Inequalities (1.1.7) and (1.1.8) together with (1.1.9) yield

$$\begin{aligned} F_n(x) - F(x) &\le F_n(x_{r,k+1}) - F(x_{r,k}+0) \\ &= [F_n(x_{r,k+1}) - F(x_{r,k+1})] + [F(x_{r,k+1}) - F(x_{r,k}+0)] \\ &\le \max_{0\le k\le r} (|F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)|) + \frac{1}{r}. \end{aligned} \tag{1.1.10}$$

Using the same argument we derive from (1.1.7)–(1.1.9) that

(1.1.11)

$$\begin{aligned} &F_n(x) - F(x) \\ &\quad \ge - \max_{0\le k\le r} \big(|F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)|\big) - \frac{1}{r}. \end{aligned}$$

Combining (1.1.10) and (1.1.11) we obtain for all $x \in (x_{r,k}, x_{r,k+1}]$

$$\begin{aligned} &|F_n(x) - F(x)| \\ &\quad \le \max_{0\le k\le r} \big(|F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)|\big) + \frac{1}{r}. \end{aligned} \tag{1.1.12}$$

Since the right-hand side of inequality (1.1.12) does not depend on k, it holds for all $x \in \mathbf{R}$. Thus

$$\begin{aligned} &\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \\ &\quad \le \max_{0\le k\le r} \big(|F_n(x_{r,k}) - F(x_{r,k})| \vee |F_n(x_{r,k}+0) - F(x_{r,k}+0)|\big) + \frac{1}{r} \end{aligned} \tag{1.1.13}$$

for all $r \ge 1$.

Since inequality (1.1.13) holds for all $r \ge 1$, we obtain

$$E \subset \left\{ \lim_{n\to\infty} \sup_{x\in\mathbf{R}} |F_n(x) - F(x)| = 0 \right\},$$

whence (1.1.5) follows in view of $\mathsf{P}(E) = 1$. □

Relations (1.1.3) and (1.1.4) as well as the Glivenko theorem indicate that the empirical distribution function $F_n(x)$ may serve as an approximation of the original distribution function $F(x)$.

Asymptotic normality of the empirical distribution function and Kolmogorov's theorem. According to definition (1.1.1) the empirical distribution function $F_n(x)$ for a fixed x is a random variable assuming values k/n for $k = 0, 1, 2, \dots, n$. Moreover

$$\mathsf{P}\{F_n(x) = k/n\} = \binom{n}{k} F^k(x)(1 - F(x))^{n-k},$$

whence

$$\mathsf{E}F_n(x) = F(x), \qquad \mathsf{D}F_n(x) = F(x)(1 - F(x))/n.$$

We say that a sequence of random variables η_n, $n = 1, 2, \dots$, is *asymptotically normal with parameters* (A_n, B_n^2) if

$$\lim_{n\to\infty} \mathsf{P}\left\{ \frac{\eta_n - A_n}{B_n} < x \right\} = \Phi(x) \quad \text{for all } x \in \mathbf{R} \tag{1.1.14}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2}\, dz \tag{1.1.15}$$

is the distribution function of the standard normal law $\mathcal{N}(0,1)$, that is, the normal law with mean 0 and variance 1.

By the central limit theorem for Bernoulli trials we obtain the following assertion on the asymptotic normality of the empirical distribution function.

THEOREM 1.1.2. *For every fixed $x \in \mathbf{R}$, the sequence of empirical distribution functions $F_n(x)$, $n = 1, 2, \ldots$, is asymptotically normal with parameters*

$$F(x) \quad \text{and} \quad \frac{F(x)(1-F(x))}{n}.$$

Consider the random variable

$$D_n = \sup_{x \in \mathbf{R}} |F_n(x) - F(x)|$$

measuring the deviation between the empirical distribution function $F_n(x)$ and the distribution function $F(x)$ in the uniform metric.

The following result by Kolmogorov (1933) allows one to estimate, for large n, that the probability D_n differs from zero.

THEOREM 1.1.3 (Kolmogorov). *If the distribution function $F(x)$ is continuous, then*

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{n} D_n < z\right\} = K(z) = \sum_{j=-\infty}^{\infty} (-1)^j \exp\left\{-2j^2 z^2\right\}$$

for all $z > 0$.

The function $K(z)$ is called the *Kolmogorov distribution function.*

1.2. Sample characteristics and their properties

Sample moments. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample, that is, $\xi_1, \ldots, \xi_n$ are independent observations of a random variable ξ with the distribution function $F(x)$. Denote by α_k the *k-th moment* of the random variable ξ (in other words, the k-th moment of the distribution function $F(x)$), that is, $\alpha_k = \mathsf{E}\xi^k$. By μ_k we denote the *k-th central moment* of the random variable ξ (in other words, the k-th central moment of the distribution function $F(x)$), that is, $\mu_k = \mathsf{E}(\xi - \alpha_1)^k$. Note that α_1 is the expectation (or *mean value*) of the random variable ξ, while μ_2 is its variance. We also note that $\mu_1 = 0$ and $\mu_2 = \alpha_2 - \alpha_1^2$. Moreover, the moments and central moments are related to each other as follows:

$$\mu_k = \sum_{j=0}^{k} \binom{k}{j} (-1)^j \alpha_1^j \alpha_{k-j}. \tag{1.2.1}$$

Similar characteristics can be introduced for the empirical distribution function $F_n(x)$ constructed from the sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$. The k-th moment of the

empirical distribution function $F_n(x)$ is called the *k-th sampling moment*, that is,

$$a_k = \int x^k \, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} \xi_i^k. \tag{1.2.2}$$

The k-th central moment of the empirical distribution function $F_n(x)$ is called the *k-th sampling central moment*, that is,

$$m_k = \int (x - a_1)^k dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - a_1)^k. \tag{1.2.3}$$

From (1.2.2) and (1.2.3) we obtain a relation between sampling moments and sampling central moments:

$$m_k = \sum_{j=0}^{k} \binom{k}{j} (-1)^j a_1^j a_{k-j}. \tag{1.2.4}$$

Expectation and variance of sampling moments. It is clear that, for all k,

$$\mathsf{E} a_k = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E} \xi_i^k = \alpha_k, \tag{1.2.5}$$

$$\mathsf{D} a_k = \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{D} \xi_i^k = \frac{1}{n} (\alpha_{2k} - \alpha_k^2) \tag{1.2.6}$$

if the corresponding moments exist. The evaluation of the expectation and variance of higher sampling central moments is a more complicated problem.

THEOREM 1.2.1. *If* $\alpha_{2k} < \infty$, *then*

$$\mathsf{E} m_k = \mu_k + O\left(\frac{1}{n}\right), \tag{1.2.7}$$

$$\mathsf{D} m_k = \frac{1}{n} \left(\mu_{2k} - 2k\mu_{k-1}\mu_{k+1} - \mu_k^2 + k^2 \mu_2 \mu_{k-1}^2\right) + O\left(\frac{1}{n^2}\right). \tag{1.2.8}$$

PROOF. Consider the random variables $\tilde{\xi}_i = \xi_i - \alpha_1$, $i = 1, 2, \ldots, n$, and put

$$\tilde{a}_j = \frac{1}{n} \sum_{i=1}^{n} \tilde{\xi}_i^j.$$

Note that $\mathsf{E}\tilde{\xi}_i = 0$, $i = 1, 2, \ldots, n$, $\mathsf{E}\tilde{a}_1 = 0$, and $\mathsf{E}\tilde{a}_j = \mu_j$. Applying (1.2.4) we get

$$m_k = \tilde{a}_k - k\tilde{a}_1 \tilde{a}_{k-1} + \sum_{j=2}^{k} \binom{k}{j} (-1)^j \tilde{a}_1^j \tilde{a}_{k-j}. \tag{1.2.9}$$

Since the random variables $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n$ are independent and $\mathsf{E}\tilde{\xi}_i = 0$, we obtain

$$\mathsf{E}\tilde{a}_1 \tilde{a}_{k-1} = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathsf{E}\tilde{\xi}_i \tilde{\xi}_j^{k-1} = \frac{1}{n} \mu_k. \tag{1.2.10}$$

By the Cauchy–Bunyakovskiĭ inequality we get that

$$\left|\mathsf{E}\tilde{a}_1^j\tilde{a}_{k-j}\right| \le \left(\mathsf{E}\tilde{a}_1^{2j}\mathsf{E}\tilde{a}_{k-j}^2\right)^{1/2} \tag{1.2.11}$$

for $j \ge 2$. On the other hand,

$$\begin{aligned}\mathsf{E}\tilde{a}_{k-j}^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{E}\tilde{\xi}_i^{2(k-j)} + \frac{1}{n^2}\sum_{i\ne l}\mathsf{E}\tilde{\xi}_i^{k-j}\tilde{\xi}_l^{k-j}\\ &= \frac{1}{n}\mu_{2(k-j)} + \frac{n-1}{n}\mu_{k-j}^2 = \mu_{k-j}^2 + \frac{1}{n}\left(\mu_{2(k-j)} - \mu_{k-j}^2\right),\end{aligned}$$

whence $\mathsf{E}\tilde{a}_{k-j}^2 \le \mu_{2(k-j)}$ in view of $\mu_{2(k-j)} - \mu_{k-j}^2 \ge 0$. Thus inequality (1.2.11) can be rewritten for an arbitrary $j \ge 2$ as follows:

$$\left|\mathsf{E}\tilde{a}_1^j\tilde{a}_{k-j}\right| \le \left(\mu_{2(k-j)}\mathsf{E}\tilde{a}_1^{2j}\right)^{1/2}. \tag{1.2.12}$$

Further,

$$\mathsf{E}\tilde{a}_1^{2j} = \frac{1}{n^{2j}}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_{2j}=1}^{n}\mathsf{E}\tilde{\xi}_{i_1}\tilde{\xi}_{i_2}\ldots\tilde{\xi}_{i_{2j}}. \tag{1.2.13}$$

Consider the terms on the right-hand side of (1.2.13) containing a factor $\tilde{\xi}_{i_l}$ whose index i_l differs from those of other factors. All such terms vanish because $\mathsf{E}\tilde{\xi}_i = 0$ and the random variables $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n$ are independent.

Now consider the terms on the right-hand side of (1.2.13) whose indices i_1, $i_2, \ldots, i_{2j}$ fall into j pairs of equal numbers. The number of elements of this set is N_1N_2 where N_1 is the number of ways to split the set $\{1, 2, \ldots, 2j\}$ into j pairs and N_2 is the number of ways to choose different j numbers from the set $\{1, 2, \ldots, n\}$. Obviously

$$N_1N_2 = \prod_{k=0}^{j}(2j-2k+1)\cdot\prod_{l=0}^{j-1}(n-l) = O\left(n^j\right), \qquad n\to\infty.$$

Note also that the cardinality of the set of all other terms in (1.2.13) can also be represented as a polynomial of n whose degree is less than j. Thus we obtain from (1.2.13) that

$$\mathsf{E}\tilde{a}_1^{2j} = O\left(n^{-j}\right) \tag{1.2.14}$$

for all $j \ge 2$. Combining (1.2.9), (1.2.10), (1.2.12), and (1.2.14) and taking into account that $\mathsf{E}\tilde{a}_k = \mu_k$, we prove (1.2.7).

Equality (1.2.7) implies that

$$\mathsf{D}m_k = \mathsf{E}(m_k-\mu_k)^2 + O\left(n^{-2}\right). \tag{1.2.15}$$

It follows from (1.2.9) that

$$\begin{aligned}\mathsf{E}(m_k-\mu_k)^2 &= \mathsf{E}(\tilde{a}_k-\mu_k)^2 + 2\sum_{j=1}^{k}\binom{k}{j}(-1)^j\mathsf{E}\tilde{a}_1^j\tilde{a}_{k-j}(\tilde{a}_k-\mu_k)\\ &\quad + \sum_{i,j=1}^{k}\binom{k}{i}\binom{k}{j}(-1)^{i+j}\mathsf{E}\tilde{a}_1^{i+j}\tilde{a}_{k-i}\tilde{a}_{k-j}.\end{aligned} \tag{1.2.16}$$

Applying (1.2.6) to $\tilde{\xi}_1, \ldots, \tilde{\xi}_n$ we obtain

$$\mathsf{E}(\tilde{a}_k - \mu_k)^2 = \mathsf{D}\tilde{a}_k = n^{-1}\left(\mu_{2k} - \mu_k^2\right), \tag{1.2.17}$$

since $\mathsf{E}\tilde{a}_k = \mu_k$. Using the same arguments as those applied for the evaluation of the right-hand side of (1.2.13) we prove that

$$\begin{aligned} \mathsf{E}\tilde{a}_1\tilde{a}_{k-1}(\tilde{a}_k - \mu_k) &= \frac{1}{n^3}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\mathsf{E}\tilde{\xi}_{i_1}\tilde{\xi}_{i_2}^{k-1}(\tilde{\xi}_{i_3} - \mu_k) \\ &= \frac{1}{n}\mu_{k-1}\mu_{k+1} + O\left(\frac{1}{n^2}\right), \end{aligned} \tag{1.2.18}$$

$$\begin{aligned} \mathsf{E}\tilde{a}_1^2\tilde{a}_{k-1}^2 &= \frac{1}{n^4}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\sum_{i_4=1}^{n}\mathsf{E}\tilde{\xi}_{i_1}\tilde{\xi}_{i_2}\tilde{\xi}_{i_3}^{k-1}\tilde{\xi}_{i_4}^{k-1} \\ &= \frac{1}{n}\mu_2\mu_{k-1}^2 + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{1.2.19}$$

The same method shows that

$$\sum_{j=2}^{k}\binom{k}{j}(-1)^j\mathsf{E}\tilde{a}_1^j\tilde{a}_{k-j}(\tilde{a}_k - \mu_k) = O\left(\frac{1}{n^2}\right), \tag{1.2.20}$$

$$\sum_{i+j\geq 3}\binom{k}{i}\binom{k}{j}(-1)^{i+j}\mathsf{E}\tilde{a}_1^{i+j}\tilde{a}_{k-i}\tilde{a}_{k-j} = O\left(\frac{1}{n^2}\right). \tag{1.2.21}$$

Combining (1.2.15)–(1.2.21) we easily obtain relation (1.2.8). □

Convergence in probability of sampling moments. We study the asymptotic behavior (as $n \to \infty$) of sampling moments a_k and m_k defined by (1.2.2) and (1.2.3), respectively. To indicate the dependence of moments a_k and m_k on the size of the sample we write a_{nk} and m_{nk}, respectively. Using (1.2.6) and the Chebyshev inequality we prove that $a_{nk} \to \alpha_k$ in probability as $n \to \infty$. A similar assertion holds for sampling central moments and even for arbitrary continuous functions of a finite number of sampling moments a_{nk} (the sampling central moment m_{nk} is a polynomial of moments $a_{n1}, a_{n2}, \ldots, a_{nk}$ in view of (1.2.4)). The following result contains a precise statement of the latter assertion.

THEOREM 1.2.2. *Let random variables $\zeta_n^{(1)}, \zeta_n^{(2)}, \ldots, \zeta_n^{(k)}$ converge in probability to some constants $c_1, c_2, \ldots, c_k$, respectively, as $n \to \infty$. Let a function $f(z_1, z_2, \ldots, z_k)$ be continuous in a neighborhood of the point $(c_1, c_2, \ldots, c_k)$. Then the random variables $\eta_n = f\left(\zeta_n^{(1)}, \zeta_n^{(2)}, \ldots, \zeta_n^{(k)}\right)$ converge to $f(c_1, c_2, \ldots, c_k)$ in probability as $n \to \infty$.*

PROOF. Let $\varepsilon > 0$ be an arbitrary number. Since $f(z_1, z_2, \ldots, z_k)$ is continuous in a neighborhood of the point $(c_1, c_2, \ldots, c_k)$, there is a number $\delta = \delta(\varepsilon)$ such that

$$|f(z_1, z_2, \ldots, z_k) - f(c_1, c_2, \ldots, c_k)| < \varepsilon$$

for $|z_i - c_i| < \delta$, $i = 1, 2, \ldots, k$.

Consider the random events $B_i = \{|\zeta_n^{(i)} - c_i| < \delta\}$, $i = 1, 2, \ldots, k$. Then $B \subset C$ where

$$B = \bigcap_{i=1}^{k} B_i, \qquad C = \{|\eta_n - f(c_1, c_2, \ldots, c_k)| < \varepsilon\}.$$

Thus

$$\mathsf{P}(C) \geq \mathsf{P}(B) = 1 - \mathsf{P}\left(\bigcup_{i=1}^{k} \overline{B}_i\right) \geq 1 - \sum_{i=1}^{k} \mathsf{P}\left(\overline{B}_i\right). \tag{1.2.22}$$

Since $\zeta_n^{(i)}$ converges in probability to c_i, for a given $\delta > 0$ and all $\gamma > 0$ there is $n_i = n_i(\gamma)$ such that $\mathsf{P}(\overline{B}_i) < \gamma/k$ for $n \geq n_i$. Then $\mathsf{P}(\overline{B}_i) < \gamma/k$ for

$$n \geq n_0 = \max(n_1, \ldots, n_k)$$

and all $i = 1, 2, \ldots, k$. Therefore $\mathsf{P}(C) \geq 1 - \gamma$ by (1.2.22) if $n \geq n_0$. □

Consider another application of Theorem 1.2.2. For continuous random variables we define the *skewness* γ_1 and *excess* γ_2 by

$$\gamma_1 = \mu_3 \mu_2^{-3/2}, \qquad \gamma_2 = \mu_4 \mu_2^{-2} - 3. \tag{1.2.23}$$

If the density of a distribution is symmetric, then $\gamma_1 = 0$. Moreover, $\gamma_2 = 0$ in the case of the Gaussian distribution. Starting from (1.2.23) we construct the sampling skewness $g_1 = g_{n1}$ and sampling excess $g_2 = g_{n2}$ from the sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$:

$$g_{n1} = m_{n3} m_{n2}^{-3/2}, \qquad g_{n2} = m_{n4} m_{n2}^{-2} - 3. \tag{1.2.24}$$

Applying Theorem 1.2.2 and equality (1.2.4), we prove that the sampling skewness and excess defined by (1.2.24) converge in probability as $n \to \infty$ to the corresponding skewness and excess defined by (1.2.23) for a given random variable.

Asymptotic normality of sampling moments. We introduce the following notation. The *law of distribution* of a random variable ξ is denoted by $\mathcal{L}(\xi)$. The law of distribution of the normal random variable ξ with expectation $\alpha = \mathsf{E}\xi = \alpha_1$ and variance $\sigma^2 = \mathsf{D}\xi = \mu_2$ is denoted by $\mathcal{L}(\xi) = \mathcal{N}(\alpha, \sigma^2)$.

We say that a sequence of random variables η_n, $n = 1, 2, \ldots$, *weakly converges* as $n \to \infty$ to a random variable η if $\mathcal{L}(\eta_n) \to \mathcal{L}(\eta)$ as $n \to \infty$ (the convergence of laws $\mathcal{L}(\eta_n) \to \mathcal{L}(\eta)$ is understood as the convergence of distribution functions $\mathsf{P}\{\eta_n < x\}$ to a distribution function $\mathsf{P}\{\eta < x\}$ at all points of continuity of the function $\mathsf{P}\{\eta < x\}$). In particular, the asymptotic normality of a sequence η_n with parameters (A_n, B_n^2) defined by (1.1.14) and (1.1.15) means that

$$\mathcal{L}((\eta_n - A_n)/B_n) \to \mathcal{N}(0, 1), \qquad n \to \infty.$$

In the latter case we also say that a sequence η_n is $\mathcal{N}(A_n, B_n^2)$ asymptotically normal and occasionally write $\mathcal{L}(\eta_n) \sim \mathcal{N}(A_n, B_n^2)$.

The sampling moment a_{nk} is the sum of n independent identically distributed random variables (see (1.2.2)).

Applying the central limit theorem, we obtain the following result.

THEOREM 1.2.3. *If $\alpha_{2k} < \infty$, then the sequence a_{nk} of order k sampling moments is $\mathcal{N}(\alpha_k, (\alpha_{2k} - \alpha_k^2)/n)$ asymptotically normal.*

PROOF. It follows from (1.2.2) that

$$\sqrt{\frac{n}{\alpha_{2k}-\alpha_k^2}}(a_{nk}-\alpha_k)=\frac{1}{\sqrt{n(\alpha_{2k}-\alpha_k^2)}}\left(\sum_{i=1}^{n}\xi_i^k-n\alpha_k\right)=\eta_n.$$

Taking into account equalities (1.2.5) and (1.2.6) and applying the central limit theorem to the sum $\sum_{i=1}^{n}\xi_i^k$, we obtain $\mathcal{L}(\eta_n)\to\mathcal{N}(0,1)$. Therefore the sequence a_{nk} is $\mathcal{N}(\alpha_k,(\alpha_{2k}-\alpha_k^2)/n)$ asymptotically normal. □

Similarly to Theorem 1.2.3 one can prove the asymptotic normality of a continuous function of a finite number of sampling moments a_{nk}. In particular,

THEOREM 1.2.4. *If $\alpha_{2k<\infty}$, then the sequence m_{nk} of order k sampling central moments is*

$$\mathcal{N}\left(\mu_k,\left(\mu_{2k}-2k\mu_{k-1}\mu_{k+1}-\mu_k^2+k^2\mu_2\mu_{k-1}^2\right)/n\right)$$

asymptotically normal.

1.3. Order statistics and their properties

The distribution of order statistics. Let ξ be a random variable with the distribution function $F(x)$, let $\xi^{(n)}=(\xi_1,\xi_2,\ldots,\xi_n)$ be observations of ξ, and let $\zeta_{n,1},\zeta_{n,2},\ldots,\zeta_{n,n}$ be order statistics constructed from the sample ξ^n defined by (1.1.2). We study the distribution $F_{n,k}(x)=\mathsf{P}\{\zeta_{n,k}<x\}$ of the k-th order statistic $\zeta_{n,k}$. It is clear that $\{\zeta_{n,k}<x\}=\{\nu_n(x)\geq k\}$ where $\nu_n(x)$ is the number of random variables in the sequence $\xi_1,\xi_2,\ldots,\xi_n$ such that $\{\xi_k<x\}$. One can treat $\nu_n(x)$ as the number of occurrences of the event $\{\xi<x\}$ in n independent Bernoulli trials (see Section 1.1). Since $\mathsf{P}\{\xi<x\}=F(x)$, the binomial distribution of $\nu_n(x)$ shows that

$$F_{n,k}(x)=\mathsf{P}\{\nu_n(x)\geq k\}=\sum_{j=k}^{n}\binom{n}{j}F^j(x)(1-F(x))^{n-j}. \tag{1.3.1}$$

The following result on the integral representation of the function $F_{n,k}$ is helpful for studies of its asymptotic properties.

THEOREM 1.3.1. *The distribution function $F_{n,k}(x)$ admits the following representation:*

$$F_{n,k}(x)=n\binom{n-1}{k-1}\int_0^{F(x)}t^{k-1}(1-t)^{n-k}\,dt. \tag{1.3.2}$$

In particular, if F has the density $f(x)=F'(x)$, then so does $F_{n,k}(x)$. Denote the density of $F_{n,k}(x)$ by $f_{n,k}(x)$ if the density f exists. Then

$$f_{n,k}(x)=n\binom{n-1}{k-1}F^{k-1}(x)(1-F(x))^{n-k}f(x). \tag{1.3.3}$$

PROOF. Evaluating the integral in (1.3.2) by parts, we obtain

$$\int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,dt = \frac{1}{k}F^k(x)(1-F(x))^{n-k} + \frac{n-k}{n}\int_0^{F(x)} t^k(1-t)^{n-k-1}\,dt. \tag{1.3.4}$$

Substituting (1.3.4) into (1.3.2) we get

$$n\binom{n-1}{k-1}\int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,dt = \binom{n}{k}F^k(x)(1-F(x))^{n-k} + n\binom{n-1}{k}\int_0^{F(x)} t^k(1-t)^{n-k-1}\,dt. \tag{1.3.5}$$

Evaluating the integral on the right-hand side of (1.3.5) by parts, we conclude that

$$n\binom{n-1}{k-1}\int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,dt = \sum_{j=k}^{n}\binom{n}{j}F^j(x)(1-F(x))^{n-j},$$

whence (1.3.2) follows by (1.3.1).

If the density $f(x) = F'(x)$ exists, then (1.3.3) follows from (1.3.2). □

The joint distribution of two order statistics, say $\zeta_{n,k}$ and $\zeta_{n,m}$ with $k < m$, is also easy to evaluate. In particular, if the density $f(x) = F'(x)$ exists, then the density of the joint distribution of $\zeta_{n,k}$ and $\zeta_{n,m}$ also exists. Denoting it by $f_{n;k,m}(x,y)$ we have for $x < y$

$$f_{n;k,m}(x,y) = n(n-1)\binom{n-2}{k-1}\binom{n-k-1}{m-1}F^{k-1}(x) \times (1-F(y))^{m-1}(F(y)-F(x))^{n-(k+m)}f(x)f(y). \tag{1.3.6}$$

To prove (1.3.6), we consider two disjoint intervals $[x, x+\Delta x)$ and $[y, y+\Delta y)$ for small Δx and Δy. Then we evaluate the probability of the event that exactly $k-1$ random variables of the sequence $\xi_1, \xi_2, \ldots, \xi_n$ belong to the interval $(-\infty, x)$; only one random variable belongs to $[x, x+\Delta x)$; $m-1$ random variables belong to $[y+\Delta y, \infty)$; only one random variable belongs to $[y, y+\Delta y)$; and all other random variables belong to $[x+\Delta x, y)$. The probability of this event equals the right-hand side of (1.3.6) multiplied by $\Delta x \Delta y$ with a remainder term of a higher order with respect to $\Delta x \Delta y$. It is not hard to show that the probabilities of other events favorable to $\{\zeta_{n,k} \in [x, x+\Delta x), \zeta_{n,m} \in [y, y+\Delta y)\}$ are of higher orders with respect to $\Delta x \Delta y$ as compared to the probability of the event discussed above.

Limit theorems for extreme order statistics. Consider the k-th order statistic $\zeta_{n,k}$ whose index $k = k(n)$ depends on n in such a way that $k(n)/n$ approaches either 0 or 1 as $n \to \infty$. Those order statistics are called *extremes* or *extreme order statistics*. Below we study the cases of $k(n) = k = \text{const}$ and $k(n) = n - m + 1$ where m does not depend on n. In other words, we study the k-th order statistic from the left and m-th order statistic from the right for fixed constants k and m. Consider the limit behavior of extremes $\zeta_{n,k}$ and $\zeta_{n,n-m+1}$.

Let

$$\eta_n = nF(\zeta_{n,k}), \qquad \kappa_n = n[1 - F(\zeta_{n,n-m+1})]. \tag{1.3.7}$$

The following characteristics of a distribution function play an important role in the theorems below:

$$\overline{x}_0 = \sup\{x: F(x) = 0\}, \qquad \underline{x}_1 = \inf\{x: F(x) = 1\}.$$

We agree that $\sup(\varnothing) = -\infty$ and $\inf(\varnothing) = \infty$.

THEOREM 1.3.2. *Assume that there exists $x' > \overline{x}_0$ such that $F(x)$ is continuous in the interval $(-\infty, x')$ and increases in the interval $(\overline{x}_0, x')$. Then*

$$\lim_{n\to\infty} \mathsf{P}\{\eta_n < y\} = \Gamma_k(y) \tag{1.3.8}$$

for all $y \in \mathbf{R}$ where the random variables η_n are defined by (1.3.7), *while $\Gamma_k(y)$ is given by*

$$\Gamma_k(y) = \begin{cases} \frac{1}{(k-1)!} \int_0^y z^{k-1} e^{-z}\, dz, & y > 0, \\ 0, & y \le 0. \end{cases} \tag{1.3.9}$$

PROOF. Equality (1.3.8) is obvious for $y \le 0$. Fix a number $y > 0$ and let n be such that $y/n < F(x')$. Then the inverse function $F^{-1}(y)$ exists in the interval $(\overline{x}_0, x')$. Applying (1.3.2) we get

$$\mathsf{P}\{\eta_n < y\} = \mathsf{P}\left\{\zeta_{n,k} < F^{-1}\left(\frac{y}{n}\right)\right\} = n\binom{n-1}{k-1} \int_0^{y/n} t^{k-1}(1-t)^{n-k}\, dt.$$

Changing the variable $z = nt$ we obtain

$$\mathsf{P}\{\eta_n < y\} = \binom{n-1}{k-1} \int_0^y \left(\frac{z}{n}\right)^{k-1} \left(1 - \frac{z}{n}\right)^{n-k} dz.$$

Note that $n^{-(k-1)}\binom{n-1}{k-1} \to 1/(k-1)!$ and $(1 - z/n)^{n-k} \to e^{-z}$ as $n \to \infty$, and moreover the convergence in the second relation is uniform with respect to $z \in (0, y)$ for an arbitrary finite y. This implies relation (1.3.8) by the Lebesgue dominated convergence theorem. □

THEOREM 1.3.3. *Assume that there exists $x'' < \underline{x}_1$ such that $F(x)$ is continuous in the interval (x'', ∞) and increases in the interval $(x'', \underline{x}_1)$. Then*

$$\lim_{n\to\infty} \mathsf{P}\{\kappa_n < y\} = \Gamma_m(y)$$

for all $y \in \mathbf{R}$ where the random variables κ_n are defined by (1.3.7), *while $\Gamma_m(y)$ is defined by* (1.3.9).

PROOF. Note that

$$\begin{aligned} \mathsf{P}\{\kappa_n < y\} &= \mathsf{P}\left\{\zeta_{n,n-m+1} > F^{-1}\left(1 - \frac{y}{n}\right)\right\} \\ &= n\binom{n-1}{m-1} \int_{1-y/n}^1 t^{n-m}(1-t)^{m-1}\, dt \\ &= n\binom{n-1}{m-1} \int_0^{y/n} z^{m-1}(1-z)^{n-m}\, dz \end{aligned}$$

if n is sufficiently large. The rest of the proof is the same as that of Theorem 1.3.2. □

The function $\Gamma_k(y)$ defined by (1.3.9) is the so-called Gamma distribution function with parameter k.

The asymptotic behavior of extremes is a complicated problem in the case of a general k. A rather complete solution of this problem is given by Gnedenko (1943) and Smirnov (1949) (see, for example, [**32**]). Below we briefly discuss some of their results.

Consider the random variables $\tilde{\zeta}_{n,k} = (\zeta_{n,k} - A_n)/B_n$ where $k = \text{const}$ and A_n and $B_n > 0$ are appropriate constants depending on n. The possible limit distributions for $\tilde{\zeta}_{n,k}$ can only be of the following three types:

$$\psi^{(1)}_{k,\alpha}(x) = \begin{cases} 0, & x \le 0, \\ \Gamma_k(x^\alpha), & x > 0, \end{cases} \qquad \psi^{(2)}_{k,\alpha}(x) = \begin{cases} \Gamma_k(|x|^{-\alpha}), & x \le 0, \\ 1, & x > 0, \end{cases}$$

$$\psi^{(3)}_k(x) = \Gamma_k(e^x), \qquad -\infty < x < \infty,$$

where $\alpha > 0$. Necessary and sufficient conditions for convergence to any of the three types of limit laws are known in terms of the distribution function $F(x)$. Similar results are also obtained for extremes $\zeta_{n,n-m+1}$ with $m = \text{const}$.

To this end, we note that one can obtain the limit distribution for the pair of random variables η_n and κ_n defined by (1.3.7). Below is the corresponding result.

THEOREM 1.3.4. *Assume that all the assumptions of Theorems* 1.3.2 *and* 1.3.3 *hold. Then*

$$\lim_{n\to\infty} \mathsf{P}\{\eta_n < x, \kappa_n < y\} = \Gamma_k(x)\Gamma_m(y)$$

for all $x < y$ where the function $\Gamma_k(x)$ is defined by (1.3.9).

Central order statistics and sampling quantiles. If $k = k(n)$ depends on n in such a way that $k(n)/n \to p$ as $n \to \infty$ and $0 < p < 1$, then the order statistic $\zeta_{n,k(n)}$ is called *central*. Sampling quantiles of a distribution can be expressed in terms of central order statistics.

Let $p \in (0,1)$. Any number x_p such that

$$F(x_p) \le p \quad \text{and} \quad F(x_p + 0) \ge p \tag{1.3.10}$$

is called a *p-quantile of a distribution $F(x)$*. It is clear that the system of inequalities (1.3.10) has at least one solution. A p-quantile of the empirical distribution function $F_n(x)$ is called a *sampling p-quantile* and is denoted by $\widehat{x}_p$.

THEOREM 1.3.5. *For all $p \in (0,1)$, the sampling p-quantile can be represented as follows*:

$$\widehat{x}_p = \begin{cases} \zeta_{n,[np]+1}, & \text{if } np \text{ is not an integer}, \\ \text{any number of the interval } [\zeta_{n,np}, \zeta_{n,np+1}], & \text{if } np \text{ is an integer}, \end{cases}$$

where $[a]$ stands for the integer part of a number a.

PROOF. Assume that np is not an integer. Generally speaking, there are order statistics $\zeta_{n,k}$ with $k \le [np] + 1$ and such that $\zeta_{n,k} = \zeta_{n,[np]+1}$, whence

$$F_n\left(\zeta_{n,[np]+1}\right) \le [np]/n < p.$$

On the other hand, there are order statistics $\zeta_{n,k}$ with $k \ge [np] + 1$ and such that $\zeta_{n,k} = \zeta_{n,[np]+1}$, thus $F_n(\zeta_{n,[np]+1} + 0) \ge ([np]+1)/n > p$. Therefore $\widehat{x}_p = \zeta_{n,[np]+1}$.

Now let np be an integer and let $\widehat{x}_p$ be any number of the interval $[\zeta_{n,np}, \zeta_{n,np+1}]$. As above $F_n(\widehat{x}_p) \le np/n = p$ and $F_n(\widehat{x}_p + 0) \ge np/n = p$, that is, $\widehat{x}_p$ is a sampling p-quantile. □

Below we use the notation $\widehat{x}_{n,p}$ for $\widehat{x}_p$ to highlight that the sampling p-quantile depends on n.

Convergence in probability of the sampling p-quantiles. The consideration below excludes the cases of $p = 0$ and $p = 1$. These two cases require special treatment. If $p = 0$, then $\widehat{x}_{n,0} \in (-\infty, \zeta_{n,1}]$ and therefore either $-\infty$ is the unique p-quantile or there are infinitely many p-quantiles.

Below we consider sampling p-quantiles $\widehat{x}_{n,p}$ for $p \in (0,1)$. The following result contains conditions for the convergence in probability of sampling quantiles to the corresponding quantiles x_p.

THEOREM 1.3.6. *If a p-quantile x_p is unique, then the sampling p-quantile $\widehat{x}_{n,p}$ converges in probability to x_p as $n \to \infty$.*

PROOF. It is obvious that $F(x_p + \varepsilon) \ge F(x_p + 0) \ge p$ for all $\varepsilon > 0$. Since a p-quantile x_p is unique, $F(x_p + \varepsilon) > p$ for all $\varepsilon > 0$. The definition of the sampling p-quantile implies that $\{F_n(x_p + \varepsilon) > p\} \subset \{\widehat{x}_{n,p} \le x_p + \varepsilon\}$. Therefore

$$\mathsf{P}\{F_n(x_p + \varepsilon) > p\} \le \mathsf{P}\{\widehat{x}_{n,p} \le x_p + \varepsilon\}. \tag{1.3.11}$$

By the law of large numbers (1.1.3)

$$\mathsf{P}\{F_n(x_p + \varepsilon) > p\} = \mathsf{P}\{F_n(x_p + \varepsilon) - F(x_p + \varepsilon) > -(F(x_p + \varepsilon) - p)\} \to 1$$

as $n \to \infty$. Using inequality (1.3.11) we get for all $\varepsilon > 0$ that

$$\lim_{n\to\infty} \mathsf{P}\{\widehat{x}_{n,p} \le x_p + \varepsilon\} = 1. \tag{1.3.12}$$

Using again the uniqueness of a p-quantile x_p and the same argument we obtain for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}\{\widehat{x}_{n,p} \ge x_p - \varepsilon\} = 1. \tag{1.3.13}$$

Relations (1.3.12) and (1.3.13) mean that $\widehat{x}_{n,p}$ converges in probability to x_p as $n \to \infty$. □

REMARK 1.3.1. In fact, $\widehat{x}_{n,p}$ converges with probability 1 to x_p as $n \to \infty$ under the assumptions of Theorem 1.3.6.

REMARK 1.3.2. Let $\zeta_{n,k(n)}$ be a central order statistic such that $k(n)/n \to p$ as $n \to \infty$. Assume that a p-quantile x_p is unique. Then one can show by using Theorem 1.3.6 that $\zeta_{n,k(n)} \to x_p$ in probability as $n \to \infty$.

Asymptotic normality of sampling quantiles. Below we provide conditions for the asymptotic normality of a sampling p-quantile for $0 < p < 1$. First we give the following central limit theorem for independent identically distributed random variables in the scheme of series.

THEOREM 1.3.7 (LINDEBERG). *Let $\xi_{n1}, \xi_{n2}, \dots, \xi_{nn}$ be independent identically distributed random variables such that*

$$\mathsf{E}\xi_{ni} = 0, \quad i = 1, \dots, n, \qquad \sum_{i=1}^{n} \mathsf{E}\xi_{ni}^2 = 1, \quad n \geq 1.$$

Then the sequence of random variables $\sum_{i=1}^{n} \xi_{ni}$ is $\mathcal{N}(0,1)$ asymptotically normal if and only if

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{E}\xi_{ni}^2 I(|\xi_{ni}| > \tau) = 0 \tag{1.3.14}$$

for all $\tau > 0$ where $I(A)$ is the indicator of a random event A.

The proof of Theorem 1.3.7 can be found in [**23**], p. 292.

THEOREM 1.3.8. *Let a distribution function F be continuous and let the equation $F(x) = p$ have a unique solution x_p. Moreover let the function $F(x)$ be differentiable at the point x_p and $F'(x_p) = f(x_p) > 0$. Then the sampling p-quantile $\widehat{x}_{n,p}$ is $\mathcal{N}(x_p, n^{-1}pqf^{-2}(x_p))$ normal as $n \to \infty$ where $q = 1 - p$.*

PROOF. In view of Theorem 1.3.5, one can restrict consideration to the case $\widehat{x}_{n,p} = \zeta_{n,k(n,p)}$ where $k(n,p) = [np] + 1$. It is obvious that it is sufficient to prove the $\mathcal{N}(0,1)$ asymptotic normality for random variables

$$\eta_n = f(x_p)\sqrt{n/(pq)}(\widehat{x}_{n,p} - x_p), \qquad n = 1, 2, \dots .$$

Note that $\{\widehat{x}_{n,p} < x\} = \{\nu_n(x) \geq k(n,p)\}$ where $\nu_n(x) = \sum_{i=1}^{n} I_{(-\infty,x)}(\xi_i)$. Then

$$\begin{aligned} \mathsf{P}\{\eta_n < x\} &= \mathsf{P}\left\{\widehat{x}_{n,p} < x_p + \sqrt{\frac{pq}{n}}\frac{x}{f(x_p)}\right\} \\ &= \mathsf{P}\left\{\nu_n\left(x_p + \sqrt{\frac{pq}{n}}\frac{x}{f(x_p)}\right) \geq k(n,p)\right\}. \end{aligned} \tag{1.3.15}$$

Consider random variables

$$\mu_{n,j} = I\left(\xi_j < x_p + x\sqrt{pq/n}/f(x_p)\right), \qquad j = 1, 2, \dots, n.$$

It follows from (1.3.15) that

$$\begin{aligned} \mathsf{P}\{\eta_n < x\} &= \mathsf{P}\left\{\sum_{j=1}^{n} \mu_{n,j} \geq k(n,p)\right\} \\ &= \mathsf{P}\left\{\frac{1}{\sigma_n\sqrt{n}}\sum_{j=1}^{n}(\mu_{n,j} - a_n) \geq \frac{k(n,p) - na_n}{\sigma_n\sqrt{n}}\right\} \end{aligned} \tag{1.3.16}$$

where

$$\begin{aligned} a_n &= \mathsf{E}\mu_{n,j} = F\left(x_p + \sqrt{pq/n}x/f(x_p)\right), \\ \sigma_n^2 &= \mathsf{D}\mu_{n,j} = a_n(1 - a_n). \end{aligned}$$

Since $F(x_p) = p$, the Taylor expansion in a neighborhood of the point x_p shows that

$$F\left(x_p + \sqrt{\frac{pq}{n}}\frac{x}{f(x_p)}\right) = p + \sqrt{\frac{pq}{n}}x + o\left(\frac{1}{\sqrt{n}}\right) \tag{1.3.17}$$

as $n \to \infty$. Thus $a_n \to p$ and $\sigma_n^2 \to pq$ as $n \to \infty$, whence

$$I\left(|\mu_{n,i} - a_n|/\sqrt{n\sigma_n^2} > \tau\right) = 0$$

for all $\tau > 0$ and all sufficiently large n. This means that condition (1.3.14) holds for $\xi_{n,j} = (\mu_{n,j} - a_n)/\sqrt{n\sigma_n^2}$. Therefore the assumptions of Theorem 1.3.7 are satisfied for the random variables $(\mu_{n,j} - a_n)/\sqrt{n\sigma_n^2}$, $j = 1, 2, \ldots, n$.

Applying Theorem 1.3.7 we derive from equality (1.3.16) that

$$\mathsf{P}\{\eta_n < x\} = 1 - \Phi\left(\frac{k(n,p) - na_n}{\sigma_n\sqrt{n}}\right) + o(1) \tag{1.3.18}$$

as $n \to \infty$. Since $k(n,p) = [np] + 1 = np + 1 + r_n$ where $|r_n| < 1$, we obtain from equality (1.3.17) that

$$\frac{k(n,p) - na_n}{\sigma_n\sqrt{n}} = \frac{np + 1 + r_n - np - \sqrt{npq}x + o(\sqrt{n}\,)}{\sqrt{npq}(1 + o(1))} = -x + o(1)$$

as $n \to \infty$.

This together with relation (1.3.18) implies that

$$\mathsf{P}\{\eta_n < x\} = 1 - \Phi(-x) + o(1) = \Phi(x) + o(1)$$

as $n \to \infty$, that is, the sequence of random variables η_n is asymptotically $\mathcal{N}(0,1)$ normal. □

In particular, Theorem 1.3.8 implies that the central order statistic $\zeta_{n,[np]+1}$ is asymptotically normal.

The study of the asymptotic behavior of the central order statistic $\zeta_{n,k(n)}$ is a complicated problem for general $k(n)$. This problem is solved by Smirnov (1949) (see [**32**]) who showed in particular that if $k(n) = np + o(\sqrt{n})$, then the limit distributions for the sequence of random variables $\tilde{\zeta}_{n,k(n)} = (\zeta_{n,k(n)} - A_n)/B_n$ can only be of the following four types:

$$\Phi_\alpha^{(1)}(x) = \begin{cases} \Phi(cx^\alpha), & x > 0, \\ 0, & x \le 0, \end{cases} \qquad \Phi_\alpha^{(2)}(x) = \begin{cases} \Phi(-c|x|^\alpha), & x \le 0, \\ 1, & x > 0, \end{cases}$$

$$\Phi_\alpha^{(3)}(x) = \begin{cases} \Phi(-c_1|x|^\alpha), & x \le 0, \\ \Phi(c_2x^\alpha), & x > 0, \end{cases} \qquad \Phi^{(4)}(x) = \begin{cases} 0, & x \le -1, \\ \frac{1}{2}, & -1 < x \le 1, \\ 1, & x > 1, \end{cases}$$

where A_n and $B_n > 0$ are some appropriate constants depending on n and p; α, c, c_1, and c_2 are some positive constants; and $\Phi(x)$ is the standard Gaussian distribution function. Smirnov also obtained necessary and sufficient conditions on the distribution function $F(x)$ for the convergence to a given type of the limit laws.

REMARK 1.3.3. More details about order statistics and further references can be found in the book by David [**10**].

1.4. The distributions of some functions of Gaussian random vectors

We consider in this section the distributions of some functions of Gaussian random vectors that are widely used in various topics of mathematics.

The normal distribution. Let $X = (X_1, \dots, X_n)'$ be a random column-vector (here and throughout the symbol $'$ stands for the transposition of matrices and vectors). By $a = (a_1, a_2, \dots, a_n)'$ we denote the vector of its expectations of X, that is, $a_i = \mathsf{E}X_i$, $i = 1, 2, \dots, n$, and by $\Lambda = (\lambda_{ij})$ we denote the $n \times n$ matrix of mixed central moments $\lambda_{ij} = \mathsf{E}(X_i - a_i)(X_j - a_j)$, $i, j = 1, 2, \dots, n$. Note that the matrix Λ is symmetric and nonnegative definite. A random vector X is called *normal* (or *Gaussian*) if its characteristic function is of the form

$$\phi(t) = \mathsf{E}e^{iX't} = \exp\left\{ia't - \frac{1}{2}t'\Lambda t\right\} \tag{1.4.1}$$

where $t = (t_1, t_2, \dots, t_n)'$. If X is a normal vector whose characteristic function is given by (1.4.1), then we write $\mathcal{L}(X) = \mathcal{N}(a, \Lambda)$, which means that X has a normal distribution. The distribution $\mathcal{L}(X) = \mathcal{N}(0, I_n)$ where I_n is the $n \times n$ unit matrix is called the *standard normal distribution*. The coordinates of the standard normal vector X are independent random variables whose distribution is $\mathcal{N}(0, 1)$. If the matrix Λ is nonsingular, then the normal distribution is called *proper* (or *nondegenerate*), in which case the distribution possesses the density

$$f(x) = (2\pi)^{-n/2}(\det\Lambda)^{-1/2}\exp\left\{-\frac{1}{2}(x - a)'\Lambda^{-1}(x - a)\right\} \tag{1.4.2}$$

where $x = (x_1, x_2, \dots, x_n)'$ and $\det\Lambda$ is the determinant of the matrix Λ.

Linear transformations of Gaussian vectors are again Gaussian vectors. The precise statement is as follows.

LEMMA 1.4.1. *Let $Y = AX$ where $\mathcal{L}(X) = \mathcal{N}(a, \Lambda)$ and A is a $k \times n$ matrix. Then $\mathcal{L}(Y) = \mathcal{N}(b, B)$ for $b = Aa$ and $B = A\Lambda A'$.*

PROOF. Let $\phi_Y(u)$ and $\phi_X(t)$ be the characteristic functions of vectors Y and X, respectively. Then

$$\phi_Y(u) = \mathsf{E}\exp\{iY'u\} = \mathsf{E}\exp\{iX'A'u\} = \phi_X(A'u).$$

Since $\mathcal{L}(X) = \mathcal{N}(a, \Lambda)$, equality (1.4.1) implies that

$$\phi_Y(u) = \exp\left\{ia'A'u - \frac{1}{2}(A'u)'\Lambda(A'u)\right\} = \exp\left\{i(Aa)'u - \frac{1}{2}u'(A\Lambda A')u\right\}.$$

Thus $\mathcal{L}(Y) = \mathcal{N}(b, B)$ for $b = Aa$ and $B = A\Lambda A'$. □

Equality (1.4.1) implies for a diagonal matrix Λ that the coordinates of the vector X are independent. If the matrix Λ is not diagonal, then there is a linear transformation $Y = AX$ such that the coordinates of the vector Y are independent. Indeed, by Lemma 1.4.1, as a matrix A one can take an orthogonal matrix (this means that $AA' = I_n$) such that $A\Lambda A'$ is diagonal. This implies that if Λ is nonsingular, then there exists a nonsingular matrix A such that the vector $Y = AX$ has the standard normal distribution $\mathcal{N}(0, I_n)$ if $a = 0$.

Chi-square distribution and its properties. Let $X = (X_1, X_2, \dots, X_n)'$ and $\mathcal{L}(X) = \mathcal{N}(0, I_n)$. The distribution of the random variable $\chi_n^2 = \sum_{i=1}^n X_i^2$ is called the *chi-square distribution with n degrees of freedom*. Put $\mathcal{L}(\chi_n^2) = \chi^2(n)$ and let us find the density of $\chi^2(n)$. Applying (1.4.2) as $\Delta r \to 0$ we get

$$\mathsf{P}\{\chi_n^2 \in [r, r+\Delta r)\} = \mathsf{P}\left\{r \le \sum_{i=1}^n X_i^2 < r + \Delta r\right\} = ke^{-r/2}\Delta\left(V_{S(\sqrt{r})}\right) + o(\Delta r)$$

where V_S is the volume of the ball $S(r) = \{x \in \mathbf{R}^n : |x| \le r\}$ of radius r. Since $V_{S(\sqrt{r})} = Cr^{n/2}$, we have $\Delta(V_{S(\sqrt{r})}) = C'r^{n/2-1}\Delta r + o(\Delta r)$. Thus the density of the distribution $\chi^2(n)$ is given by

$$k_n(x) = K_n x^{n/2-1} e^{-x/2}, \qquad x > 0, \tag{1.4.3}$$

where $K_n = (2^{n/2}\Gamma(n/2))^{-1}$ and $\Gamma(\cdot)$ is the Gamma function.

The characteristic function of the distribution $\chi^2(n)$ is

$$\phi(t; n) = \mathsf{E}\exp\{it\chi_n^2\} = K_n \int_0^\infty x^{n/2-1}\exp\{-x(1-2it)/2\}\, dx.$$

Differentiating with respect to t we obtain

$$\phi'(t; n) = \frac{in}{1-2it}\phi(t; n). \tag{1.4.4}$$

Solving equation (1.4.4) subject to the condition $\phi(0; n) = 1$ yields

$$\phi(t; n) = (1-2it)^{-n/2}. \tag{1.4.5}$$

This equality allows one to find the moments of the distribution $\chi^2(n)$:

$$\mathsf{E}\chi_n^2 = \frac{1}{i}\phi'(0; n) = n, \qquad \mathsf{D}\chi_n^2 = \frac{1}{i^2}\phi''(0; n) - (\mathsf{E}\chi_n^2)^2 = 2n. \tag{1.4.6}$$

We also mention the following important property of the distribution $\chi^2(n)$. Let random variables $\chi_{n_1}^2$ and $\chi_{n_2}^2$ be independent and let $\mathcal{L}(\chi_{n_i}^2) = \chi^2(n_i)$, $i = 1, 2$. In view of equality (1.4.5) the characteristic function of the sum $\chi_{n_1}^2 + \chi_{n_2}^2$ is $\phi(t; n_1 + n_2)$, that is, $\mathcal{L}(\chi_{n_1}^2 + \chi_{n_2}^2) = \chi^2(n_1 + n_2)$. This means that the sum of independent chi-square random variables is again a chi-square random variable and its degree of freedom is equal to the sum of degrees of freedom of terms.

Linear and quadratic forms of normal random variables. Let

$$X = (X_1, X_2, \dots, X_n)'$$

be a random vector with the standard normal $\mathcal{L}(0, I_n)$ distribution. Consider a quadratic form

$$Q = \sum_{i,j=1}^n a_{ij} X_i X_j = X'AX$$

where $A = (a_{ij})$ and $A' = A$. We also consider m linear forms

$$Y_k = \sum_{i=1}^n b_{ki} X_i, \qquad k = 1, 2, \dots, m.$$

Using matrix notation we rewrite the latter relations in a compact form as $Y = BX$ where B is a rectangle $m \times n$ matrix and $Y = (Y_1, \dots, Y_m)'$. By O we denote the

matrix with zero entries. The following result contains conditions for the independence of functions Q and Y.

LEMMA 1.4.2. *If $BA = O$, then the functions Q and Y are independent.*

PROOF. Since the real matrix A is symmetric, there exists an orthogonal matrix U such that $U'AU = D$ where D is a diagonal matrix with diagonal entries $\lambda_i \geq 0$, $i = 1, 2, \ldots, n$. The numbers $\lambda_1, \ldots, \lambda_n$ are characteristic numbers of the matrix A, that is, they are the roots of the characteristic equation $\det(A - \lambda I_n) = 0$. The columns u_k of the matrix $U = \|u_1 \ldots u_n\|$ are eigenvectors of the matrix A, that is, $Au_k = \lambda_k u_k$, $k = 1, 2, \ldots, n$.

Let r be the rank of the matrix A and let $\lambda_1, \ldots, \lambda_r$ be nonzero characteristic numbers. The equality $A = UDU'$ can be viewed as the matrix form of the spectral representation of the matrix A, namely

$$A = \sum_{k=1}^{r} \lambda_k u_k u_k'. \tag{1.4.7}$$

By the assumptions of the lemma, $O = BA = \sum_{k=1}^{n} \lambda_k B u_k u_k'$. Multiplying this equality on the right by the vector u_s we get

$$Bu_s = 0, \qquad s = 1, 2, \ldots, r, \tag{1.4.8}$$

since the vectors u_j are orthogonal. Put $Z = (Y_1, \ldots, Y_m, u_1'X, \ldots, u_r'X)$. It is clear that $Z = CX$ for some matrix C and thus Lemma 1.4.1 implies that the distribution of the vector Z is normal with $\mathsf{E}Z = 0$. According to representation (1.4.7)

$$Q = \sum_{k=1}^{n} \lambda_k (X'u_k)(u_k'X) = \sum_{k=1}^{n} \lambda_k (u_k'X)^2.$$

Thus the equalities $\mathsf{E}Y_i u_s'X = 0$, $i = 1, 2, \ldots, m$, $s = 1, 2, \ldots, r$, complete the proof of the lemma, since they mean that Y_i and $u_s'X$ are independent in view of the normal distribution of the vector Z. To prove the above equalities we denote by b_i' the rows of the matrix B, $i = 1, 2, \ldots, m$. Then we have by (1.4.8)

$$\mathsf{E}Y_i u_s'X = \mathsf{E}b_i'X u_s'X = \mathsf{E}b_i'XX'u_s = b_i'(\mathsf{E}XX')u_s = b_i'I_n u_s = 0. \qquad \square$$

Consider two quadratic forms $Q_1 = X'AX$ and $Q_2 = X'BX$.

LEMMA 1.4.3. *If $AB = BA = O$, then Q_1 and Q_2 are independent.*

PROOF. Let the matrix A admit the representation (1.4.7), and let the spectral representation of the matrix B be $B = \sum_{i=1}^{s} \nu_i v_i v_i'$ where $s = \operatorname{rank} B$ is the rank of the matrix B. By the assumptions, $O = AB = \sum_{k,i} \lambda_k \nu_i u_k (u_k'v_i) v_i'$. Multiplying this equality on the left by u_i' and on the right by v_j, we get $u_i'v_j = 0$, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Since the joint distribution of random variables $u_i'X$ and $v_j'X$ is normal, we prove as above that these random variables are independent. Now it follows from $Q_1 = \sum_{k=1}^{r} \lambda_k (u_k'X)^2$ and $Q_2 = \sum_{i=1}^{s} \nu_i (v_i'X)^2$ that the random variables Q_1 and Q_2 are independent. $\square$

The distributions of quadratic forms of normal random variables. By $\operatorname{tr} A$ we denote the *trace* of a quadratic matrix A, that is, the sum of its diagonal entries.

LEMMA 1.4.4. *Let* $Q = X'AX$ *where* $\mathcal{L}(X) = \mathcal{N}(0, I_n)$ *and* $\operatorname{rank} A = r \leq n$. *If the matrix* A *is idempotent, that is,* $A^2 = A$, *then* $\mathcal{L}(Q) = \chi^2(r)$ *and* $r = \operatorname{tr} A$.

PROOF. Let the matrix A admit the representation (1.4.7). Since A is symmetric and idempotent, $\lambda_1 = \cdots = \lambda_r = 1$. Thus $Q = \sum_{k=1}^{r} (u_k' X)^2$. The vectors u_k are orthonormal, thus the random variables $u_k' X$, $k = 1, 2, \ldots, r$, are independent and $\mathcal{N}(0,1)$ normal. Hence $\mathcal{L}(Q) = \chi^2(r)$. Since $A = UDU'$, we obtain from $\operatorname{tr}(BA) = \operatorname{tr}(AB)$ that $\operatorname{tr} A = \operatorname{tr}(U'UD) = \operatorname{tr} D = \lambda_1 + \cdots + \lambda_r = r$. □

The following result, which is a corollary of the preceding assertions, is of its own interest as well.

THEOREM 1.4.1. *Let an* n*-dimensional vector* Y *have a nondegenerate* $\mathcal{N}(\mu, \Sigma)$ *distribution. Then the distribution of the quadratic form* $Q = (Y - \mu)'\Sigma^{-1}(Y - \mu)$ *is* $\chi^2(n)$.

PROOF. Let U be an orthogonal matrix such that $U'\Sigma U = D$ where D is a diagonal matrix. Since Σ is nondegenerate, all diagonal entries λ_k of the matrix D are positive. Thus $D^{-1/2}$ is well defined as the diagonal matrix with diagonal entries $\lambda_k^{-1/2}$. Consider the random vector $Z = D^{-1/2}U'(Y - \mu)$. By Lemma 1.4.1

$$\mathcal{L}(Z) = \mathcal{N}(0, I_n).$$

On the other hand, $Y - \mu = UD^{1/2}Z$. Thus $Q = Z'D^{1/2}U'\Sigma^{-1}UD^{1/2}Z = Z'Z$, whence $\mathcal{L}(Q) = \chi^2(n)$. □

The following important result of the sampling theory is proved by Fisher (1925).

THEOREM 1.4.2. *Let* $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ *be a sample from the* $\mathcal{N}(\mu, \sigma^2)$ *distribution. Then the sampling moments* a_1 *and* m_2 *are independent. Moreover* $\mathcal{L}(\sqrt{n}(a_1 - \mu)/\sigma) = \mathcal{N}(0,1)$ *and* $\mathcal{L}(nm_2/\sigma^2) = \chi^2(n-1)$.

PROOF. Consider a sample $\tilde{\xi}^{(n)} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ where

$$\tilde{\xi}_i = (\xi_i - \mu)/\sigma, \qquad i = 1, 2, \ldots, n.$$

Put

$$\tilde{a}_1 = \frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}_i, \qquad \tilde{m}_2 = \frac{1}{n}\sum_{i=1}^{n} (\tilde{\xi}_i - \tilde{a}_1)^2.$$

Then $\tilde{a}_1 = (a_1 - \mu)/\sigma$ and $\tilde{m}_2 = m_2/\sigma$. Thus it is sufficient to prove that $\tilde{a}_1$ and $\tilde{m}_2$ are independent, since $\mathcal{L}(\sqrt{n}\tilde{a}_1) = \mathcal{N}(0,1)$ and $\mathcal{L}(n\tilde{m}_2) = \chi^2(n-1)$. Consider an n-dimensional vector-column $b = (1/n, \ldots, 1/n)'$ and $n \times n$ matrix $B = \|b \cdots b\|$. It is clear that $\tilde{a}_1 = b'\tilde{\xi}^{(n)}$ and $n\tilde{m}_2 = (\tilde{\xi}^{(n)} - B\tilde{\xi}^{(n)})'(\tilde{\xi}^{(n)} - B\tilde{\xi}^{(n)}) = (\tilde{\xi}^{(n)})'A\tilde{\xi}^{(n)}$ where $A = I_n - B$. Since $b'A = b' - b'B = b' - b' = 0$, the random variables $\tilde{a}_1$ and $\tilde{m}_2$ are independent by Lemma 1.4.2.

It is obvious that the distribution of a_1 is normal. Note that the matrix A is idempotent and $\operatorname{tr} A = \operatorname{tr} I_n - \operatorname{tr} B = n - 1$. Then $\mathcal{L}(n\tilde{m}_2) = \chi^2(n-1)$ by Theorem 1.4.1. □

Student and Snedekor distributions. Let two random variables ξ and χ^2 be independent and let $\mathcal{L}(\xi) = \mathcal{N}(0,1)$ and $\mathcal{L}(\chi^2) = \chi^2(n)$. Then the distribution of the random variable $t = \xi/\sqrt{\chi^2/n}$ is called the *Student distribution with n degrees of freedom* and is denoted by $S(n)$. The density $s_n(x)$ of the distribution $S(n)$ is

$$s_n(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \frac{1}{(1+x^2/n)^{(n+1)/2}}, \qquad x \in \mathbf{R}.$$

Let random variables χ_1^2 and χ_2^2 be independent and let $\mathcal{L}(\chi_i^2) = \chi^2(n_i)$, $i = 1, 2$. Then the distribution of the ratio

$$F = (\chi_1^2/n_1) \,/\, (\chi_2^2/n_2)$$

is called the *Snedekor distribution with n_1 and n_2 degrees of freedom* and is denoted by $S(n_1, n_2)$. This distribution is sometimes called the *F-distribution* or *Fisher distribution*. The density $s_{n_1,n_2}(x)$ of the distribution $S(n_1, n_2)$ is

$$s_{n_1,n_2}(x) = \left(\frac{n_1}{n_2}\right)^{n_1/2} \Gamma\left(\frac{n_1+n_2}{2}\right) \left(\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\right)^{-1} \times x^{n_1/2-1} \left(1 + \frac{n_1 x}{n_2}\right)^{-(n_1+n_2)/2}, \qquad x > 0.$$

The distributions $S(n)$ and $S(n_1, n_2)$ play an important role in the sampling theory.

REMARK 1.4.1. Properties of normal distributions and those related to normal distributions are treated in many textbooks on probability theory. A comprehensive text on properties and applications of normal distributions in statistical problems can be found in [**1, 28**].

Samples from Multidimensional Distributions

2.1. Empirical distribution function, sampling moments, and their properties

Empirical distribution function and its properties. Let (ξ, η) be a two-dimensional random vector with real coordinates ξ and η. Denote its distribution function by $F(x,y) = \mathsf{P}\{\xi < x, \eta < y\}$, $x \in \mathbf{R}$, $y \in \mathbf{R}$. Assume that there are n independent observations of the vector (ξ, η):

$$(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots, (\xi_n, \eta_n). \tag{2.1.1}$$

The set of observations is called a *sample from the two-dimensional distribution* $F(x,y)$.

For fixed $x \in \mathbf{R}$ and $y \in \mathbf{R}$ consider the following random variables:

$$\nu_n(x,y) = \sum_{i=1}^{n} I(\{\xi_i < x, \eta_i < y\}). \tag{2.1.2}$$

Then

$$F_n(x,y) = \frac{1}{n}\nu_n(x,y), \qquad x \in \mathbf{R},\ y \in \mathbf{R}, \tag{2.1.3}$$

is called the *empirical distribution function of the sample* (2.1.1). Note that the empirical distribution function $F_n(x,y)$ possesses all the properties of regular two-dimensional distribution functions.

Equality (2.1.2) implies that $\nu_n(x,y)$ is the total number of occurrences of the event

$$\{\xi < x, \eta < y\}$$

in n independent trials, while (2.1.3) shows that the empirical distribution function $F_n(x,y)$ is the relative frequency of the event $\{\xi < x, \eta < y\}$ in n independent trials. Like the one-dimensional case, the Bernoulli law of large numbers implies that the empirical distribution function $F_n(x,y)$ approaches $F(x,y)$ in probability as $n \to \infty$ for all $x \in \mathbf{R}$ and $y \in \mathbf{R}$, that is,

$$\lim_{n\to\infty} \mathsf{P}\{|F_n(x,y) - F(x,y)| > \varepsilon\} = 0 \qquad \text{for all } \varepsilon > 0.$$

Moreover, the Borel strong law of large numbers implies that $F_n(x,y)$ approaches $F(x,y)$ with probability 1 as $n \to \infty$ for all $x \in \mathbf{R}$ and $y \in \mathbf{R}$, that is,

$$\mathsf{P}\left\{\lim_{n\to\infty} F_n(x,y) = F(x,y)\right\} = 1.$$

Therefore the empirical distribution function $F_n(x,y)$ may serve as an approximation of the distribution function $F(x,y)$.

According to definition (2.1.3) the empirical distribution function $F_n(x,y)$ is a random variable for all fixed x and y. It assumes values k/n, $k=0,1,2,\dots,n$, and moreover

$$\mathsf{P}\left\{F_n(x,y)=\frac{k}{n}\right\}=\binom{n}{k}F^k(x,y)(1-F(x,y))^{n-k}.$$

Therefore

$$\mathsf{E}F_n(x,y)=F(x,y),\qquad \mathsf{D}F_n(x,y)=F(x,y)(1-F(x,y))/n.$$

Applying the De Moivre–Laplace central limit theorem we obtain the following result on the asymptotic normality of the empirical distribution function $F_n(x,y)$.

THEOREM 2.1.1. *The sequence of empirical distribution functions*

$$F_n(x,y),\qquad n=1,2,\dots,$$

is asymptotically normal with parameters $\big(F(x,y),F(x,y)(1-F(x,y))/n\big)$ *for all fixed* $x\in\mathbf{R}$ *and* $y\in\mathbf{R}$.

Moments of two-dimensional distributions. Let (ξ,η) be a real two-dimensional random vector. The number $\alpha_{ij}=\mathsf{E}\xi^i\eta^j$ is called the *mixed moment of order* $i+j$ (or, $(i+j)$*-th mixed moment*) of the random vector (ξ,η). The number $\mu_{ij}=\mathsf{E}(\xi-\alpha_{10})^i(\eta-\alpha_{01})^j$ is called the *mixed central moment of order* $i+j$ (or $(i+j)$*-th mixed central moment*) of the random vector (ξ,η). It is easy to see that

$$\mu_{ij}=\sum_{k=0}^{i}\sum_{l=0}^{j}\binom{i}{k}\binom{j}{l}(-1)^{k+l}\alpha_{10}^k\alpha_{01}^l\alpha_{i-k,j-l}. \tag{2.1.4}$$

Note that α_{i0} is the i-th moment of the random variable ξ, while α_{0j} is the j-th moment of the random variable η. Analogously, μ_{i0} is the i-th central moment of the random variable ξ, while μ_{0j} is the j-th central moment of the random variable η. Note further that μ_{20} is the variance of ξ, while μ_{02} is the variance of η. We often use the notation $\sigma_1^2=\mu_{20}$ and $\sigma_2^2=\mu_{02}$. It is clear that $\mu_{20}=\alpha_{20}-\alpha_{10}^2$, $\mu_{02}=\alpha_{02}-\alpha_{01}^2$, and $\mu_{11}=\alpha_{11}-\alpha_{10}\alpha_{01}$.

If $\mu_{11}=0$, then the random variables ξ and η are called *uncorrelated*. In this case $\alpha_{11}=\alpha_{10}\alpha_{01}$, that is, $\mathsf{E}\xi\eta=\mathsf{E}\xi\mathsf{E}\eta$. If ξ and η are independent, then $\mu_{11}=0$, that is, independent random variables are uncorrelated. The converse is, in general, false. In a particular case where the vector (ξ,η) has a normal distribution, the random variables ξ and η are independent if and only if they are uncorrelated.

Let $z=(t,u)'$ where t and u are real numbers. Consider a quadratic form

$$Q(z)=\mathsf{E}[t(\xi-\alpha_{10})+u(\eta-\alpha_{01})]^2=\mu_{20}t^2+2\mu_{11}tu+\mu_{02}u^2. \tag{2.1.5}$$

Since $Q(z)$ is the expectation of a square of a random variable, $Q(z)\ge 0$ for all vectors z, whence it follows that the quadratic form $Q(z)$ is nonnegative definite. Definition (2.1.5) implies that $Q(z)=z'Mz$ where M is the matrix of central moments of second order:

$$M=\begin{pmatrix}\mu_{20} & \mu_{11}\\ \mu_{11} & \mu_{02}\end{pmatrix}.$$

The matrix M also is nonnegative definite (this follows from the same property of the quadratic form $Q(z)$). Hence

$$\mu_{02}\mu_{20}-\mu_{11}^2\ge 0. \tag{2.1.6}$$

Rewriting (2.1.6) we obtain $\mu_{11}^2 \leq \mu_{02}\mu_{20}$, which is known as *the Cauchy–Bunyakovskiĭ inequality*.

Denote the rank of M by r. The possible values of r are 0, 1, or 2. If $r = 2$, then (2.1.6) becomes a strict inequality, while (2.1.6) becomes an equality if $r = 0$ or $r = 1$. The following result contains some simple properties of the distribution of the vector (ξ, η) related to the rank r.

THEOREM 2.1.2. *The following are true*:

1) $r = 0$ *if and only if the distribution of the vector* (ξ, η) *assigns unit mass to a single point*;
2) $r = 1$ *if and only if the distribution of the vector* (ξ, η) *is concentrated on a straight line, but not at a single point*;
3) $r = 2$ *if and only if the support of the distribution of the vector* (ξ, η) *does not coincide with a straight line or with a single point.*

PROOF. We consider only the first two cases when $r = 0$ and $r = 1$. The case when $r = 2$ follows from 1) and 2).

If $r = 0$, then $\mu_{20} = \mu_{02} = 0$, so that the distribution of the random variables ξ and η is concentrated at the points α_{10} and α_{01}, respectively. Then the distribution of the vector (ξ, η) is concentrated at the single point $(\alpha_{10}, \alpha_{01})$. Conversely, if the distribution of the vector (ξ, η) is concentrated at a single point, then $\mu_{20} = \mu_{02} = 0$, so that $\mu_{11} = 0$ by (2.1.6). Therefore the rank of the matrix M is equal to zero.

If $r = 1$, then the quadratic form $Q(z)$ is not positive definite. Thus there is a vector $z_0 = (t_0, u_0) \neq 0$ such that $Q(z_0) = 0$. It means by (2.1.5) that with probability 1

$$t_0(\xi - \alpha_{10}) + u_0(\xi - \alpha_{01}) = 0, \tag{2.1.7}$$

which implies that the distribution of the vector (ξ, η) is concentrated at a straight line $t_0(x - \alpha_{10}) + u_0(y - \alpha_{01}) = 0$. Conversely, let the distribution of the vector (ξ, η) be concentrated at a straight line but not be concentrated at a single point. Obviously this line passes the point $(\alpha_{10}, \alpha_{01})$, whence it follows that equality (2.1.7) holds with probability 1 for some constants t_0 and u_0. By (2.1.5) we have $Q(z_0) = 0$ for $z_0 = (t_0, u_0)'$, that is, the quadratic form $Q(z)$ is not positive definite. Since the distribution is not concentrated at a single point of the line (2.1.7), at least one of the numbers μ_{20} or μ_{02} is nonzero. Thus the rank of the matrix M equals 1. □

If $\mu_{20} \neq 0$ and $\mu_{02} \neq 0$, then the rank of the matrix M equals either 1 or 2. Let

$$\rho = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \frac{\mu_{11}}{\sigma_1\sigma_2}. \tag{2.1.8}$$

We have by (2.1.6) that $\rho^2 \leq 1$, that is, $|\rho| \leq 1$. It is clear that $|\rho| = 1$ if and only if the rank of the matrix M is equal to 1, that is, if and only if the support of the distribution of the vector (ξ, η) belongs to a straight line. In particular, if ξ and η are independent, then $\mu_{11} = 0$, whence $\rho = 0$. On the other hand, the equality $\rho = 0$ means that the random variables ξ and η are uncorrelated.

The number ρ defined by (2.1.8) is called the *coefficient of correlation* (or simply *correlation*) of random variables ξ and η.

Sampling moments. Consider a sample (2.1.1) consisting of n independent observations of a vector (ξ, η) and denote its distribution function by $F(x, y)$. Let $F_n(x, y)$ be the empirical distribution function of the sample (2.1.1) defined by (2.1.3). The $(i + j)$-th mixed moment of the empirical distribution function $F_n(x, y)$ is called the *$(i + j)$-th sampling mixed moment*, that is,

$$a_{ij} = \int\int x^i y^j \, dF_n(x, y) = \frac{1}{n} \sum_{k=1}^{n} \xi_k^i \eta_k^j. \tag{2.1.9}$$

The $(i+j)$-th mixed central moment of the empirical distribution function $F_n(x, y)$ is called the *$(i + j)$-th sampling mixed central moment*, that is,

$$\begin{aligned} m_{ij} &= \int\int (x - a_{10})^i (y - a_{01})^j \, dF_n(x, y) \\ &= \frac{1}{n} \sum_{k=1}^{n} (\xi_k - a_{10})^i (\eta_k - a_{01})^j. \end{aligned} \tag{2.1.10}$$

It follows from (2.1.9) and (2.1.10) that an analog of (2.1.4) holds for sampling mixed moments, namely

$$m_{ij} = \sum_{k=0}^{i} \sum_{l=0}^{j} \binom{i}{k} \binom{j}{l} (-1)^{k+l} a_{10}^k a_{01}^l a_{i-k,j-l}. \tag{2.1.11}$$

It follows from (2.1.9) that

$$\mathsf{E} a_{ij} = \frac{1}{n} \sum_{k=1}^{n} \mathsf{E} \xi_k^i \eta_k^j = \alpha_{ij}, \tag{2.1.12}$$

$$\mathsf{D} a_{ij} = \frac{1}{n^2} \sum_{k=1}^{n} \mathsf{D} \xi_k^i \eta_k^j = \frac{1}{n} (\alpha_{2i,2j} - \alpha_{ij}^2). \tag{2.1.13}$$

As in the proof of Theorem 1.2.1, we derive from (2.1.11) that

$$\mathsf{E} m_{ij} = \mu_{ij} + O\left(\frac{1}{n}\right), \tag{2.1.14}$$

$$\mathsf{D} m_{ij} = \frac{1}{n} R + O\left(\frac{1}{n^2}\right) \tag{2.1.15}$$

where R is a constant depending on mixed central moments μ_{kl}. Relations (2.1.12)–(2.1.15) hold if all expectations involved are finite.

The most important mixed moments are of order less than or equal to 2. We often use the notation $m_{20} = s_1^2$ and $m_{02} = s_2^2$. For sampling mixed central moments m_{ij} of order $i + j = 2$ we easily obtain that

$$\mathsf{E} m_{ij} = \frac{n-1}{n} \mu_{ij}, \qquad \mathsf{D} m_{ij} = \frac{\mu_{2i,2j} - \mu_{ij}^2}{n} + O\left(\frac{1}{n^2}\right),$$

which is a refinement of (2.1.14) and (2.1.15).

The number

$$r = \frac{m_{11}}{\sqrt{m_{20} m_{02}}} = \frac{m_{11}}{s_1 s_2}$$

is called the *sampling coefficient of correlation* (or simply *sampling correlation*). Since r is the coefficient of correlation of the distribution function $F_n(x,y)$, we have $|r| \le 1$. The sampling coefficient of correlation r attains values ± 1 if and only if all the sampling points $(\xi_1,\eta_1),(\xi_2,\eta_2),\dots,(\xi_n,\eta_n)$ lie on a straight line. One can show that

$$\mathsf{E}r = \rho + O\left(\frac{1}{n}\right), \tag{2.1.16}$$

$$\mathsf{D}r = \frac{\rho^2}{4n}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right) + O\left(n^{-3/2}\right). \tag{2.1.17}$$

By the law of large numbers, the moments a_{ij} approach α_{ij} in probability as $n \to \infty$. The latter result also follows from (2.1.12) and (2.1.13) by the Chebyshev inequality if the moment $\alpha_{2i,2j}$ exists. Now we apply Theorem 1.2.2 to prove that $m_{ij} \to \mu_{ij}$ and $r \to \rho$ in probability as $n \to \infty$.

Provided $\alpha_{2i,2j} < \infty$ we use relations (2.1.12) and (2.1.13) and the central limit theorem for sums of independent identically distributed random variables $\xi_k^i \eta_k^j$, $k = 1,2,\dots,n$, and prove that sampling mixed moments a_{ij} are asymptotically $\mathcal{N}(\alpha_{ij},(\alpha_{2i,2j} - \alpha_{ij}^2)/n)$ normal. Applying (2.1.4) and (2.1.5) one can show that the sampling mixed central moments m_{ij} are asymptotically $\mathcal{N}(\mu_{ij}, R/n)$ normal under appropriate assumptions where R is a function of central moments involved with asymptotic equality (2.1.15). One can also see from (2.1.16) and (2.1.17) that the sampling correlation coefficient r is asymptotically $\mathcal{N}(\rho, C^2(\rho)/n)$ normal where

$$C^2(\rho) = \frac{\rho^2}{4}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right).$$

REMARK 2.1.1. More details about two-dimensional sampling vectors can be found in the classical book by Cramér [9].

Samples from k-dimensional distributions for $k > 2$. Let $(\xi_1,\xi_2,\dots,\xi_k)$ be a k-dimensional vector with real coordinates $\xi_1,\xi_2,\dots,\xi_k$ and the distribution function

$$F(x_1,x_2,\dots,x_k) = \mathsf{P}\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_k < x_k\}.$$

The *moments* of this distribution are defined by

$$\alpha_{i_1 i_2 \dots i_k} = \mathsf{E}\xi_1^{i_1}\xi_2^{i_2}\dots\xi_k^{i_k}.$$

The number $i_1 + i_2 + \dots + i_k$ is called the *order* of the moment. We use the notation $\alpha_1^{(i)} = \mathsf{E}\xi_i$, $i = 1,2,\dots,k$, for moments of first order. In particular, $(\alpha_1^{(1)},\dots,\alpha_1^{(k)}) = (\mathsf{E}\xi_1,\dots,\mathsf{E}\xi_k)$. The *central moments* are defined by

$$\mu_{i_1 i_2 \dots i_k} = \mathsf{E}\left(\xi_1 - \alpha_1^{(1)}\right)^{i_1}\left(\xi_2 - \alpha_1^{(2)}\right)^{i_2}\dots\left(\xi_k - \alpha_1^{(k)}\right)^{i_k}$$

where $i_1 + i_2 + \dots + i_k$ is the *order* of the moment. The general notation is inconvenient for use if $k > 2$. Thus we sometimes use another notation for moments of second order, namely

$$\lambda_{ii} = \sigma_i^2 = \mathsf{E}\left(\xi_i - \alpha_1^{(i)}\right)^2, \qquad \lambda_{ij} = \mathsf{E}\left(\xi_i - \alpha_1^{(i)}\right)\left(\xi_j - \alpha_1^{(j)}\right) = \rho_{ij}\sigma_i\sigma_j.$$

Here σ_i^2 is the variance of the random variable ξ_i and λ_{ij} is the mixed central moment of second order of random variables ξ_i and ξ_j. The coefficient of correlation ρ_{ij} for random variables ξ_i and ξ_j is well defined if $\sigma_i \neq 0$ and $\sigma_j \neq 0$. Moreover $\rho_{ij} = \lambda_{ij}/(\sigma_i\sigma_j)$. The *matrix of central moments of second order* $\Lambda = (\lambda_{ij})$ is nonnegative definite. The matrix of coefficients of correlation

$$\boldsymbol{\rho} = (\rho_{ij})$$

(well defined if all σ_i are positive) also is nonnegative definite. There is a relationship between matrices Λ and $\boldsymbol{\rho}$, namely

$$\Lambda = \Sigma\boldsymbol{\rho}\Sigma$$

where Σ is a diagonal matrix with diagonal entries $\sigma_1, \ldots, \sigma_n$.

In particular, if $\lambda_{ij} = 0$ for all $i \neq j$, then the random variables $\xi_1, \xi_2, \ldots, \xi_k$ are uncorrelated and the matrix Λ is diagonal, thus $\det(\Lambda) = \lambda_{11}\lambda_{22}\ldots\lambda_{kk}$. If additionally all numbers σ_i are positive, then the matrix $\boldsymbol{\rho}$ is well defined and moreover $\boldsymbol{\rho} = I_k$.

Consider n independent observations of a random vector $(\xi_1, \ldots, \xi_k)$. This means that there is a sample $(\xi_{1i}, \ldots, \xi_{ki})$, $i = 1, \ldots, n$. Denote by $F_n(x_1, \ldots, x_k)$ the empirical distribution function of this sample defined in the same way as in the case of two-dimensional vectors. The *sampling moments* are defined in this case by

$$a_{i_1 i_2 \ldots i_k} = \frac{1}{n}\sum_{j=1}^{n} \xi_{1j}^{i_1}\xi_{2j}^{i_2}\cdots\xi_{kj}^{i_k}.$$

The number $i_1 + \cdots + i_k$ is called the *order* of the moment. We use the notation $a_1^{(i)} = n^{-1}\sum_{j=1}^{n}\xi_{ij}$, $i = 1, 2, \ldots, k$, for sampling moments of first order. The *central sampling moments* are defined by

$$m_{i_1 i_2 \ldots i_k} = \frac{1}{n}\sum_{j=1}^{n}\left(\xi_{1j} - a_1^{(1)}\right)^{i_1}\left(\xi_{2j} - a_1^{(2)}\right)^{i_2}\cdots\left(\xi_{kj} - a_1^{(k)}\right)^{i_k}.$$

The number $i_1 + \cdots + i_k$ is called the *order* of the moment. For moments of second order we use the simpler notation

$$l_{ii} = s_i^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\xi_{ij} - a_1^{(i)}\right)^2, \qquad i = 1, 2, \ldots, k,$$

$$l_{ij} = \frac{1}{n}\sum_{m=1}^{n}\left(\xi_{im} - a_1^{(i)}\right)\left(\xi_{jm} - a_1^{(j)}\right) = r_{ij}s_i s_j.$$

Here s_i^2 is the sampling variance of the random variable ξ_i constructed from observations of the i-th coordinate of the vector, while $r_{ij} = l_{ij}/(s_i s_j)$ is the sampling coefficient of correlation between random variables ξ_i and ξ_j. Let $L = (l_{ij})$ be the matrix of sampling central moments of second order, and let $R = (r_{ij})$ be the matrix of sampling coefficients of correlation. It is clear that $L = SRS$ for the diagonal matrix S whose diagonal entries are $s_1, s_2, \ldots, s_k$.

Asymptotic behavior of the empirical distribution function, sampling moments, and sampling coefficients of correlation for $k > 2$ are analogous to those in the cases of $k = 2$ and $k = 1$. Further results and other properties can be found in a classical

book by Cramér [9] as well as in other books devoted to the multidimensional case, for example in [1, 4, 18, 28].

2.2. Sampling regression and its properties

General regression. Let ξ and η be two random variables with the joint distribution function $F(x, y)$. We denote the conditional expectations by

$$m_1(y) = \mathsf{E}\{\xi/\eta = y\}, \qquad m_2(x) = \mathsf{E}\{\eta/\xi = x\}. \tag{2.2.1}$$

The function $m_1(y)$ is called the *regression of* ξ *on* η, while the function $m_2(x)$ is called the *regression of* η *on* ξ. Regressions (2.2.1) possess an important property of minimality explained in the following result.

THEOREM 2.2.1. *If* $\mathsf{E}\xi^2 < \infty$, *then for any Borel function* f

$$\mathsf{E}(\xi - m_1(\eta))^2 \le \mathsf{E}(\xi - f(\eta))^2. \tag{2.2.2}$$

Analogously if $\mathsf{E}\eta^2 < \infty$, *then for any Borel function* g

$$\mathsf{E}(\eta - m_2(\xi))^2 \le \mathsf{E}(\eta - g(\xi))^2. \tag{2.2.3}$$

PROOF. We prove inequality (2.2.3), the proof of inequality (2.2.2) is analogous.

Let g be an arbitrary Borel function. Inequality (2.2.3) is trivial if

$$\mathsf{E}(\eta - g(\xi))^2 = \infty.$$

Consider the case $\mathsf{E}(\eta - g(\xi))^2 < \infty$. Then

$$\begin{aligned} \mathsf{E}(\eta - g(\xi))^2 &= \mathsf{E}((\eta - m_2(\xi)) + (m_2(\xi) - g(\xi)))^2 \\ &= \mathsf{E}(\eta - m_2(\xi))^2 + 2\mathsf{E}(\eta - m_2(\xi))(m_2(\xi) - g(\xi)) \\ &\quad + \mathsf{E}(m_2(\xi) - g(\xi))^2. \end{aligned} \tag{2.2.4}$$

On the other hand,

$$\begin{aligned} \mathsf{E}(\eta - m_2(\xi))(m_2(\xi) - g(\xi)) &= \mathsf{E}\mathsf{E}\left\{(\eta - m_2(\xi))(m_2(\xi) - g(\xi)) \,/\, \xi\right\} \\ &= \mathsf{E}(m_2(\xi) - g(\xi))\mathsf{E}\left\{\eta - m_2(\xi) \,/\, \xi\right\} = 0, \end{aligned} \tag{2.2.5}$$

since

$$\mathsf{E}\left\{\eta - m_2(\xi) \,/\, \xi\right\} = \mathsf{E}\left\{\eta \,/\, \xi\right\} - m_2(\xi) = m_2(\xi) - m_2(\xi) = 0.$$

It follows from (2.2.4) and (2.2.5) that

$$\mathsf{E}(\eta - g(\xi))^2 = \mathsf{E}(\eta - m_2(\xi))^2 + \mathsf{E}(m_2(\xi) - g(\xi))^2 \ge \mathsf{E}(\eta - m_2(\xi))^2. \tag{2.2.6}$$

□

REMARK 2.2.1. Inequality (2.2.6) becomes an equality if and only if

$$\mathsf{E}(m_2(\xi) - g(\xi))^2 = 0,$$

that is, if $\mathsf{P}\{g(\xi) = m_2(\xi)\} = 1$. Thus inequality (2.2.3) becomes an equality if and only if $\mathsf{P}\{g(\xi) = m_2(\xi)\} = 1$. Similarly inequality (2.2.2) becomes an equality if and only if $\mathsf{P}\{f(\eta) = m_1(\eta)\} = 1$. This implies that the regressions $m_1(y)$ and $m_2(x)$ can be defined as functions minimizing the right-hand sides of inequalities (2.2.2) and (2.2.3), respectively. More precisely, every Borel function f^*, such that

$$\mathsf{E}(\xi - f^*(\eta))^2 = \min \mathsf{E}(\xi - f(\eta))^2 \tag{2.2.7}$$

where the minimum is taken over all Borel functions f, is the regression of ξ on η. Similarly, every Borel function g^*, such that

$$\mathsf{E}(\eta - g^*(\xi))^2 = \min \mathsf{E}(\eta - g(\xi))^2 \tag{2.2.8}$$

where the minimum is taken over all Borel functions g, is the regression of η on ξ.

Linear regression. We solved problem (2.2.7) in the class of all Borel functions and found a function $f(y)$ such that the random variable $f(\eta)$ as a function of η is the best *mean square approximation* of the random variable ξ. In other words, we found a function $f^*(y)$ minimizing the mean square error $\mathsf{E}(\xi - f(\eta))^2$. Problem (2.2.7) is also of interest in the cases where we consider a narrower class of functions $f(y)$ instead of the class of all Borel functions. Say, one can solve problem (2.2.7) in the class of all linear functions or, more generally, in the class of polynomials of a fixed degree, etc. A similar remark concerns problem (2.2.8), too.

Let $L = \{\alpha + \beta x;\ \alpha, \beta \in (-\infty, \infty)\}$ be the class of all linear functions on $\mathbf{R}$. A function $g^*(x) = \alpha^* + \beta^* x$ such that

$$\mathsf{E}(\eta - g^*(\xi))^2 = \min_{g \in L} \mathsf{E}(\eta - g(\xi))^2 \tag{2.2.9}$$

is called the *linear regression of η on ξ*.

A function $f^*(y) = \alpha^* + \beta^* y$ such that

$$\mathsf{E}(\xi - f^*(\eta))^2 = \min_{f \in L} \mathsf{E}(\xi - f(\eta))^2 \tag{2.2.10}$$

is called the *linear regression of ξ on η*.

Below we find the linear regression of η on ξ, that is, we find a function

$$g^*(x) = \alpha^* + \beta^* x$$

solving problem (2.2.9). We assume that $\mu_{20} > 0$ and $\mu_{02} > 0$. Therefore we exclude the case of $\mu_{20} = 0$ and $\mu_{02} = 0$ for which the distribution of the vector (ξ, η) is concentrated at the point $(\alpha_{10}, \alpha_{01})$.

Let $G(\alpha, \beta) = \mathsf{E}(\eta - g(\xi))^2$ where $g(x) = \alpha + \beta x$ is an arbitrary linear function. Then

$$\begin{aligned} G(\alpha, \beta) &= \mathsf{E}((\eta - \alpha_{01}) - \beta(\xi - \alpha_{10}) + (\alpha_{01} - \beta\alpha_{10} - \alpha))^2 \\ &= \mu_{20}\beta^2 - 2\mu_{11}\beta + \mu_{02} + (\alpha_{01} - \beta\alpha_{10} - \alpha)^2. \end{aligned} \tag{2.2.11}$$

To solve the regression problem, it is sufficient to find the minimum of the function $G(\alpha, \beta)$. It follows from (2.2.11) that

$$\beta^* = \beta_{21} = \frac{\mu_{11}}{\mu_{20}} = \rho\frac{\sigma_2}{\sigma_1}, \qquad \alpha^* = \alpha_{01} - \beta^*\alpha_{10} \tag{2.2.12}$$

where ρ is the coefficient of correlation between random variables ξ and η defined by (2.1.8). The number β_{21} defined by the first equality in (2.2.12) is called the *coefficient of linear regression of η on ξ*.

Substituting coefficients (2.1.12), the regression equation $y = g^*(x) = \alpha^* + \beta^* x$ becomes of the form

$$y = \alpha_{01} + \beta_{21}(x - \alpha_{10}). \tag{2.2.13}$$

This is the equation determining a straight line passing through the point $(\alpha_{10}, \alpha_{01})$. This equation can also be written in the form

$$\frac{y-\alpha_{01}}{\sigma_2} = \rho\frac{x-\alpha_{10}}{\sigma_1}. \tag{2.2.14}$$

Equation (2.2.14) is called the *canonical equation of the linear regression of* η *on* ξ.

The number $\mathsf{E}(\eta-\alpha^*-\beta^*\xi)^2$ is the minimal mean square error in the problem (2.2.9) and is called the *least variance* of the random variable η. In view of (2.2.11) and (2.2.12) we get

$$\mathsf{E}(\eta-\alpha^*-\beta^*\xi)^2 = \sigma_2^2(1-\rho^2). \tag{2.2.15}$$

It follows from (2.2.15) that $|\rho| = 1$ if and only if $\eta = \alpha^* + \beta^*\xi$ with probability 1 where α^* and β^* are defined by (2.2.12), that is, $\eta = \alpha_{01} + \beta_{21}(\xi - \alpha_{10})$ with probability 1. Therefore we obtained the straight line for the assertion 2) of Theorem 2.1.2, that is, we found the coefficients t_0 and u_0 in equality (2.1.7).

If $\rho = 0$, then it follows from (2.2.14) that the linear regression of η on ξ is of the form $y = \alpha_{01}$. Note that this is the straight line parallel to the x-axis and passing through the point $(\alpha_{10}, \alpha_{01})$. Moreover we obtain from (2.2.15) that $\mathsf{E}(\eta-\alpha^*-\beta^*\xi)^2 = \sigma_2^2$, that is, the variance of the random variable η does not decrease after subtracting the linear regression $\alpha^* + \beta^*\xi$.

It is not hard to prove that if the general regression $y = m_2(x)$ of η on ξ is linear, then it coincides with the linear regression given by (2.2.13).

Solving the analogous problem (2.2.10), we find the linear regression $f^*(y) = \alpha^* + \beta^* y$ of ξ on η whose coefficients are given by

$$\beta^* = \beta_{12} = \frac{\mu_{11}}{\mu_{02}} = \rho\frac{\sigma_1}{\sigma_2}, \qquad \alpha^* = \alpha_{10} - \beta^*\alpha_{01}. \tag{2.2.16}$$

The number β_{12} is called the *coefficient of the linear regression of* ξ *on* η. Therefore the equation of the linear regression of ξ on η is of the form

$$x = \alpha_{10} + \beta_{12}(y - \alpha_{01}). \tag{2.2.17}$$

The regression can be rewritten in the canonical form:

$$\frac{x-\alpha_{10}}{\sigma_1} = \rho\frac{y-\alpha_{01}}{\sigma_2}. \tag{2.2.18}$$

The least variance of the random variable ξ is equal in this case to

$$\mathsf{E}(\xi-\alpha^*-\beta^*\eta)^2 = \sigma_1^2\left(1-\rho^2\right). \tag{2.2.19}$$

If $\rho = 0$, then it follows from (2.2.16) and (2.2.17) that the linear regression of ξ on η is $x = \alpha_{10}$; this is the straight line passing through the point $(\alpha_{10}, \alpha_{01})$ and parallel to the y-axis. It follows from (2.2.19) that $\mathsf{E}(\xi-\alpha^*-\beta^*\eta)^2 = \sigma_1^2$, that is, the least variance of ξ coincides with the variance σ_1^2.

If $|\rho| = 1$, then we obtain from (2.2.19) that $\xi = \alpha^* + \beta^*\eta$ with probability 1 where α^* and β^* are defined by (2.2.16). Moreover, we obtain from (2.2.14) and (2.2.18) that the linear regression of η on ξ and that of ξ on η coincide.

If $0 < |\rho| < 1$, then linear regression (2.2.18) can be rewritten in the form

$$\frac{y-\alpha_{01}}{\sigma_2} = \frac{1}{\rho}\,\frac{x-\alpha_{10}}{\sigma_1} \tag{2.2.20}$$

(cf. (2.2.14)). It follows from (2.2.14) and (2.2.20) that the linear regression of η on ξ and that of ξ on η coincide if and only if $|\rho| = 1$. Otherwise they do not coincide; in the case $\rho = 0$ they are perpendicular and each of them is parallel to the corresponding coordinate axis.

Parabolic regression. Let P be the family of polynomials

$$g(x) = c_0 + c_1 x + \cdots + c_k x^k$$

of degree k whose coefficients $c_0, c_1, \ldots, c_k$ are real. A polynomial

$$g^*(x) = c_0^* + c_1^* x + \cdots + c_k^* x^k$$

such that

$$\mathsf{E}(\eta - g^*(\xi))^2 = \min_{g \in P} \mathsf{E}(\eta - g(\xi))^2 \tag{2.2.21}$$

is called the *parabolic regression of* η *on* ξ. Assuming that all the moments occurring in (2.2.21) are finite we obtain the following condition for the minimum in (2.2.21):

$$\frac{1}{2}\frac{\partial G}{\partial c_i} = \mathsf{E}(\xi^i(g(\xi) - \eta)) = c_0\alpha_{i,0} + \cdots + c_k\alpha_{i+k,0} - \alpha_{i,1} = 0, \tag{2.2.22}$$

$i = 0, 1, \ldots, k$ (here $G = G(c_0, c_1, \ldots, c_k) = \mathsf{E}(\eta - g(\xi))^2$ for $g(x) = c_0 + c_1 x + \cdots + c_k x^k$). If the moments α_{ij} are known, the coefficients $c_0^*, c_1^*, \ldots, c_k^*$ can be determined from the above $k + 1$ equations. The evaluation of the coefficients can be simplified if the polynomial $g(x)$ is represented as a linear combination of *orthogonal polynomials* $p_i(x)$ *of degree* i *related to the distribution of* ξ and such that

$$\mathsf{E}p_m(\xi)p_l(\xi) = \begin{cases} 1, & m = l, \\ 0, & m \neq l. \end{cases} \tag{2.2.23}$$

Any polynomial $g(x)$ of degree k can be represented as

$$g(x) = c_0 p_0(x) + c_1 p_1(x) + \cdots + c_k p_k(x)$$

for some coefficients $c_0, c_1, \ldots, c_k$. According to (2.2.23) the condition for the minimum becomes of the form

$$\frac{1}{2}\frac{\partial G}{\partial c_i} = \mathsf{E}(p_i(\xi)(g(\xi) - \eta)) = c_i - \mathsf{E}\eta p_i(\xi), \qquad i = 0, 1, \ldots, k. \tag{2.2.24}$$

Using (2.2.24) we determine the coefficients $c_i^* = \mathsf{E}\eta p_i(\xi)$, whence the parabolic regression of η on ξ is

$$g^*(x) = c_0^* p_0(x) + c_1^* p_1(x) + \cdots + c_k^* p_k(x). \tag{2.2.25}$$

Note that the coefficients c_i^* do not depend on the degree k of the polynomial $g(x)$. Therefore one can apply the recursion to find the regression. Namely if the parabolic regression $g^*(x)$ of degree k is known in the form (2.2.25), then the parabolic regression of degree $k + 1$ can be obtained in the form

$$c_0^* p_0(x) + \cdots + c_k^* p_k(x) + c_{k+1}^* p_{k+1}(x)$$

by evaluating only one extra number $c_{k+1}^* = \mathsf{E}\eta p_{k+1}(\xi)$. By (2.2.23) we get

$$\mathsf{E}(\eta - g^*(\xi))^2 = \mathsf{E}\eta^2 - (c_0^*)^2 - \cdots - (c_k^*)^2 \tag{2.2.26}$$

where $g^*(x) = c_0^* p_0(x) + \cdots + c_k^* p_k(x)$. It is seen from (2.2.26) that a larger degree of the parabolic regression results in a smaller error of approximation $\mathsf{E}(\eta - g^*(\xi))$.

Note that the above relations hold not only for polynomials $p_i(x)$. In fact, an arbitrary sequence of functions $p_i(x)$ satisfying condition (2.2.23) can be used to construct the function $g(x) = \sum_{i=1}^{k} c_i p_i(x)$. Note that relations (2.2.24) and (2.2.26) remain true in this case, too.

To this end we note that the parabolic regression of ξ on η can be evaluated in the same way as in the case of problem (2.2.21). This regression possesses the same properties as that of η on ξ.

Sampling linear regression. Let $F_n(x, y)$ be the empirical distribution function constructed from sample (2.1.1) according to (2.1.2) and (2.1.3) where sample (2.1.1) consists of n independent observations of the random vector (ξ, η) with the distribution function $F(x, y)$.

The *sampling linear regression of η on ξ* is called a function $g^*(x) = \alpha^* + \beta^* x$ such that

$$\sum_{i=1}^{n} (\eta_i - g^*(\xi_i))^2 = \min_{g \in L} \sum_{i=1}^{n} (\eta_i - g(\xi_i))^2 \tag{2.2.27}$$

where L is the family of all linear functions. The *sampling linear regression of ξ on η* is called a function $f^*(y) = \alpha^* + \beta^* y$ such that

$$\sum_{i=1}^{n} (\xi_i - f^*(\eta_i))^2 = \min_{f \in L} \sum_{i=1}^{n} (\xi_i - f(\eta_i))^2. \tag{2.2.28}$$

To determine a linear regression $g^*(x) = \alpha^* + \beta^* x$ solving problem (2.2.27) it is sufficient to find $\alpha = \alpha^*$ and $\beta = \beta^*$ for which the function

$$G(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (\eta_i - \alpha - \beta \xi_i)^2 \tag{2.2.29}$$

attains the minimum. Taking into account (2.1.9) and (2.1.10) and applying the same argument as in the case of (2.2.11) we obtain

$$G(\alpha, \beta) = m_{20}\beta^2 - 2m_{11}\beta + m_{02} + (a_{01} - \beta a_{10} - \alpha)^2, \tag{2.2.30}$$

whence

$$\beta^* = b_{21} = \frac{m_{11}}{m_{20}} = r\frac{s_2}{s_1}, \qquad \alpha^* = a_{01} - \beta^* a_{10} \tag{2.2.31}$$

where r is the sampling coefficient of regression (see Section 2.1). The random variable b_{21} defined by (2.2.31) is called the *sampling coefficient of linear regression of η on ξ*. Using (2.2.31) we rewrite the regression equation $y = g^*(x)$ in the form

$$\frac{y - a_{01}}{s_2} = r\frac{x - a_{10}}{s_1}. \tag{2.2.32}$$

Equation (2.2.32) is called the *canonical equation of sampling linear regression of η on ξ*.

As above we get

$$G(\alpha^*, \beta^*) = \frac{1}{n} \sum_{i=1}^{n} (\eta_i - \alpha^* - \beta^* \xi_i)^2 = s_2^2 \left(1 - r^2\right). \tag{2.2.33}$$

Reasoning in the same way as in the case of the least variance (2.2.15) we obtain the same results about $G(\alpha^*, \beta^*)$ defined by (2.2.33). The only difference is that these results involve sampling characteristics instead of the corresponding characteristics of the random vector (ξ, η).

The method for obtaining sampling linear regressions as solutions of extremal problems (2.2.27) and (2.2.28) is called the *least squares method*. For example, the extremal problem (2.2.27) is to find the minimum of the sum of squares of distances evaluated in the y-direction between sampling points (ξ_i, η_i) and straight lines $y = \alpha + \beta x$. The same is true for the extremal problem (2.2.28) with the difference that the distances are evaluated in the x-direction between sampling points (ξ_i, η_i) and straight lines $y = \alpha + \beta x$.

Solving extremal problem (2.2.28) and using representations, similar to (2.2.29) and (2.2.30), we get the sampling linear regression $f^*(y) = \alpha^* + \beta^* y$ where

$$\beta^* = b_{12} = \frac{m_{11}}{m_{02}} = r\frac{s_1}{s_2}, \qquad \alpha^* = a_{10} - \beta^* a_{01}. \tag{2.2.34}$$

The random variable b_{12} defined by (2.2.34) is called the *coefficient of the sampling linear regression of ξ on η*. The canonical equation $x = \alpha^* + \beta^* y$ of the sampling linear regression of ξ on η is

$$\frac{x - a_{10}}{s_1} = r\frac{y - a_{01}}{s_2} \tag{2.2.35}$$

in view of (2.2.34). Reasoning in the same way as in the case of equalities (2.2.18) and (2.2.20) we prove the same results for sampling regressions (2.2.32) and (2.2.35).

The asymptotic behavior of coefficients b_{12} and b_{21} as $n \to \infty$ is as follows. If $\sigma_1 \neq 0$ and $\sigma_2 \neq 0$, then $r \to \rho$ in probability as $n \to \infty$, thus by (2.2.31) and (2.2.34) we get $b_{21} \to \beta_{21}$ and $b_{12} \to \beta_{12}$ in probability as $n \to \infty$. Further applying (2.1.14) and (2.1.16) we obtain

$$\mathsf{E}b_{12} = \beta_{12} + O\left(\frac{1}{n}\right), \qquad \mathsf{E}b_{21} = \beta_{21} + O\left(\frac{1}{n}\right)$$

provided that all necessary moments are finite. Moreover one can show that

$$\mathsf{D}b_{12} = \frac{c_{12}}{n} + O\left(n^{-3/2}\right), \qquad \mathsf{D}b_{21} = \frac{c_{21}}{n} + O\left(n^{-3/2}\right)$$

where c_{12} and c_{21} are some positive constants.

One can also show that b_{12} is asymptotically $\mathcal{N}(\beta_{12} n^{-1} c_{12})$ normal, and b_{21} is asymptotically $\mathcal{N}(\beta_{21}, n^{-1} c_{21})$ normal.

Sampling parabolic regression. Let P be a family of polynomials

$$g(x) = \gamma_0 + \gamma_1 x + \cdots + \gamma_k x^k$$

of degree k whose coefficients $\gamma_0, \gamma_1, \ldots, \gamma_k$ are real. A polynomial

$$g^*(x) = \gamma_0^* + \gamma_1^* x + \cdots + \gamma_k^* x^k$$

such that

$$\sum_{i=1}^{n} (\eta_i - g^*(\xi_i))^2 = \min_{g \in P} \sum_{i=1}^{n} (\eta_i - g(\xi_i))^2 \tag{2.2.36}$$

is called the *sampling parabolic regression of* η *on* ξ. Let $g(x) = \gamma_0 + \gamma_1 x + \ldots + \gamma_k x^k$ and

$$G(\gamma_0, \gamma_1, \ldots, \gamma_k) = \frac{1}{n}\sum_{i=1}^{n}(\eta_i - g(\xi_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\eta_i - \gamma_0 - \gamma_1\xi_i - \cdots - \gamma_k\xi_i^k\right)^2.$$

Then condition (2.2.36) can be rewritten in the form

$$\frac{1}{2}\frac{\partial G}{\partial \gamma_j} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^j(g(\xi_i) - \eta_i) = \gamma_0 a_{j,0} + \gamma_1 a_{j+1,0} + \cdots + \gamma_k a_{j+k,0} - a_{j,1} = 0, \qquad j = 0, 1, \ldots, k. \tag{2.2.37}$$

Condition (2.2.37) becomes simpler if the polynomial $g(x)$ is represented as a linear combination of orthogonal polynomials $p_m(x)$ of degree m related to the sampling distribution of ξ and such that

$$\frac{1}{n}\sum_{i=1}^{n}p_m(\xi_i)p_l(\xi_i) = \begin{cases} 1, & m = l, \\ 0, & m \neq l. \end{cases} \tag{2.2.38}$$

Every polynomial $g(x)$ of degree k can be represented as

$$g(x) = c_0 p_0(x) + \cdots + c_k p_k(x)$$

with some real coefficients $c_0, c_1, \ldots, c_k$. Using property (2.2.38) we rewrite condition (2.2.37) in the form

$$\frac{1}{2}\frac{\partial G}{\partial c_j} = \frac{1}{n}\sum_{i=1}^{n}p_j(\xi_i)(g(\xi_i) - \eta_i) = c_j - \frac{1}{n}\sum_{i=1}^{n}\eta_i p_j(\xi_i) = 0, \qquad j = 0, 1, \ldots, k. \tag{2.2.39}$$

This system of equations is easy to solve; its solution is given by

$$c_j^* = \frac{1}{n}\sum_{i=1}^{n}\eta_i p_j(\xi_i), \qquad j = 0, 1, \ldots, k. \tag{2.2.40}$$

Thus the sampling parabolic regression of η on ξ is

$$g^*(x) = c_0^* p_0(x) + c_1^* p_1(x) + \cdots + c_k^* p_k(x). \tag{2.2.41}$$

In view of (2.2.40) the coefficients c_j^* depend only on the polynomial $p_j(x)$. This implies that if the sampling parabolic regression $g^*(x)$ of degree k is represented in the form of (2.2.41) and one needs to obtain the regression of a higher degree, say of the degree $k+1$, that is,

$$g^*(x) = c_0^* p_0(x) + c_1^* p_1(x) + \cdots + c_k^* p_k(x) + c_{k+1}^* p_{k+1}(x),$$

then one can use coefficients c_j^*, $j = 0, 1, \ldots, k$, known from the preceding regression. The only extra work is to evaluate the coefficient c_{k+1}^* by using (2.2.40) for $j = k+1$. For the regression (2.2.41) we have

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i - g^*(\xi_i))^2 = a_{20} - (c_0^*)^2 - \cdots - (c_k^*)^2 \tag{2.2.42}$$

where a_{20} is the sampling moment of order 2 defined by (2.1.9). It follows from equality (2.2.42) that a larger degree of the regression results in a smaller error of approximation.

To this end we note that the sampling parabolic regression of ξ on η is defined similarly to (2.2.21). This regression possesses all the properties as does that of η on ξ.

REMARK 2.2.2. In this section we considered the regression analysis only for two-dimensional vectors. For the higher-dimensional case see [**1, 18, 28**].

Estimation of Unknown Parameters of Distributions

3.1. Statistical estimators and their quality measures

Parametric families of distributions and statistical estimators of parameters. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$. Let the probability distribution of the random element ξ be either unknown or partially known. Let $\{\mathsf{P}_\theta, \theta \in \Theta\}$ be a family of probability measures on $(X, \mathcal{B})$ and let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ be a k-dimensional parameter belonging to a subset $\Theta \subset \mathbf{R}^k$, $k \geq 1$. We assume that the distribution of the random element ξ depends on the parameter θ which is unknown for the statistician. Thus the measure P_θ is the distribution of ξ if the unknown parameter is equal to θ, that is, $\mathsf{P}_\theta\{\xi \in A\} = \mathsf{P}_\theta(A)$ for all $A \in \mathcal{B}$. The problem is to estimate the unknown parameter θ or a function $\phi(\theta)$ of the parameter θ with the help of the observation $\xi = x$.

The space X is called the *sampling space*. Every measurable function $T = T(x)$ mapping $(X, \mathcal{B})$ onto a measurable space $(Y, \mathcal{S})$ is called a *statistic*. If Θ is a Borel set of $\mathbf{R}^k$ and $\mathcal{B}(\Theta)$ is the σ-algebra of Borel subsets of Θ, then, in the case of $(Y, \mathcal{S}) = (\Theta, \mathcal{B}(\Theta))$, a statistic $T = T(x)$ is called a *statistical estimator* (or just *estimator*) of an unknown parameter θ constructed from an observation $\xi = x$. In the case of $(Y, \mathcal{S}) = (\mathbf{R}^k, \mathcal{B}^k)$ and $\mathbf{R}^k \neq \Theta$, we sometimes refer to a statistic $T = T(x)$ as an estimator of a parameter θ.

The notion of a statistical estimator of a function of a parameter θ can be introduced in an analogous way. The random variable $T = T(\xi)$ is also called an estimator of a parameter (or, an estimator of a function of a parameter).

Statistical estimators of a parameter θ introduced above are sometimes called *point estimators*. A point estimator T constructed from an observation $\xi = x$ provides a single value $t = T(x)$ which we treat as an approximation of the parameter. However the true value of the parameter can be (and usually is) different from an estimator. Therefore it is very important to know the error of approximation appearing due to a specific estimator. For this purpose, statisticians usually also indicate a region (an interval, if $k = 1$) such that the probability that the true value of a parameter θ belongs to the region is large.

Let $k = 1$; thus θ is a one-dimensional (scalar) parameter. Let $T_1 = T_1(x)$ and $T_2 = T_2(x)$ be two statistics with values in $\mathbf{R}^1$. Assume that $T_1 < T_2$ and for a given $\gamma \in (0, 1)$

$$\mathsf{P}_\theta\{T_1(\xi) < \theta < T_2(\xi)\} \geq \gamma \quad \text{for all } \theta \in \Theta. \tag{3.1.1}$$

The interval $(T_1(\xi), T_2(\xi))$ is called a *γ-confidence interval* or a *confidence interval of level γ* for the parameter θ. The number γ is often called the *confidence*

probability or *confidence level*. The numbers $T_1(\xi)$ and $T_2(\xi)$ are called the *lower* and *upper confidence bounds*, respectively.

Now let $k > 1$. Then a parameter θ is a k-dimensional vector and instead of γ-confidence intervals we define *γ-confidence regions* $G = G(\xi) \subset \mathbf{R}^k$ with the help of a condition similar to (3.1.1), namely

$$\mathsf{P}_\theta\{\theta \in G(\xi)\} \geq \gamma \quad \text{for all } \theta \in \Theta. \tag{3.1.2}$$

The γ-confidence intervals and γ-confidence regions defined by (3.1.1) and (3.1.2) can be constructed by using point estimators of the unknown parameter θ. For example, if T is a point estimator of a parameter θ, then, as a confidence interval, one can take $(T - \delta, T + \delta)$ where $\delta > 0$ is found from condition (3.1.1).

In what follows we often treat a sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ as an observed random element ξ. For this case, the sampling space is $(\mathbf{R}^n, \mathcal{B}^n)$, while the family of distributions of the sample $\xi^{(n)}$ is $\{\mathsf{P}_\theta^n, \theta \in \Theta\}$. The point estimation of parameters for this case is considered in the book by Lehmann [**21**].

Unbiased and consistent estimators. Let ξ be an observed random element, let $\theta = (\theta_1, \ldots, \theta_k)$ be an unknown parameter of the distribution, and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ be a statistical estimator of the parameter θ constructed from the observation ξ. An estimator $\widehat{\theta}$ is called *unbiased* if

$$\mathsf{E}_\theta \widehat{\theta} = \theta \quad \text{for all } \theta \in \Theta \tag{3.1.3}$$

where E_θ is the expectation with respect to the distribution P_θ.

In the case of estimation of a function $\phi(\theta)$ of a parameter θ, a statistic $\widehat{\phi}$ is called an unbiased estimator of the function $\phi(\theta)$ if

$$\mathsf{E}_\theta \widehat{\phi} = \phi(\theta) \quad \text{for all } \theta \in \Theta. \tag{3.1.4}$$

By $\widehat{\theta}_n = \widehat{\theta}(\xi^{(n)})$ we denote statistical estimators of a parameter θ in the case where an observed random element is a sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$. In such a case we deal with a sequence of estimators $\widehat{\theta}_n$, $n = 1, 2, \ldots$.

If an estimator $\widehat{\theta}$ does not satisfy condition (3.1.3), then we consider the *bias* of the estimator $\widehat{\theta}$ defined by

$$b(\theta) = \mathsf{E}_\theta \widehat{\theta} - \theta.$$

A sequence of estimators $\widehat{\theta}_n$, $n = 1, 2, \ldots$, is called an *asymptotically unbiased estimator* of a parameter θ if

$$\lim_{n \to \infty} \mathsf{E}_\theta \widehat{\theta}_n = \theta \quad \text{for all } \theta \in \Theta \tag{3.1.5}$$

or, in other words,

$$\lim_{n \to \infty} b_n(\theta) = 0 \quad \text{for all } \theta \in \Theta$$

where $b_n(\theta) = \mathsf{E}_\theta \widehat{\theta}_n - \theta$ is the bias of the estimator $\widehat{\theta}_n$. The notion of asymptotically unbiased estimators of a function $\phi(\theta)$ of a parameter can be introduced in a similar way.

When analyzing data, statisticians often restrict themselves to the case of unbiased estimators, since there exists a simple and useful theory of unbiased estimators where the quality of an estimator is measured by its variance.

On the other hand, there are many cases where the requirement that an estimator should be unbiased is too restrictive. For example, it is possible that unbiased estimators do not exist at all or are useless in practice for a given parametric model. To see this we consider the following examples.

EXAMPLE 3.1.1. Let ξ be a Poisson random variable with parameter $\theta > 0$, that is, $\mathsf{P}_\theta\{\xi = x\} = \theta^x e^{-x}/x!$, $x = 0, 1, 2, \ldots$. Assume that we want to estimate the function $\phi(\theta) = 1/\theta$ of the parameter θ by an observation ξ. Let $T = T(\xi)$ be an unbiased estimator of $\phi(\theta)$, that is, condition (3.1.4) holds. Then it can be rewritten as

$$\sum_{x=0}^{\infty} T(x)\frac{\theta^x}{x!}e^{-x} = \frac{1}{\theta} \quad \text{for all } \theta \in (0, \infty)$$

or, in other words,

$$\sum_{x=0}^{\infty} T(x)\frac{\theta^{x+1}}{x!} = e^{\theta} = \sum_{s=0}^{\infty}\frac{\theta^s}{s!} \quad \text{for all } \theta \in (0, \infty). \tag{3.1.6}$$

It is obvious that there is no function $T(x)$ that satisfies condition (3.1.6) for all $\theta \in (0, \infty)$ and does not depend on θ. This means that there is no unbiased estimator of $\phi(\theta) = 1/\theta$.

EXAMPLE 3.1.2. Let ξ have the geometric distribution with parameter

$$\theta \in (0, 1),$$

that is, $\mathsf{P}_\theta\{\xi = x\} = \theta^x(1-\theta)$, $x = 0, 1, 2, \ldots$. Assume that we want to estimate the parameter θ. Then the condition that $T = T(\xi)$ is an unbiased estimator is given by

$$\sum_{x=0}^{\infty} T(x)\theta^x = \frac{\theta}{1-\theta} = \sum_{s=1}^{\infty}\theta^s \qquad \text{for all } \theta \in (0, 1).$$

Comparing the coefficients for degrees of θ we see that the only unbiased estimator of θ is the statistic $T(x)$ such that $T(0) = 0$ and $T(x) = 1$ for $x = 1, 2, \ldots$. Since this statistic does not belong to the set $\Theta = (0, 1)$ of possible values of the parameter, it gives a wrong approximation of the true value of θ and the estimator is useless for practice.

The following example shows that, at least in some cases, an estimator with a small bias and small mean square error is better that an unbiased estimator with a large variance.

EXAMPLE 3.1.3. Let $\xi^{(n)}$ be a sample from the normal $\mathcal{N}(\theta_1, \theta_2^2)$ distribution, $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$. The unknown parameter is $\theta = (\theta_1, \theta_2)$. Consider the problem on estimating the function $\phi(\theta) = \theta_2^2$.

Consider the sampling variance

$$s^2 = m_2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - a_1)^2 \tag{3.1.7}$$

as an estimator where a_1 is the first order sampling moment defined by (1.2.2). Applying (1.2.9) for $k = 2$ we get $s^2 = \widetilde{a}_2 - \widetilde{a}_1^2$ where $\widetilde{a}_1$ and $\widetilde{a}_2$ are defined in the proof of Theorem 1.2.1. Note that $\mathsf{E}_\theta\widetilde{a}_2 = \theta_2^2$. In view of (1.2.10) for $k = 2$ we have $\mathsf{E}_\theta\widetilde{a}_1^2 = \theta_2^2/n$, whence $\mathsf{E}_\theta s^2 = \theta_2^2(n-1)/n$, that is, the estimator s^2 is biased

(however it is asymptotically unbiased). This implies that an unbiased estimator is given by

$$\widetilde{s}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - a_1)^2. \tag{3.1.8}$$

Theorem 1.4.2 implies that $\mathcal{L}((n-1)\widetilde{s}^2/\theta_2^2) = \chi^2(n-1)$. This together with (1.4.6) yields

$$\mathsf{D}_\theta \widetilde{s}^2 = \frac{2\theta_2^4}{n-1}. \tag{3.1.9}$$

Consider the class of estimators $T_\lambda = \lambda\widetilde{s}^2$, $\lambda \in (0,\infty)$. Since

$$\mathsf{E}_\theta T_\lambda = \lambda \mathsf{E}_\theta \widetilde{s}^2 = \lambda\theta_2^2,$$

there is only one unbiased estimator $\widetilde{s}^2$ of the function $\phi(\theta)$ in this class. The mean square error of the estimator T_λ equals

$$\mathsf{E}_\theta\left(T_\lambda - \theta_2^2\right)^2 = \left(\frac{2\lambda^2}{n-1} + (1-\lambda)^2\right)\theta_2^4. \tag{3.1.10}$$

The right-hand side of (3.1.10) attains its minimum at $\lambda^* = (n-1)/(n+1)$. Taking into account (3.1.9) we obtain

$$\mathsf{E}_\theta\left(T_{\lambda^*} - \theta_2^2\right)^2 = \frac{2}{n+1}\theta_2^4 < \frac{2}{n-1}\theta_2^4 = \mathsf{E}_\theta\left(\widetilde{s}^2 - \theta_2^2\right)^2.$$

Therefore the estimator T_{λ^*} has a smaller mean square error than that of the unbiased estimator $\widetilde{s}^2$. Since $\mathsf{E}_\theta T_{\lambda^*} = (n-1)\theta_2^2/(n+1)$, the estimator T_{λ^*} is asymptotically unbiased. Note that s^2 also is an asymptotically unbiased estimator, but $s^2 = T_{\lambda'}$ for $\lambda' = (n-1)/n \neq \lambda^*$. This means that the estimator s^2 is worse than T_{λ^*} in the sense of the minimum of the mean square error.

Moreover,

$$\mathsf{E}_\theta\left(T_{\lambda^*} - \theta_2^2\right)^2 < \mathsf{E}_\theta\left(s^2 - \theta_2^2\right)^2 < \mathsf{E}_\theta\left(\widetilde{s}^2 - \theta_2^2\right)^2. \tag{3.1.11}$$

Let $\widehat{\phi}_n$, $n = 1, 2, \ldots$, be a sequence of estimators of a function $\phi(\theta)$ of a parameter θ constructed from a sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$. A sequence of estimators $\widehat{\phi}_n$, $n = 1, 2, \ldots$, is called a *consistent sequence of estimators* of a function $\phi(\theta)$ if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}_\theta\{|\widehat{\phi}_n - \phi(\theta)| > \varepsilon\} = 0 \quad \text{for all } \theta \in \Theta. \tag{3.1.12}$$

For brevity we also say that an estimator $\widehat{\phi}_n$ satisfying condition (3.1.12) is a *consistent estimator* of the function $\phi(\theta)$.

Note that s^2 and $\widetilde{s}^2$ introduced in Example 3.1.3 by formulas (3.1.7) and (3.1.8) respectively, as well as $T_{\lambda^*} = \lambda^*\widetilde{s}^2$ are consistent estimators of the function

$$\phi(\theta) = \theta_2^2,$$

since they satisfy condition (3.1.12) in view of the Chebyshev inequality and relations (3.1.9) and (3.1.11).

We say that a sequence of estimators $\widehat{\phi}_n$, $n = 1, 2, \ldots$, is a *strongly consistent sequence of estimators of a function* $\phi(\theta)$ (alternatively, $\widehat{\phi}_n$ is a *strongly consistent estimator of a function* $\phi(\theta)$) if

$$\mathsf{P}_\theta \left\{ \lim_{n\to\infty} \widehat{\phi}_n = \phi(\theta) \right\} = 1. \tag{3.1.13}$$

For example, it follows from (1.1.4) that the empirical distribution function $F_n(x)$ is a strongly consistent estimator of the distribution function $F(x)$.

Consistent estimators with minimal variance. It is natural to compare unbiased estimators according to their variances. Let ξ be an observed random element with values in a measurable space $(X, \mathcal{B})$ and with a distribution belonging to a parametric family of probability measures $\{\mathsf{P}_\theta,\ \theta \in \Theta\}$.

Consider the problem of estimating a real function $g(\theta)$. Consider the following classes of estimators:

(1) the class U_g^θ of unbiased estimators $T = T(\xi)$ of the function $g(\theta)$ at a given point θ and such that $\mathsf{E}_\theta T^2 < \infty$,

(2) the class U_0^θ of unbiased estimators $T = T(\xi)$ of zero at a given point θ and such that $\mathsf{E}_\theta T^2 < \infty$.

Therefore

$$U_g^\theta = \left\{T\colon \mathsf{E}_\theta T = g(\theta),\ \mathsf{E}_\theta T^2 < \infty\right\},$$
$$U_0^\theta = \left\{T\colon \mathsf{E}_\theta T = 0,\ \mathsf{E}_\theta T^2 < \infty\right\}.$$

We also consider the following classes:

$$U_g = \bigcap_{\theta\in\Theta} U_g^\theta, \qquad U_0 = \bigcap_{\theta\in\Theta} U_0^\theta.$$

The following result contains necessary and sufficient conditions for an estimator to be optimal in the sense of minimum of the variance in the classes U_g^θ and U_g.

THEOREM 3.1.1. *The variance of an estimator $T \in U_g$ (respectively, $T \in U_g^\theta$) is minimal in the class U_g (respectively, in U_g^θ) if and only if $\mathsf{E}_\theta Th = 0$ for all $h \in U_0$ and $\theta \in \Theta$ (respectively, $\mathsf{E}_\theta Th = 0$ for a given $\theta \in \Theta$ and for all $h \in U_0^\theta$).*

PROOF. *Necessity.* Let an estimator $T \in U_g^\theta$ have minimal variance in the class U_g^θ for a fixed $\theta \in \Theta$ and $h \in U_0^\theta$. It is clear that $T + \lambda h \in U_g^\theta$ for all constants λ. Then

$$\mathsf{D}_\theta(T + \lambda h) = \mathsf{D}_\theta T + 2\lambda \mathsf{E}_\theta Th + \lambda^2 \mathsf{D}_\theta h.$$

If $\mathsf{E}_\theta Th \neq 0$, then there exists a number λ such that

$$2\lambda \mathsf{E}_\theta Th + \lambda^2 \mathsf{D}_\theta h < 0.$$

This implies that $\mathsf{D}_\theta(T + \lambda h) < \mathsf{D}_\theta T$, which contradicts the assumption that the estimator T has minimal variance in the class U_g^θ. Therefore $\mathsf{E}_\theta Th = 0$ for all $h \in U_0^\theta$.

Sufficiency. Let condition $\mathsf{E}_\theta Th = 0$ hold for all $h \in U_0^\theta$ and for an estimator $T \in U_g^\theta$ where $\theta \in \Theta$ is fixed. Let $T' \in U_g^\theta$ be another estimator. Then

$$T' - T = h \in U_0^\theta$$

and

$$\mathsf{D}_\theta T' = \mathsf{D}_\theta T + 2\mathsf{E}_\theta Th + \mathsf{D}_\theta h \geq \mathsf{D}_\theta T,$$

since $\mathsf{E}_\theta Th = 0$. This means that the estimator T has minimal variance in the class U_g^θ.

The proof for the class U_g is analogous. □

Theorem 3.1.1 is convenient for applications if a family of distributions contains a sufficiently wide class of unbiased estimators of zero.

Optimal estimators. As in the preceding section we consider the problem of estimating a function $g(\theta)$ from an observation ξ assuming values in $(X, \mathcal{B})$ and whose distribution belongs to a parametric family $\{\mathsf{P}_\theta,\ \theta \in \Theta\}$.

Let $T = T(\xi)$ be an estimator of a function $g(\theta)$ and let $r(T, g)$ be a nonnegative loss function (a loss appears because we approximate $g = g(\theta)$ by an estimator T). A function

$$R(T;\theta) = \mathsf{E}_\theta r(T(\xi), g(\theta)), \qquad \theta \in \Theta, \tag{3.1.14}$$

is called a *risk function of an estimator* $T = T(\xi)$ if the true value of the parameter is θ.

Examples of loss functions are presented by the quadratic function

$$r(T, g) = (T - g)^2$$

used in the preceding section, by the Laplace function $r(T, g) = |T - g|$, and by the function

$$r(T, g) = \begin{cases} 0, & |T - g| \leq b, \\ 1, & |T - g| > b, \end{cases}$$

where $b > 0$. The latter function appears in the interval estimation of parameters.

Sometimes we treat a risk function as a measure of quality of estimators. We consider the general case where an estimator is not necessarily unbiased and the loss function is not necessarily quadratic.

An estimator $T' = T'(\xi)$ belonging to a class $\mathcal{K}$ of estimators of a function $g(\theta)$ is called *admissible* for the class $\mathcal{K}$ with respect to a loss function $r(T, g)$ if there is no estimator $T \in \mathcal{K}$ such that

$$R(T;\theta) \leq R(T';\theta) \quad \text{for all } \theta \in \Theta \tag{3.1.15}$$

and inequality (3.1.15) is strict for at least one $\theta \in \Theta$. An estimator $T' = T'(\xi)$ that is admissible for the class of all estimators is called an *absolutely admissible* estimator of the function $g(\theta)$.

A statistic $T^* = T^*(\xi) \in \mathcal{K}$ is called an *optimal estimator of a function* $g(\theta)$ *in the class* $\mathcal{K}$ with respect to a loss function $r(T, g)$ if for all $T \in \mathcal{K}$

$$R(T^*;\theta) \leq R(T;\theta) \quad \text{for all } \theta \in \Theta.$$

A statistic $T' = T'(\xi) \in \mathcal{K}$ is called an *optimal estimator* (or, *locally optimal estimator*) *of a function* $g(\theta)$ *at a point* θ_0 *in the class* $\mathcal{K}$ with respect to a loss function $r(T, g)$ if for all $T \in \mathcal{K}$

$$R(T';\theta_0) \leq R(T;\theta_0).$$

It is clear that the set of risk functions $R(T;\theta)$ is unordered in the class of all estimators T. For this reason we consider narrower classes of either estimators or

distributions. One of the possible approaches here is to exclude estimators that are not admissible.

EXAMPLE 3.1.4. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample where $\xi_1, \xi_2, \dots, \xi_n$ are independent identically distributed random variables depending on an unknown parameter θ. Assume that $\theta = \mathsf{E}_\theta \xi_1$. Let $T = T(\xi^{(n)}) = \xi_1$. It is obvious that T is an unbiased estimator of the parameter θ. Let $\mathsf{E}_\theta \xi_1^2 < \infty$ for all θ and let $r(T, \theta) = |T - \theta|^2$ be the loss function. Then $R(T; \theta) = \mathsf{E}_\theta |\xi_1 - \theta|^2 = \mathsf{D}_\theta \xi_1$. The estimator $T' = T'(\xi) = n^{-1} \sum_{i=1}^n \xi_i$ also is an unbiased estimator of the parameter θ. In this case

$$R(T'; \theta) = \mathsf{E}_\theta |T' - \theta|^2 = \frac{1}{n} \mathsf{D}_\theta \xi_1 \le R(T; \theta).$$

Moreover the inequality becomes strict if $n \ge 2$. Therefore T is not an admissible estimator in the class of unbiased estimators of the parameter θ.

Below we consider the Bayes and minimax approaches allowing one to avoid the problem that the set of risk functions is unordered.

The Bayes and minimax approaches. Let Θ be an open set of $\mathbf{R}^k$ and let $\mathbf{Q}$ be a σ-finite measure on Θ. Without loss of generality we can extend the measure Θ to the whole space $\mathbf{R}^k$ by putting $\mathbf{Q}(\mathbf{R}^k \setminus \Theta) = 0$. Let $T = T(\xi)$ be some estimator of a function $g(\theta)$ and let $R(T; \theta)$ be a risk function of the estimator T defined by (3.1.14). The number

$$R(T) = \int R(T; \theta)\, \mathbf{Q}(d\theta) \tag{3.1.16}$$

is called the *risk of the estimator* T. The measure $\mathbf{Q}$ is called the *a priori measure*. An estimator T^* is called a *Bayes* estimator of a function $g(\theta)$ with respect to a loss function $r(T, g)$ and the a priori measure $\mathbf{Q}$ if

$$R(T^*) \le R(T)$$

for all estimators T where the risk $R(T)$ is defined by (3.1.16). In other words,

$$R(T^*) = \min_T R(T)$$

for a Bayes estimator.

Sometimes an estimator is called Bayes only in the case where $\mathbf{Q}$ is a probability measure. Otherwise an estimator minimizing the risk (3.1.16) is called a *generalized Bayes estimator*.

Note that we can think of θ as a random parameter with distribution $\mathbf{Q}$ if $\mathbf{Q}$ is a probability measure. Then all Bayes estimators T^* are of the form

$$T^* = \mathsf{E}\left\{ g(\theta) \,/\, \xi \right\}$$

in the case of $r(T, g) = (T - g)^2$ where the conditional expectation is evaluated with respect to the conditional distribution of the parameter θ subject to ξ. In its turn, the latter distribution can be found by the Bayes formula and this explains why these estimators are called Bayes. In this case, a Bayes estimator minimizes the possibility that the risk is the mean square error

$$R(T) = \mathsf{E}(T(\xi) - g(\theta))^2 = \mathsf{E} R(T; \theta).$$

Another approach is based on the comparison of maximums of risk functions for estimators $\sup_{\theta\in\Theta} R(T;\theta)$. A statistic $T' = T'(\xi)$ is called a *minimax* estimator of a function $g(\theta)$ with respect to a loss function $r(T,g)$ if for all estimators T

$$\sup_{\theta\in\Theta} R(T';\theta) \le \sup_{\theta\in\Theta} R(T;\theta).$$

In other words,

$$\sup_{\theta\in\Theta} R(T';\theta) = \inf_{T} \sup_{\theta\in\Theta} R(T;\theta)$$

for a minimax estimator T'. There are many relations between minimax and Bayes estimators; some of them are given below.

THEOREM 3.1.2. *Let T^* be a Bayes estimator of a function $g(\theta)$ with respect to a loss function $r(T,g)$ and the a priori probability measure $\mathbf{Q}$. If there is an estimator T' such that*

$$R(T';\theta) \le \int R(T^*;t)\,\mathbf{Q}(dt) \tag{3.1.17}$$

for all $\theta\in\Theta$, then the estimator T' is minimax.

PROOF. Let T be an arbitrary estimator of a function $g(\theta)$. Then for all $t\in\Theta$

$$\sup_{\theta\in\Theta} R(T;\theta) \ge \int R(T;t)\,\mathbf{Q}(dt) \ge \int R(T^*;t)\,\mathbf{Q}(dt) \ge R(T';t). \qquad \square$$

Assume that the measure $\mathbf{Q}$ possesses the density $q(t)$. Consider the set

$$N_Q = \{t: q(t) > 0\}.$$

Note that inequality (3.1.17) becomes an equality for almost all $\theta\in N_Q$, since otherwise

$$\int R(T';\theta)q(\theta)\,d\theta < \int R(T^*;\theta)q(\theta)\,d\theta,$$

contradicting the assumption that the estimator T^* is Bayes. This remark allows one to obtain the following criterion, which is equivalent to Theorem 3.1.2.

THEOREM 3.1.3. *Assume that an estimator T exists such that*

1) *T is a Bayes estimator with respect to some probability measure $\mathbf{Q}$ possessing density $q(t)$;*
2) *$R(T;t) = c = \text{const}$ for $t\in N_Q$;*
3) *$R(T;t) \le c$ for $t\notin N_Q$.*

Then the estimator T is minimax.

If an estimator T' is minimax and is Bayes with respect to a probability measure $\mathbf{Q}$ with density $q(t)$, then

$$\sup_{t\in\Theta} R(T';t) = \int R(T';t)q(t)\,dt.$$

Therefore any minimax estimator is a Bayes estimator that smooths the risk function. This means that the a priori measure $\mathbf{Q}'$ related to this estimator suggests that statisticians pay the same attention to all possible parameters θ, instead of the approach suggested by Bayes estimators $T^* = T^*_Q$ that corresponds to other a priori

measures $\mathbf{Q} \neq \mathbf{Q}'$, namely to pay special attention to some (the most probable) values of θ. Therefore

$$\int R(T_Q^*; t)\, \mathbf{Q}(dt) \leq \int R(T'; t)\, \mathbf{Q}'(dt).$$

This inequality explains why the distribution $\mathbf{Q}'$ in Theorem 3.1.3 corresponding to a minimax estimator T' is often called the *worse* or *least favorable*.

The least favorable distribution $\mathbf{Q}'$ does not always exists, thus one can use the following criterion for minimax estimators.

THEOREM 3.1.4. *Assume that there are an estimator T' and a sequence of distributions $\mathbf{Q}_m$, $m = 1, 2, \ldots$, possessing the densities $q_m(t)$ such that*

$$\mathsf{E}_t r(T'; g(t)) \leq \lim_{m\to\infty} \sup \int \mathsf{E}_t(T_m^*; g(t)) q_m(t)\, dt \tag{3.1.18}$$

for all $t \in \Theta$ where T_m^* is a Bayes estimator of a function $g(\theta)$ with respect to a loss function $r(T; g(t))$ and the a priori distribution $\mathbf{Q}_m$. Then the estimator T' is minimax.

PROOF. For all estimators T of a function $g(\theta)$,

$$\sup_t \mathsf{E}_t r(T; g(t)) \geq \int \mathsf{E}_t r(T; g(t)) q_m(t)\, dt \geq \int \mathsf{E}_t r(T_m^*; g(t)) q_m(t)\, dt.$$

According to (3.1.18) this implies that

$$\sup_t \mathsf{E}_t r(T; g(t)) \geq \lim_{m\to\infty} \sup \int \mathsf{E}_t r(T_m^*; g(t)) q_m(t)\, dt \geq \sup_t \mathsf{E}_t r(T'; g(t)),$$

whence it follows that the estimator T' is minimax. □

EXAMPLE 3.1.5. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample from the normal $\mathcal{N}(\theta, 1)$ distribution. Let the a priori measure also be normal $\mathbf{Q}_m = \mathcal{N}(0, m)$ distribution where m is the variance. Then the Bayes estimator of the parameter θ with respect to the quadratic loss function and the a priori $\mathcal{N}(0, m)$ distribution coincides with the a posteriori mean $\mathsf{E}(\theta/\xi^{(n)} = x) = \theta_m^*(x)$. Simple calculations show that

$$\theta_m^*(x) = \frac{1}{n}\sum_{i=1}^n x_i \Big/ \left(1 + \frac{1}{nm}\right).$$

The variance of the a posteriori distribution is

$$\mathsf{D}\left(\theta \,/\, \xi^{(n)} = x\right) = \frac{m}{1 + nm},$$

whence it follows that the mean square error of the estimator θ_m^* is

$$\mathsf{E}(\theta_m^* - \theta)^2 = \mathsf{ED}\left(\theta \,/\, \xi^{(n)}\right) = \frac{m}{1+nm} = \int \mathsf{E}_t(\theta_m^* - t) q_m(t)\, dt$$

where $q_m(t)$ is the density of $\mathbf{Q}_m$ with respect to the Lebesgue measure. This implies for the estimator $\widetilde{\theta}_n(x) = n^{-1}\sum_{i=1}^n x_i$ that

$$\mathsf{E}_t(\widetilde{\theta}_n - t)^2 = \frac{1}{n} = \lim_{m\to\infty} \int \mathsf{E}_t(\theta_m^* - t) q_m(t)\, dt.$$

Thus the estimator $\widetilde{\theta}_n$ is minimax by Theorem 3.1.4. Note also that the least favorable distribution does not exist in this case.

We show in Example 3.2.1 below that $\widetilde{\theta}_n$ is a Pitman estimator of the parameter θ. By (3.2.9) this estimator is Bayes with respect to the quadratic loss function and the Lebesgue measure taken as a priori measure.

The following is an example where the least favorable distribution exists.

EXAMPLE 3.1.6. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the binomial distribution with parameter $\theta \in \Theta = [0, 1]$, that is, the random variables $\xi_1, \dots, \xi_n$ are independent, identically distributed, and assume values 1 and 0 with probabilities θ and $1 - \theta$, respectively. For the estimator $\widetilde{\theta}_n(x) = n^{-1} \sum_{i=1}^n x_i$,

$$\mathsf{E}_\theta(\widetilde{\theta}_n - \theta)^2 = \theta(1-\theta)/n.$$

Hence the assumption of Theorem 3.1.3 does not hold for this estimator. Consider another estimator

$$\theta_n^*(x) = \left(\widetilde{\theta}_n(x) + \frac{1}{2\sqrt{n}}\right)\left(1 + \frac{1}{\sqrt{n}}\right)^{-1} \tag{3.1.19}$$

for which

$$\mathsf{E}_\theta(\theta_n^* - \theta)^2 = \left(1 + \frac{1}{\sqrt{n}}\right)^{-2} \mathsf{E}_\theta\left(\widetilde{\theta}_n - \theta + \frac{1}{2\sqrt{n}} - \frac{\theta}{\sqrt{n}}\right)^2 = \frac{1}{4(1+\sqrt{n})^2},$$

that is, the risk function of the estimator θ_n^* does not depend on θ.

Let B_{λ_1,λ_2} be the beta distribution with density

$$\frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, t^{\lambda_1 - 1}(1-t)^{\lambda_2 - 1}, \qquad 0 \le t \le 1, \tag{3.1.20}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are two parameters of the distribution. Let the a priori distribution $\mathbf{Q}$ be a Beta $B_{N+1,N+1}$ distribution. It can be proved in this case that the a priori distribution coincides with the Beta distribution (3.1.20) with parameters $\lambda_1 = N + n\widetilde{\theta}_n(x) + 1$ and $\lambda_2 = N + n(1 - \widetilde{\theta}_n(x)) + 1$. Since the mean value of the distribution B_{λ_1,λ_2} is $\lambda_1/(\lambda_1 + \lambda_2)$, the Bayes estimator with respect to the a priori distribution $\mathbf{Q} = B_{N+1,N+1}$ and the quadratic loss function is

$$\theta_n^{\mathbf{Q}}(x) = \frac{N + n\widetilde{\theta}_n(x) + 1}{2N + n + 2} = \frac{\widetilde{\theta}_n(x) + (N+1)/n}{1 + 2(N+1)/n}.$$

If $N + 1 = \sqrt{n}/2$, then the latter estimator coincides with the estimator $\theta_n^*(x)$ defined by (3.1.19). By Theorem 3.1.3 θ_n^* is minimax. On the other hand, it is known that this is a Bayes estimator with respect to the a priori distribution

$$\mathbf{Q} = B_{N+1,N+1}$$

for $N = \sqrt{n}/2 - 1$. This means that a priori distribution is the least favorable. If n increases, then the support of this distribution tends to concentrate in a neighborhood of the least favorable value of the parameter $\theta = 1/2$ for which the variance $\theta(1-\theta)/n = 1/(4n)$ of the estimator $\widetilde{\theta}_n$ is maximal. The estimator $\widetilde{\theta}_n$ itself is not minimax, since

$$\sup_\theta \frac{\theta(1-\theta)}{n} = \frac{1}{4n} > \frac{1}{4(1+\sqrt{n})^2}.$$

It is also clear that for all θ outside a small enough neighborhood of the point $\theta = 1/2$ the estimator $\widetilde{\theta}_n$ is better than θ_n^Q. The small neighborhood of the point $\theta = 1/2$ mentioned above is determined by the inequality

$$\theta(1-\theta) < \frac{1}{4(1+1/\sqrt{n})^2}.$$

REMARK 3.1.1. In the general case, it is not always possible to give explicit expressions for Bayes and minimax estimators.

3.2. Estimation of a location parameter

In this and the next sections we show how to estimate unknown location and scale parameters of a distribution in the cases where optimal estimators exist.

Location parameters. Equivariant estimators. Let an observed element $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a vector whose coordinates are, generally speaking, dependent random variables, let $(\mathbf{R}^n, \mathcal{B}^n)$ be a sampling space, and let $(\mathsf{P}_\theta^n, \theta \in \Theta)$ be a family of distributions to which the distribution of the vector $\xi^{(n)}$ belongs. Assume that the parameter θ is one-dimensional and $\Theta = \mathbf{R}^1$. If measures P_θ^n depend on the parameter θ in such a way that

$$\mathsf{P}_\theta^n(A) = \mathsf{P}_0^n(A-\theta) \quad \text{for all } A \in \mathcal{B}^n, \tag{3.2.1}$$

then θ is called a *location parameter*. We use the notation

$$A - \theta = \{x - \theta = (x_1 - \theta, \dots, x_n - \theta) : x = (x_1, \dots, x_n) \in A\}$$

for all $A \in \mathcal{B}^n$ in equality (3.2.1). One of the models leading to distributions (3.2.1) is the so-called scheme of direct observations where

$$\xi_i = \theta + \varepsilon_i, \qquad i = 1, \dots, n,$$

and $\varepsilon_1, \dots, \varepsilon_n$ are, generally speaking, dependent random variables with the joint distribution defined by the measure P_0^n.

Let $\mathsf{P}_\theta\{\xi^{(n)} \in A\} = \mathsf{P}_\theta^n(A)$ for all $A \in \mathcal{B}^n$. Then condition (3.2.1) can be rewritten in an equivalent form as

$$\mathsf{P}_\theta\left\{\xi^{(n)} - \theta \in A\right\} = \mathsf{P}_0\left\{\xi^{(n)} \in A\right\} \quad \text{for all } A \in \mathcal{B}^n \tag{3.2.2}$$

where $\xi^{(n)} - \theta = (\xi_1 - \theta, \dots, \xi_n - \theta)$. Condition (3.2.2) means that if the true value of the parameter is θ, then the vector $\xi^{(n)} - \theta$ has the same distribution as the vector $\xi^{(n)}$ corresponding to the zero value of the parameter.

There is a natural class of estimators used in the estimation of location parameters, namely

$$\mathcal{T} = \left\{\widetilde{\theta}_n = \widetilde{\theta}_n(x) : \widetilde{\theta}_n(x+c) = \widetilde{\theta}_n(x) + c \text{ for all } x \in \mathbf{R}^n \text{ and all } c \in \mathbf{R}^1\right\}. \tag{3.2.3}$$

Estimators of the class $\mathcal{T}$ are called *equivariant* estimators of a location parameter. Some authors call such estimators "invariant".

Let $r(\widetilde{\theta}_n; \theta) = r(\widetilde{\theta}_n - \theta)$ be a nonnegative loss function depending on the difference of arguments $\widetilde{\theta}_n - \theta$, and let $R(\widetilde{\theta}_n; \theta) = \mathsf{E}_\theta r(\widetilde{\theta}_n - \theta)$ be the risk function of

the estimator $\widetilde{\theta}_n$ (here E_θ stands for the expectation with respect to the measure P_θ). If $\widetilde{\theta}_n \in \mathcal{T}$, then

$$R(\widetilde{\theta}_n;\theta) = \mathsf{E}_\theta r(\widetilde{\theta}_n - \theta) = \mathsf{E}_\theta r\left(\widetilde{\theta}_n\left(\xi^{(n)} - \theta\right)\right) = \mathsf{E}_0 r(\widetilde{\theta}_n) = \text{const} \tag{3.2.4}$$

by (3.2.2), that is, the risk function $R(\widetilde{\theta}_n;\theta)$ does not depend on θ. Thus the estimator $\widetilde{\theta}_n \in \mathcal{T}$ is either optimal in $\mathcal{T}$ or not admissible in $\mathcal{T}$ for such loss functions.

An optimal in the class $\mathcal{T}$ estimator $\widehat{\theta}_n$ is called the *Pitman estimator* of a location parameter θ corresponding to a loss function $r(\widetilde{\theta}_n - \theta)$ if

$$R(\widehat{\theta}_n;\theta) = \min_{\widetilde{\theta}_n \in \mathcal{T}} R(\widetilde{\theta}_n;\theta) \quad \text{for all } \theta \in \mathbf{R}^1.$$

Below we show that the Pitman estimator exists and find it for some loss functions.

We mention another useful property of equivariant estimators. Let $\widetilde{\theta}_n$ and $\widetilde{\theta}'_n$ be two equivariant estimators. Then by the definition of equivariant estimators

$$\widetilde{\theta}_n(x) - \widetilde{\theta}'_n(x) = \psi(y), \qquad y = (x_2 - x_1, x_3 - x_1, \dots, x_n - x_1), \tag{3.2.5}$$

where $\psi(y)$ is some measurable function. Indeed,

$$\begin{aligned}\widetilde{\theta}_n(x) - \widetilde{\theta}'_n(x) &= [\widetilde{\theta}_n(x) - x_1] - [\widetilde{\theta}'_n(x) - x_1] = \widetilde{\theta}_n(x - x_1) - \widetilde{\theta}'_n(x - x_1) \\ &= \widetilde{\theta}_n(0, x_2 - x_1, \dots, x_n - x_1) - \widetilde{\theta}'_n(0, x_2 - x_1, \dots, x_n - x_1),\end{aligned}$$

whence (3.2.5) follows.

The Pitman estimator of a location parameter. In what follows we use the following notation. If $T = T(x)$, $x \in \mathbf{R}^n$, is a statistic, then $\mathsf{E}_\theta(T/y)$ denotes the conditional expectation

$$\mathsf{E}_\theta(T/y) = \mathsf{E}_\theta\left\{T\left(\xi^{(n)}\right) / \eta = y\right\} \tag{3.2.6}$$

where $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$, the vector y is defined by (3.2.5), and

$$\eta = (\xi_2 - \xi_1, \xi_3 - \xi_1, \dots, \xi_n - \xi_1). \tag{3.2.7}$$

The following result establishes the Pitman estimator of a location parameter with respect to the quadratic loss function.

THEOREM 3.2.1. *Assume that* $\mathsf{E}_0\xi_i^2 < \infty$ *for all* $i = 1, 2, \dots, n$. *Let*

$$l(x) = \sum_{i=1}^n c_i x_i, \qquad x = (x_1, \dots, x_n),$$

be a linear statistic such that $\sum_{i=1}^n c_i = 1$. *Then*

1) *the estimator*

$$\widehat{\theta}_n(x) = l(x) - \mathsf{E}_0(l/y), \qquad x \in \mathbf{R}^n, \tag{3.2.8}$$

is the Pitman estimator of a location parameter θ *with respect to the quadratic loss function* $r(\widetilde{\theta}_n, \theta) = |\widetilde{\theta}_n - \theta|^2$;

2) *if the measure* P_0 *is absolutely continuous with respect to the Lebesgue measure and its density is* $f(x)$, $x \in \mathbf{R}^n$, *then the Pitman estimator is of the form*

$$\widehat{\theta}_n(x) = \int_{-\infty}^{\infty} v f(x-v)\,dv \left(\int_{-\infty}^{\infty} f(x-v)\,dv \right)^{-1}, \qquad x \in \mathbf{R}^n. \tag{3.2.9}$$

PROOF. It is clear that $\widehat{\theta}_n \in \mathcal{T}$. Let $\widetilde{\theta}_n$ be an arbitrary equivariant estimator of $\mathcal{T}$. Then $\widetilde{\theta}_n(x) = \widehat{\theta}_n(x) + \psi(y)$ by (3.2.5). If $\mathsf{E}_0\widetilde{\theta}_n^2 = \infty$, then

$$\mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 \le \mathsf{E}_\theta(\widetilde{\theta}_n - \theta)^2$$

for all $\theta \in \mathbf{R}^1$. On the other hand, if $\mathsf{E}_0\widetilde{\theta}_n^2 < \infty$, then

$$\begin{aligned} \mathsf{E}_\theta(\widetilde{\theta}_n - \theta)^2 &= \mathsf{E}_\theta(\widetilde{\theta}_n - \widehat{\theta}_n + \widehat{\theta}_n - \theta)^2 \\ &= \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 + 2\mathsf{E}_\theta(\widetilde{\theta}_n - \widehat{\theta}_n)(\widehat{\theta}_n - \theta) + \mathsf{E}_\theta(\widetilde{\theta}_n - \widehat{\theta}_n)^2. \end{aligned} \tag{3.2.10}$$

It follows from (3.2.2) and (3.2.5) that

$$\begin{aligned} \mathsf{E}_\theta(\widetilde{\theta}_n - \widehat{\theta}_n)(\widehat{\theta}_n - \theta) &= \mathsf{E}_0(\widetilde{\theta}_n - \widehat{\theta}_n)\widehat{\theta}_n = \mathsf{E}_0\mathsf{E}_0\{(\widetilde{\theta}_n - \widehat{\theta}_n)\widehat{\theta}_n \,/\, \eta\} \\ &= \mathsf{E}_0(\widetilde{\theta}_n - \widehat{\theta}_n)\mathsf{E}_0\{\widehat{\theta}_n \,/\, \eta\} = 0 \end{aligned} \tag{3.2.11}$$

in view of (3.2.6) and (3.2.7) where $\widehat{\theta}_n$ is defined by (3.2.8). Thus

$$\mathsf{E}_0(\widehat{\theta}_n/y) = \mathsf{E}_0(l/y) - \mathsf{E}_0(l/y) = 0.$$

Relations (3.2.10) and (3.2.11) imply that for all $\theta \in \mathbf{R}^1$ we have

$$\mathsf{E}_\theta(\widetilde{\theta}_n - \theta)^2 = \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 + \mathsf{E}_\theta(\widetilde{\theta}_n - \widehat{\theta}_n)^2 \ge \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2.$$

Therefore we proved that the estimator $\widehat{\theta}_n$ defined by (3.2.8) is the Pitman estimator of a location parameter with respect to the quadratic loss function.

Now we prove equality (3.2.9). Let $l(x) = x_1$. Then estimator (3.2.8) is of the form

$$\widehat{\theta}_n(x) = x_1 - \mathsf{E}_0(\xi_1/y), \qquad x \in \mathbf{R}^n. \tag{3.2.12}$$

Consider random variables $\zeta_1 = \xi_1$, $\zeta_2 = \xi_2 - \xi_1, \ldots, \zeta_n = \xi_n - \xi_1$. Let $p(z_1, z_2, \ldots, z_n)$ be the probability density of the vector $(\zeta_1, \zeta_2, \ldots, \zeta_n)$ for $\theta = 0$. It is not hard to show that

$$p(z_1, z_2, \ldots, z_n) = f(z_1, z_1 + z_2, \ldots, z_1 + z_n)$$

where $f(x)$, $x \in \mathbf{R}^n$, is the density of the distribution P_0. It is obvious that

$$
\begin{aligned}
\mathsf{E}_0(\xi_1/y) &= \mathsf{E}_0\left\{\zeta_1 \,/\, \zeta_2 = x_2 - x_1, \ldots, \zeta_n = x_n - x_1\right\} \\
&= \int_{-\infty}^{\infty} zp(z, x_2 - x_1, \ldots, x_n - x_1)\, dz \\
&\quad \times \left(\int_{-\infty}^{\infty} p(z, x_2 - x_1, \ldots, x_n - x_1)\, dz\right)^{-1} \\
&= \int_{-\infty}^{\infty} zf(z, z + x_2 - x_1, \ldots, z + x_n - x_1)\, dz \\
&\quad \times \left(\int_{-\infty}^{\infty} f(z, z + x_2 - x_1, \ldots, z + x_n - x_1)\, dz\right)^{-1} \\
&= x_1 + \int_{-\infty}^{\infty} vf(x_1 - v, x_2 - v, \ldots, x_n - v)\, dv \\
&\quad \times \left(\int_{-\infty}^{\infty} f(x_1 - v, x_2 - v, \ldots, x_n - v)\, dv\right)^{-1}
\end{aligned}
\tag{3.2.13}
$$

Combining equalities (3.2.12) and (3.2.13) we obtain representation (3.2.9). □

REMARK 3.2.1. When proving representation (3.2.9) we put $l(x) = x_1$ in (3.2.8). Note in the general case that

$$\mathsf{E}_0(l/y) = l(x) - x_1 + \mathsf{E}_0(\xi_1/y), \qquad x \in \mathbf{R}^n,$$

where we used the property that $\sum_{i=1}^n c_i = 1$. Thus estimator (3.2.8) becomes of the form $\widehat{\theta}_n(x) = x_1 - \mathsf{E}_0(\xi_1/y)$ and this result is used in the above proof.

REMARK 3.2.2. If the measure P_0 is absolutely continuous with respect to the Lebesgue measure, then it is seen from equality (3.2.9) that the Pitman estimator is the Bayes estimator with respect to the quadratic loss function and the a priori measure $\mathbf{Q}$ coinciding with the Lebesgue measure. In other words, the Pitman estimator is a generalized Bayes estimator.

REMARK 3.2.3. In fact, statement 1) of Theorem 3.2.1 is a particular case of Theorem 3.1.1. Namely, the optimality of the estimator $\widehat{\theta}_n$ in the class $\mathcal{T}$ means that $\widehat{\theta}_n$ is orthogonal to unbiased estimators of zero and the latter are of the form $h(x_2 - x_1, \ldots, x_n - x_1)$.

The following result establishes the Pitman estimator of a location parameter with respect to the Laplace loss function $r(\widetilde{\theta}, \theta) = |\widetilde{\theta} - \theta|$. We use the notation $\mathrm{med}_0(l/y)$ for a median of the conditional distribution $l(\xi)$ in the case of $\theta = 0$ given condition $\eta = y$ for which $\mathrm{med}_0(l/y)$ is a statistic.

THEOREM 3.2.2. *Let* $\mathsf{E}_0|\xi_i| < \infty$ *for all* $i = 1, 2, \ldots, n$. *Let* $l(x) = \sum_{i=1}^n c_i x_i$, $x = (x_1, x_2, \ldots, x_n)$, *be some linear statistic such that* $\sum_{i=1}^n c_i = 1$. *Then*

$$\overline{\theta}_n(x) = l(x) - \mathrm{med}_0(l/y), \qquad x \in \mathbf{R}^n, \tag{3.2.14}$$

is the Pitman estimator of a location parameter with respect to the Laplace loss function.

PROOF. Let $\widetilde{\theta}_n \in \mathcal{T}$. Then $\widetilde{\theta}_n(x) = l(x) + \psi(y)$ by (3.2.5), since $l \in \mathcal{T}$. Thus

$$\begin{aligned} \mathsf{E}_\theta|\widetilde{\theta}_n - \theta| &= \mathsf{E}_0|l + \psi(\eta)| = \mathsf{E}_0\mathsf{E}_0(|l + \psi(\eta)|/\eta) \\ &\geq \mathsf{E}_0\mathsf{E}_0(|l - \operatorname{med}_0(l/\eta)|/\eta) = \mathsf{E}_0|\overline{\theta}_n| = \mathsf{E}_\theta|\overline{\theta}_n - \theta|, \end{aligned} \tag{3.2.15}$$

since for all y

$$\mathsf{E}_0(|l + \psi(y)|/y) \geq \mathsf{E}_0(|l - \operatorname{med}_0(l/y)|/y). \tag{3.2.16}$$

It follows from (3.2.15) that estimator (3.2.14) is the Pitman estimator of a location parameter with respect to the Laplace loss function. □

REMARK 3.2.4. In the proof of Theorem 3.2.2 we used inequality (3.2.16) which is a well-known property of a median of a distribution (see [9]).

Below are some examples of the evaluation of Pitman estimators.

EXAMPLE 3.2.1. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed $\mathcal{N}(\theta, 1)$ random variables. Then the assumption of statement 2) of Theorem 3.2.1 holds, and moreover

$$f(x) = (2\pi)^{-n/2}\exp\left\{-2^{-1}\sum_{i=1}^{n} x_i^2\right\}, \qquad x = (x_1, \ldots, x_n).$$

Substituting this density into (3.2.9), we show that the Pitman estimator is of the form $\widehat{\theta}_n(x) = n^{-1}\sum_{i=1}^{n} x_i$. This implies that $\mathsf{D}_\theta\widehat{\theta}_n = \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 = n^{-1}$.

EXAMPLE 3.2.2. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent exponential random variables with the density $e^{-x+\theta}$, $x \geq \theta$. Again the assumption of statement 2) of Theorem 3.2.1 holds, and moreover

$$f(x) = \exp\left\{-\sum_{i=1}^{n} x_i\right\} I_{[0,\infty)}\left(\min_{1\leq i\leq n} x_i\right), \qquad x = (x_1, \ldots, x_n).$$

Substituting this density into (3.2.9) we obtain for the Pitman estimator

$$\widehat{\theta}_n(x) = \min_{1\leq i\leq n} x_i - \frac{1}{n}, \qquad x = (x_1, \ldots, x_n).$$

Using (1.3.3) for the density of the first order statistic $\zeta_{1,n}$ we obtain that

$$\mathsf{D}_\theta\widehat{\theta}_n = \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 = 2n^{-2}.$$

This example is remarkable, since the mean square error of the Pitman estimator is of order n^{-2} which is higher than that in the preceding example. This can be explained by the discontinuity of the density with respect to the parameter.

The optimal estimator of a location parameter in the class of linear unbiased estimators. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ where $\xi_1, \ldots, \xi_n$ are independent random variables with the distribution functions $F_1(x-\theta), \ldots, F_n(x-\theta)$, respectively. Assume that

$$\int x\, dF_j(x) = 0, \qquad j = 1, \ldots, n, \tag{3.2.17}$$

$$0 < \int x^2 dF_j(x) = \sigma_j^2 < \infty, \qquad j = 1, \ldots, n. \tag{3.2.18}$$

Thus θ is a location parameter and $\theta = \mathsf{E}_\theta \xi_i$, $i = 1, \dots, n$, by (3.2.17). Let L be the class of linear unbiased estimators l of the parameter θ that are of the form $l(x) = \sum_{i=1}^n c_i x_i$ where $\sum_{i=1}^n c_i = 1$, $x = (x_1, \dots, x_n)$. For any estimator $l \in L$ we have

$$\mathsf{E}_\theta (l - \theta)^2 = \mathsf{E}_\theta \left(\sum_{i=1}^n c_i (\xi_i - \theta) \right)^2 = \sum_{i=1}^n c_i^2 \sigma_i^2. \tag{3.2.19}$$

Solving the extremum problem for the function (3.2.19) subject to $\sum_{j=1}^n c_j = 1$ we find that

$$c_j = c_j^* = \sigma_j^{-2} \left(\sum_{i=1}^n \sigma_i^{-2} \right)^{-1}, \qquad j = 1, \dots, n, \tag{3.2.20}$$

at the point of extremum.

Thus we have proved the following result.

THEOREM 3.2.3. *Let conditions* (3.2.17) *and* (3.2.18) *hold. Then the optimal estimator of a location parameter in the class L of linear unbiased estimators is given by $l^*(x) = \sum_{j=1}^n c_j^* x_j$, $x = (x_1, \dots, x_n)$, where the coefficients c_j^*, $j = 1, \dots, n$, are defined by* (3.2.20).

The following result contains necessary and sufficient conditions that an estimator $l^*(x)$ is admissible in the class of unbiased estimators.

THEOREM 3.2.4. *Assume that $\xi_1, \dots, \xi_n$, $n \geq 3$, are independent observations with the distribution functions $F_1(x - \theta), \dots, F_n(x - \theta)$, respectively, for which conditions* (3.2.17) *and* (3.2.18) *hold. An optimal estimator of a parameter θ in the class of linear unbiased estimators $l^*(x) = \sum_{i=1}^n c_i^* x_i$ is admissible for the quadratic loss function in the class of all unbiased estimators of the parameter θ if and only if all the distribution functions $F_j(x)$ are normal.*

The proof of this theorem can be found in [**15**], Theorem 7.4.1. The book [**15**] contains further results on the estimation of a location parameter.

REMARK 3.2.5. If conditions (3.2.17) and (3.2.18) hold and moreover the variances σ_j^2 are equal to each other, that is, $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2$, then $c_j^* = n^{-1}$, $j = 1, \dots, n$, in (3.2.20). Thus the optimal estimator of the location parameter in the class L of linear unbiased estimators is of the form $l^*(x) = n^{-1} \sum_{i=1}^n x_i$ and this obviously is the sampling mean.

REMARK 3.2.6. If condition (3.2.18) holds, while condition (3.2.17) does not hold, then one can consider the class of linear unbiased estimators l of the form $l(x) = \sum_{j=1}^n c_j (x_j - b_j)$ where $\sum_{j=1}^n c_j = 1$ and $b_j = \int x \, dF_j(x)$, $j = 1, \dots, n$. As before we obtain that the optimal estimator is of the form $l^*(x) = \sum_{j=1}^n c_j^* (x_j - b_j)$ where the coefficients c_j^* are defined by (3.2.20).

REMARK 3.2.7. If the assumptions of Theorem 3.2.4 hold, then the optimal estimator θ in the class of linear unbiased estimators $l^*(x) = \sum_{j=1}^n c_j^* x_j$ is absolutely admissible with respect to the quadratic loss function if and only if all the distribution functions $F_j(x)$ are normal (see [**15**], Theorem 7.4.2). Note that if all the functions $F_j(x)$ are equal to each other, then the optimal estimator is of the form $l^*(x) = n^{-1} \sum_{i=1}^n x_i$ and therefore this estimator is absolutely admissible in the case of Gaussian distributions.

The problem of the confidence estimation of a location parameter. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be an observation where $\xi_1, \dots, \xi_n$ are independent identically distributed random variables with the distribution function $F(x-\theta)$, $\theta \in \mathbf{R}^1$. Consider the problem of the confidence estimation of the parameter θ with respect to the loss function

$$r(\widetilde{\theta}_n, \theta) = \begin{cases} 0, & |\widetilde{\theta}_n - \theta| \le b, \\ 1, & |\widetilde{\theta}_n - \theta| > b, \end{cases} \tag{3.2.21}$$

where b is some positive constant. The corresponding risk function of the estimator $\widetilde{\theta}_n$ is of the form

$$R(\widetilde{\theta}_n; \theta) = \mathsf{E}_\theta r(\widetilde{\theta}_n; \theta) = \mathsf{P}_\theta\{|\widetilde{\theta}_n - \theta| > b\} = \mathsf{P}_\theta\{\theta \notin [\widetilde{\theta}_n - b, \widetilde{\theta}_n + b]\}.$$

This risk function is the probability of the event that the confidence interval

$$[\widetilde{\theta}_n - b, \widetilde{\theta}_n + b]$$

does not contain the unknown value θ.

Given an arbitrary $F(x)$, the risk function

$$R\left(\overline{\xi}^{(n)}; \theta\right) = 1 - \mathsf{P}_\theta\left\{-b \le \overline{\xi}^{(n)} - \theta \le b\right\} = \gamma, \qquad \gamma = \gamma(b),$$

for the estimator $\overline{\xi}^{(n)} = n^{-1}\sum_{i=1}^n \xi_i$ does not depend on θ and therefore

$$\left[\overline{\xi}^{(n)} - b, \overline{\xi}^{(n)} + b\right]$$

is a confidence interval of level γ (depending, of course, on $F(x)$).

The following result claims that the sampling mean $\overline{\xi}^{(n)}$ is admissible with respect to the loss function (3.2.21) (in other words, it claims that the confidence interval $[\overline{\xi}^{(n)} - b, \overline{\xi}^{(n)} + b]$ is admissible).

THEOREM 3.2.5. *Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, $n \ge 3$, where $\xi_1, \dots, \xi_n$ are independent identically distributed random variables with the distribution function $F(x-\theta)$ whose density $f(x-\theta)$ is bounded. If, for a given sequence of numbers $b_j \to 0$, the sampling mean $\overline{\xi}^{(n)}$ is an admissible estimator of the parameter $\theta \in \mathbf{R}^1$ with respect to the loss function* (3.2.21) *for $b = b_j$, then $F(x)$ is the Gaussian distribution function.*

The proof of Theorem 3.2.5 can be found in [**31**] (see Theorem 7.9.3 therein).

REMARK 3.2.8. If random variables $\xi_1, \xi_2, \dots, \xi_n$ are independent and identically distributed according to the normal $\mathcal{N}(\theta, \sigma^2)$ law, then one can show that the sampling mean $\overline{\xi}^{(n)}$ is an admissible estimator of the location parameter θ with respect to the loss function (3.2.21) for all $b > 0$. Thus the converse statement to Theorem 3.2.5 is also true.

The minimax property of the Pitman estimator of a location parameter. As mentioned above, the Pitman estimator of a location parameter with respect to the quadratic loss function is a Bayes estimator of the location parameter with respect to the Lebesgue a priori measure $\mathbf{Q}$ if there exists the density of the observation. If all the assumptions of Theorem 3.2.1 hold, then the Pitman

estimator $\widehat{\theta}_n$ of the location parameter θ with respect to the quadratic loss function is of the form (3.2.9).

Let the a priori measure $\mathbf{Q}_N$ coincide with the uniform distribution on $[-N, N]$, so that the density is $q_N(t) = (2N)^{-1} I_{[-N,N]}(t)$. Then the Bayes estimator of the parameter θ with respect to the a priori measure $\mathbf{Q}_N$ and the quadratic loss function is given by

$$\theta^*_{Q_N}(x) = \frac{\int v f(x-v) q_N(v)\, dv}{\int f(x-v) q_N(v)\, dv} = \int_{-N}^{N} v f(x-v)\, dv \left(\int_{-N}^{N} f(x-v)\, dv \right)^{-1},$$

$x \in \mathbf{R}^n$, where $f(x)$ is the density mentioned in statement 2) of Theorem 3.2.1. It is clear that $\widehat{\theta}_n(x) = \lim_{N\to\infty} \theta^*_{Q_N}(x)$ for all $x \in \mathbf{R}^n$. One can show that

$$\lim_{N\to\infty} \mathsf{E}_\theta \left(\theta^*_{Q_N} - \theta\right)^2 = \mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 \tag{3.2.22}$$

for all $\theta \in [-N + \sqrt{N}, N - \sqrt{N}]$. Moreover the convergence is uniform in θ in the above interval. Since $\mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2$ does not depend on θ and convergence (3.2.22) is uniform in the interval $[-N + \sqrt{N}, N - \sqrt{N}]$, we get

$$\begin{aligned}
\limsup_{N\to\infty} \int \mathsf{E}_t \left(\theta^*_{Q_N} - t\right)^2 \mathbf{Q}_N(dt) &\geq \limsup_{N\to\infty} \frac{1}{2N} \int_{-N+\sqrt{N}}^{N-\sqrt{N}} \mathsf{E}_t \left(\theta^*_{Q_N} - t\right)^2 dt \\
&\geq \limsup_{N\to\infty} \frac{2(N-\sqrt{N})}{2N} \left(\mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 - \varepsilon \right) \\
&= \mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 - \varepsilon
\end{aligned}$$

for all $\varepsilon > 0$. This implies for all $\theta \in \mathbf{R}^1$ that

$$\mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 \leq \limsup_{N\to\infty} \int \mathsf{E}_t \left(\theta^*_{Q_N} - t\right)^2 \mathbf{Q}_N(dt).$$

By Theorem 3.1.4 this means that the Pitman estimator $\widehat{\theta}_n$ is minimax. Thus we proved the following result.

THEOREM 3.2.6. *If all the assumptions of Theorem* 3.2.1 *hold, then the Pitman estimator* $\widehat{\theta}_n$ *of a location parameter* θ *with respect to the quadratic loss function is minimax.*

3.3. Estimation of a scale parameter

Scale parameters. Equivariant estimators. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be an observed random element assuming values in the space $(\mathbf{R}^k, \mathcal{B}^k)$ and having the distribution belonging to a parametric set of measures $(\mathsf{P}_\sigma, \sigma \in (0, \infty))$ where σ is an unknown parameter. If the measure P_σ depends on the parameter σ such that

$$\mathsf{P}_\sigma(A) = \mathsf{P}_1(A/\sigma), \qquad A \in \mathcal{B}^k, \tag{3.3.1}$$

then σ is called a *scale parameter*. We put

$$A/\sigma = \left\{ x/\sigma = (x_1/\sigma, \ldots, x_n/\sigma) : x = (x_1, \ldots, x_n) \in A \right\}$$

for all $A \in \mathcal{B}^k$. Condition (3.3.1) can be rewritten as

$$\mathsf{P}_\sigma \left\{ \xi^{(n)} \in A \right\} = \mathsf{P}_\sigma \left\{ \xi^{(n)}/\sigma \in A/\sigma \right\} = \mathsf{P}_1 \left\{ \xi^{(n)} \in A/\sigma \right\}, \qquad A \in \mathcal{B}^k. \tag{3.3.2}$$

Distributions (3.3.1) arise in the case where observations are of the form $\xi_i = \sigma\varepsilon_i$, $i = 1, \dots, n$, for some $\sigma > 0$ and if the vector $(\varepsilon_1, \dots, \varepsilon_n)$ has the distribution defined by a measure P_1. Generally speaking, the random variables $\varepsilon_1, \dots, \varepsilon_n$ are dependent, thus the random variables $\xi_1, \dots, \xi_n$ are dependent, too.

It is natural to consider the following class of estimators in the case of the estimation of a scale parameter σ:

$$\Sigma = \{\widetilde{\sigma}_n(x): \sigma_n(\lambda x) = \lambda\widetilde{\sigma}_n(x) \text{ for all } \lambda > 0 \text{ and } x \in \mathbf{R}^n\} \tag{3.3.3}$$

where we put $\lambda x = (\lambda x_1, \dots, \lambda x_n)$ for all $x = (x_1, \dots, x_n)$. Estimators of the class Σ are called *equivariant estimators*. Consider a loss function $r(\widetilde{\sigma}_n; \sigma)$ such that

$$r(\widetilde{\sigma}_n; \sigma) = r(\widetilde{\sigma}_n - \sigma), \qquad r(\lambda u) = \lambda^m r(u) \tag{3.3.4}$$

for all $\lambda > 0$ and some $m > 0$. Using (3.3.2) and the definition (3.3.3) we get that the risk of any estimator $\widetilde{\sigma}_n \in \Sigma$ is such that

$$R(\widetilde{\sigma}_n; \sigma) = \mathsf{E}_\sigma r(\widetilde{\sigma}_n - \sigma) = \sigma^m \mathsf{E}_1 r(\widetilde{\sigma}_n - 1) = \sigma^m R(\widetilde{\sigma}_n; 1).$$

Therefore an estimator $\widetilde{\sigma}_n \in \Sigma$ is either optimal in the class Σ or is not admissible in this class provided the loss function satisfies condition (3.3.4).

An optimal estimator $\widehat{\sigma}_n \in \Sigma$ of a parameter σ in the class Σ, that is, the one such that

$$R(\widehat{\sigma}_n; \sigma) = \min_{\widetilde{\sigma}_n \in \Sigma} R(\widetilde{\sigma}_n; \sigma) \quad \text{for all } \sigma \in (0, \infty),$$

is called the *Pitman estimator of a scale parameter σ with respect to the loss function* (3.3.4).

In what follows we assume that an observed random element $\xi^{(n)}$ assumes values in the space

$$\mathbf{R}^n_{+,0} = \mathbf{R}^n_+ \setminus \{0\} = \{x = (x_1, \dots, x_n): x_i > 0 \text{ for all } i = 1, \dots, n\}.$$

Alternatively one can think that the sampling space is $\mathbf{R}^n$ and assume that

$$\mathsf{P}_\sigma\left(\mathbf{R}^n_{+,0}\right) = 1$$

for all $\sigma > 0$.

Equivariant estimators possess the following useful property. If

$$l(x) = \sum_{j=1}^n c_j x_j, \qquad x \in \mathbf{R}^n_{+,0},$$

is a linear statistic with $c_j > 0$ for all $j = 1, 2, \dots, n$, then

$$\widetilde{\sigma}_n(x) = l(x)\psi(y), \qquad x = (x_1, \dots, x_n) \in \mathbf{R}^n_{+,0}, \tag{3.3.5}$$

for all $\widetilde{\sigma}_n \in \Sigma$ where $\psi(y)$ is some measurable function of the vector

$$y = \left(\frac{x_2}{x_1}, \frac{x_3}{x_1}, \dots, \frac{x_n}{x_1}\right). \tag{3.3.6}$$

Using (3.3.3) we also get

$$\widetilde{\sigma}_n(x) = l(x)\widetilde{\sigma}_n(x/l(x)) = l(x)\widetilde{\sigma}_n\big((x/x_1)/l(x/x_1)\big),$$

since $l(x) > 0$, $x \in \mathbf{R}^n_{+,0}$, whence (3.3.5) follows for the function $\psi(y)$ specified above.

Pitman estimators. We use the following notation for an arbitrary statistic $T(x)$, $x \in \mathbf{R}^n$:

$$\mathsf{E}_1(T/y) = \mathsf{E}_1\left\{T\left(\xi^{(n)}\right) / \eta = y\right\} \tag{3.3.7}$$

where y is the vector defined by (3.3.6), while the vector η is given by

$$\eta = \left(\frac{\xi_2}{\xi_1}, \frac{\xi_3}{\xi_1}, \ldots, \frac{\xi_n}{\xi_1}\right). \tag{3.3.8}$$

The following result provides the explicit form of Pitman estimators of a scale parameter with respect to the quadratic loss function.

THEOREM 3.3.1. *Let $\mathsf{E}_1\xi_j^2 < \infty$ for all $j = 1, 2, \ldots, n$ and let $l(x) = \sum_{j=1}^n c_j x_j$, $x \in \mathbf{R}^n_{+,0}$, be some linear statistic with $c_j > 0$, $j = 1, 2, \ldots, n$. Then*

1) *the estimator*

$$\widehat{\sigma}_n(x) = l(x)\frac{\mathsf{E}_1(l/y)}{\mathsf{E}_1(l^2/y)}, \qquad x \in \mathbf{R}^n_{+,0}, \tag{3.3.9}$$

is the Pitman estimator of a scale parameter σ with respect to the quadratic loss function where y is the vector given by (3.3.6), *and $\mathsf{E}_1(l/y)$ and $\mathsf{E}_1(l^2/y)$ are conditional expectations defined by* (3.3.7) *and* (3.3.8);

2) *if the measure P_1 is absolutely continuous with respect to the Lebesgue measure and its density is $f(x)$, $x \in \mathbf{R}^n_{+,0}$, then the Pitman estimator is of the form*

$$\widehat{\sigma}_n(x) = \int_0^\infty u^n f(ux)\,du \left(\int_0^\infty u^{n+1} f(ux)\,du\right)^{-1}, \qquad x \in \mathbf{R}^n_{+,0}. \tag{3.3.10}$$

PROOF. Let $\widetilde{\sigma}_n$ be an arbitrary estimator of the class Σ. Using (3.3.5) we get

$$\mathsf{E}_\sigma(\widetilde{\sigma}_n - \sigma)^2 = \sigma^2\mathsf{E}_1(l\psi(\eta) - 1)^2 = \sigma^2\mathsf{E}_1\mathsf{E}_1\left((l\psi(\eta) - 1)^2/\eta\right).$$

If $\eta = y$ is fixed, then

$$\min_c \mathsf{E}_1\left((lc - 1)^2/y\right) = \mathsf{E}_1\left((lc^* - 1)^2/y\right)$$

where

$$c^* = \psi^*(y) = \frac{\mathsf{E}_1(l/y)}{\mathsf{E}_1(l^2/y)}.$$

Thus

$$\min_\psi \mathsf{E}_\sigma(l\psi(\eta) - \sigma)^2 = \mathsf{E}_\sigma(l\psi^*(\eta) - \sigma)^2 = \mathsf{E}_\sigma(\widehat{\sigma}_n - \sigma)^2$$

where $\widehat{\sigma}_n$ is the estimator defined by (3.3.9). Therefore estimator (3.3.9) is the Pitman estimator of the scale parameter σ with respect to the quadratic loss function.

Now we prove equality (3.3.10). Let $l(x) = n^{-1}\sum_{i=1}^n x_i = \overline{x}$. Then the Pitman estimator (3.3.9) can be rewritten as follows (here we put $\overline{\xi}^{(n)} = n^{-1}\sum_{i=1}^n \xi_i$):

$$\begin{aligned}\widehat{\sigma}_n(x) &= \overline{x}\frac{\mathsf{E}_1\left(\overline{\xi}^{(n)}/y\right)}{\mathsf{E}_1\left(\left(\overline{\xi}^{(n)}\right)^2/y\right)} = \overline{x}\frac{\mathsf{E}_1\left(\xi_1\left(\overline{\xi}^{(n)}/\xi_1\right)/y\right)}{\mathsf{E}_1\left(\xi_1^2\left(\overline{\xi}^{(n)}/\xi_1\right)^2/y\right)} \\ &= \overline{x}\frac{(\overline{x}/x_1)\mathsf{E}_1(\xi_1/y)}{(\overline{x}/x_1)^2\mathsf{E}_1\left(\xi_1^2/y\right)} = x_1\frac{\mathsf{E}_1(\xi_1/y)}{\mathsf{E}_1(\xi_1^2/y)}.\end{aligned} \tag{3.3.11}$$

Consider the random variables $\mathfrak{z}_1 = \xi_1, \mathfrak{z}_2 = \xi_2/\xi_1, \dots, \mathfrak{z}_n = \xi_n/\xi_1$ and denote by $p(z_1, z_2, \dots, z_n)$ the probability density of the vector $(\mathfrak{z}_1, \mathfrak{z}_2, \dots, \mathfrak{z}_n)$ for $\sigma = 1$. It is not hard to show that

$$p(z_1, z_2, \dots, z_n) = z_1^{n-1} f(z_1, z_1 z_2, \dots, z_1 z_n)$$

where $f(x)$ is the probability density of P_1 with respect to the Lebesgue measure. It is clear that

$$\begin{aligned}
&\mathsf{E}_1 (\xi_1^m / y) \\
&\quad = \mathsf{E}_1 \left\{ \mathfrak{z}_1^m \,/\, \mathfrak{z}_2 = \frac{x_2}{x_1}, \dots, \mathfrak{z}_n = \frac{x_n}{x_1} \right\} \\
&\quad = \int_0^\infty z^m p\left(z, \frac{x_2}{x_1}, \dots, \frac{x_n}{x_1}\right) dz \left(\int_0^\infty p\left(z, \frac{x_2}{x_1}, \dots, \frac{x_n}{x_1}\right) dz \right)^{-1} \\
&\quad = \int_0^\infty z^{m+n-1} f\left(z, z\frac{x_2}{x_1}, \dots, z\frac{x_n}{x_1}\right) dz \\
&\qquad \times \left(\int_0^\infty z^{n-1} f\left(z, z\frac{x_2}{x_1}, \dots, z\frac{x_n}{x_1}\right) dz \right)^{-1} \\
&\quad = x_1^m \int_0^\infty u^{m+n-1} f(x_1 u, \dots, x_n u)\, du \\
&\qquad \times \left(\int_0^\infty u^{n-1} f(x_1 u, \dots, x_n u)\, du \right)^{-1}.
\end{aligned} \tag{3.3.12}$$

Substituting (3.3.12) into (3.3.11) we obtain (3.3.10). □

Let Σ_u be the class of unbiased equivariant estimators of a scale parameter σ. It is clear that $\Sigma_u \subset \Sigma$. The following result provides the explicit form of an optimal estimator in the class Σ_u with respect to the quadratic loss function.

THEOREM 3.3.2. *Let all the assumptions of Theorem* 3.3.1 *hold. An optimal in the class Σ_u estimator of a scale parameter σ with respect to the quadratic loss function is of the form*

$$\widehat{\sigma}_{n,u} = C_u \widehat{\sigma}_n \tag{3.3.13}$$

where the constant C_u is such that $C_u \mathsf{E}_1 \widehat{\sigma}_n = 1$.

PROOF. Let $\widetilde{\sigma}_n$ be an arbitrary estimator of the class Σ_u. If $\mathsf{E}_1 \widetilde{\sigma}_n^2 = \infty$, then $\mathsf{E}_\sigma \widetilde{\sigma}_n^2 = \infty$ for all $\sigma > 0$. Thus the estimator $\widetilde{\sigma}_n$ is worse than $\widehat{\sigma}_{n,u}$. If $\mathsf{E}_1 \widetilde{\sigma}_n^2 < \infty$, then

$$\begin{aligned}
\mathsf{E}_\sigma (\widetilde{\sigma}_n - \sigma)^2 &= \sigma^2 \mathsf{E}_1 (\widetilde{\sigma}_n - 1)^2 = \sigma^2 \mathsf{E}_1 (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u} + \widehat{\sigma}_{n,u} - 1)^2 \\
&= \sigma^2 \big\{ \mathsf{E}_1 (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u})^2 \\
&\qquad + 2\mathsf{E}_1 (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u})(\widehat{\sigma}_{n,u} - 1) + \mathsf{E}_1 (\widehat{\sigma}_{n,u} - 1)^2 \big\}.
\end{aligned} \tag{3.3.14}$$

Since $\widetilde{\sigma}_n \in \Sigma_u$ and $\widehat{\sigma}_{n,u} \in \Sigma_u$, we have $\mathsf{E}_1 (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u}) = 0$. Since $\widehat{\sigma}_n$ is optimal in the class Σ, we obtain $\mathsf{E}_1 \widehat{\sigma}_n h = 0$ for an arbitrary unbiased estimator of zero $h \in \Sigma$ such that $\mathsf{E}_1 h^2 < \infty$ (the proof is the same as that in Theorem 3.1.1). In particular, $\mathsf{E}_1 \widehat{\sigma}_{n,u} (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u}) = 0$. Thus (3.3.14) implies

$$\mathsf{E}_\sigma (\widetilde{\sigma}_n - \sigma)^2 = \mathsf{E}_\sigma (\widetilde{\sigma}_n - \widehat{\sigma}_{n,u})^2 + \mathsf{E}_\sigma (\widehat{\sigma}_{n,u} - \sigma)^2 \geq \mathsf{E}_\sigma (\widehat{\sigma}_{n,u} - \sigma)^2. \qquad \square$$

Below is an example of the evaluation of a Pitman estimator.

EXAMPLE 3.3.1. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ where $\xi_1, \dots, \xi_n$ are independent identically distributed exponential random variables. Then their common probability density is $\sigma^{-1}\exp(-x/\sigma)$, $x > 0$, $\sigma > 0$. Assumptions of assertion 2) in Theorem 3.3.1 are satisfied and

$$f(x) = \exp\left(-\sum_{i=1}^{n} x_i\right) I_{[0,\infty)}\left(\min_{1\le i\le n} x_i\right), \qquad x = (x_1, \dots, x_n).$$

Substituting this density into (3.3.10), the Pitman estimator of the parameter σ becomes of the form

$$\widehat{\sigma}_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} x_i, \qquad x = (x_1, \dots, x_n).$$

Since $\mathsf{E}_1\widehat{\sigma}_n = n/(n+1)$, Theorem 3.3.2 implies that the optimal estimator of the parameter σ in the class of the unbiased equivariant estimator Σ_u is of the form

$$\widehat{\sigma}_{n,u}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad x = (x_1, \dots, x_n).$$

Straightforward calculations yield

$$\mathsf{E}_\sigma(\widehat{\sigma}_n - \sigma)^2 = \frac{\sigma^2}{n+1} < \mathsf{E}_\sigma(\widehat{\sigma}_{n,u} - \sigma)^2 = \frac{\sigma^2}{n},$$

that is, the estimator $\widehat{\sigma}_n$ is better than $\widehat{\sigma}_{n,u}$, and this result is natural.

The optimal estimator of a scale parameter in the class of linear unbiased estimators. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ where $\xi_1, \dots, \xi_n$ are nondegenerate independent random variables with the distribution functions $F_1(x/\sigma), \dots, F_n(x/\sigma)$, respectively. The distribution functions depend on a scale parameter $\sigma \in (0, \infty)$ and are such that

$$F_j(0+) = 0, \qquad \mathsf{E}_1\xi_j^2 < \infty, \qquad j = 1, \dots, n. \tag{3.3.15}$$

Let L be the class of linear unbiased estimators of the parameter σ of the form $l(x) = \sum_{i=1}^{n} c_i x_i$, $x = (x_1, \dots, x_n)$, where $c_i > 0$ for all $i = 1, \dots, n$. Put

$$\alpha_{1j} = \mathsf{E}_1\xi_j, \qquad \alpha_{2j} = \mathsf{E}_1\xi_j^2, \qquad \sigma_j^2 = \alpha_{2j} - \alpha_{1j}^2.$$

Since ξ_j are nondegenerate, it follows that $\sigma_j^2 > 0$. If $l(x) = \sum_{j=1}^{n} c_j x_j$, then $\sum_{j=1}^{n} c_j\alpha_{1j} = 1$. Further $\mathsf{E}_\sigma(l-\sigma)^2 = \sigma^2\sum_{j=1}^{n} c_j^2\sigma_j^2$. This implies that the optimal estimator of the parameter σ in the class L with respect to the quadratic loss function is of the form $l^*(x) = \sum_{j=1}^{n} c_j^* x_j$ where the coefficients c_j^* are such that

$$c_j^* = \frac{\alpha_{1j}}{\sigma_j^2}\left(\sum_{i=1}^{n}\frac{\alpha_{1i}}{\sigma_i^2}\right)^{-1}, \qquad j = 1, \dots, n. \tag{3.3.16}$$

Therefore we have proved the following result.

THEOREM 3.3.3. *Let random variables $\xi_1, \ldots, \xi_n$ be independent and nondegenerate with the distribution functions $F_1(x/\sigma), \ldots, F_n(x/\sigma)$, respectively. If condition* (3.3.15) *holds, then the optimal linear unbiased estimator of the parameter σ with respect to the quadratic loss function is of the form $l^*(x) = \sum_{j=1}^n c_j^* x_j$ where the coefficients c_j^* are defined by* (3.3.16).

The following result provides necessary and sufficient conditions for an optimal linear unbiased estimator $l^*(x) = \sum_{j=1}^n c_j^* x_j$ of a scale parameter σ to be admissible.

THEOREM 3.3.4. *Let all the assumptions of Theorem* 3.3.3 *hold. Then an optimal linear unbiased estimator $l^*(x) = \sum_{i=1}^n c_i^* x_i$ of the parameter σ is admissible in the class of unbiased estimators with respect to the quadratic loss function for some two different values of n, say $n = n_1$ and $n = n_2$, $n_2 > n_1 \geq 3$, if and only if the random variables ξ_j have the Gamma distribution*

$$F_j(x) = \frac{\alpha_j^{\gamma_j}}{\Gamma(\gamma_j)} \int_0^x t^{\gamma_j - 1} e^{-\alpha_j t}\, dt, \qquad x > 0,$$

for some $\gamma_j > 0$ and $\alpha_j > 0$, $j = 1, \ldots, n$.

The proof of Theorem 3.3.4 can be found in [**15**] (see Theorem 7.12.2 therein).

REMARK 3.3.1. If an estimator $l^*(x)$ is admissible in the class of unbiased estimators of the scale parameter σ, then by Theorem 3.3.2 it is optimal in the class Σ_u. Moreover by Theorem 3.3.4 the distribution of the random variables ξ_j is the Gamma distribution in this case (see Example 3.3.1).

REMARK 3.3.2. Further results concerning the estimation of a scale parameter can be found in [**15**].

3.4. The Cramér–Rao inequality and efficient estimators

In the preceding sections we studied the quality of statistical estimators of unknown parameters and obtained several qualitative results. Moreover we introduced two classes of parameters, namely the classes of scale and location parameters, for which one can construct optimal estimators in appropriate classes of estimators. In this section we use a somewhat different approach to construct optimal estimators. We also obtain the minimal mean square error of the estimation that can be achieved in an experiment.

Regularity conditions for families of distributions. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$ and whose distribution belongs to a parametric set $\{\mathsf{P}_\theta, \theta \in \Theta\}$ where Θ is a subset of $\mathbf{R}^k$, $k \geq 1$. Throughout this section we assume that for all $\theta \in \Theta$ the measure P_θ is absolutely continuous with respect to some σ-finite measure μ on $(X, \mathcal{B})$, that is, $\mathsf{P}_\theta \ll \mu$, and that $f(x;\theta)$ is the density of the measure P_θ with respect to the measure μ. In particular, if $(X, \mathcal{B}) = (\mathbf{R}^m, \mathcal{B}^m)$ for some $m \geq 1$, then as the measure μ one can take the Lebesgue measure.

We consider the case of a one-dimensional parameter θ in this section, that is, we consider the case $k = 1$.

Below we use the following set of *regularity conditions*, called (CR):

(i) Θ is a finite or infinite interval in $\mathbf{R}^1$;

(ii) the derivative $\partial f(x;\theta)/\partial\theta$ exists and is finite P_θ-almost everywhere for all $\theta \in \Theta$;

(iii) $|\partial^i f(x;\theta)/\partial\theta^i| \le g_i(x)$ for all $\theta \in \Theta$ and $i = 1, 2$ where $g_1(x)$ and $g_2(x)$ are nonnegative Borel functions such that $\int g_i(x)\mu(dx) << \infty$, $i = 1, 2$;

(iv) $0 < \mathsf{E}_\theta(\partial \ln f(\xi;\theta)/\partial\theta)^2 < \infty$ for all $\theta \in \Theta$.

The regularity conditions (CR) are also called the *Cramér–Rao regularity conditions*. If conditions (CR) hold, then the family of distributions $\{\mathsf{P}_\theta,\ \theta \in \Theta\}$ is called *CR-regular*.

Let $S(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$, $x \in X$, and put

$$I(\theta) = \mathsf{E}_\theta S^2(\xi;\theta), \qquad \theta \in \Theta. \tag{3.4.1}$$

The function $I(\theta)$ is called the *Fisher information*. This function is treated as the amount of information about the parameter θ contained in the observation ξ. The notion of the information $I(\theta)$ will become clear after the proof of the Cramér–Rao inequality. Note that condition (iv) above can be rewritten as follows: $0 < I(\theta) < \infty$ for all $\theta \in \Theta$.

First we prove an auxiliary result.

LEMMA 3.4.1. *If regularity conditions* (CR) *hold, then*

$$\mathsf{E}_\theta S(\xi;\theta) = 0 \quad \textit{for all } \theta \in \Theta. \tag{3.4.2}$$

PROOF. First we differentiate the identity

$$\int f(x;\theta)\,\mu(dx) = 1 \quad \text{for all } \theta \in \Theta$$

with respect to θ and obtain

$$\frac{\partial}{\partial\theta}\int f(x;\theta)\,\mu(dx) = 0 \quad \text{for all } \theta \in \Theta. \tag{3.4.3}$$

Conditions (ii) and (iii) allow one to interchange the differentiation and integration in (3.4.3). Thus

$$\int \frac{\partial}{\partial\theta} f(x;\theta)\,\mu(dx) = 0 \quad \text{for all } \theta \in \Theta, \tag{3.4.3$'$}$$

whence

$$\mathsf{E}_\theta S(\xi;\theta) = \int \frac{\partial}{\partial\theta} f(x;\theta)\,\mu(dx) = 0 \quad \text{for all } \theta \in \Theta. \qquad \square$$

The following result gives another representation for the Fisher information.

LEMMA 3.4.2. *If regularity conditions* (CR) *hold, then*

$$I(\theta) = -\mathsf{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f(\xi;\theta)\right) \tag{3.4.4}$$

for all $\theta \in \Theta$.

PROOF. It is obvious that

$$\mathsf{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f(\xi;\theta)\right) = \mathsf{E}_\theta\left(\frac{\partial^2 f(\xi;\theta)/\partial\theta^2}{f(\xi;\theta)} - \left(\frac{\partial f(\xi;\theta)/\partial\theta}{f(\xi;\theta)}\right)^2\right)$$
$$= \int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,\mu(dx) - I(\theta). \tag{3.4.5}$$

It follows from (iv) that $I(\theta) < \infty$ for all $\theta \in \Theta$. Condition (iii) implies that

$$\frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,\mu(dx) = \int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,\mu(dx) = 0 \tag{3.4.6}$$

for all $\theta \in \Theta$. Combining (3.4.5) and (3.4.6) we obtain (3.4.4). □

Sometimes we also consider the following set of regularity conditions, called (R):

(i′) Θ is a finite or infinite interval in $\mathbf{R}^1$;

(ii′) the function $(f(x;\theta))^{1/2}$ is continuously differentiable with respect to θ for μ-almost all x;

(iii′) $0 < \mathsf{E}_\theta S^2(\xi;\theta) = I(\theta) < \infty$ for all $\theta \in \Theta$ and the function $I(\theta)$ is continuous with respect to θ.

In what follows we need an assertion on the continuity of integrals of functions depending on a parameter.

Let $\psi(t,y)$, $t \in \Theta$, be a family of measurable functions defined on a measurable space $(Y, \mathcal{B}_Y)$ equipped with a measure ν. We consider some conditions under which

$$\int \psi(t,y)\,\nu(dy) \to \int \psi(\theta,y)\,\nu(dy), \qquad t \to \theta. \tag{3.4.7}$$

Let $A(t) = A(t,\theta)$, $t \in \Theta$, be a family of sets belonging to $\mathcal{B}_Y$. Put

$$\overline{A}(t) = Y \setminus A(t).$$

The following result is a generalization of a well-known Lebesgue theorem.

LEMMA 3.4.3. *Let $A(t)$, $t \in \Theta$, be a family such that*

1) $\psi(t,y)I_{A(t)}(y) \to \psi(\theta,y)$ *as* $t \to \theta$ *for* ν*-almost all* y *for which* $\psi(\theta,y) \neq 0$;
2) $\sup_t |\psi(t,y)I_{A(t)}(y)| \le \psi(y)$ *where* $\psi(y)$ *is a function integrable with respect to the measure* ν, *that is,* $\int \psi(y)\,\nu(dy) < \infty$.

Then relation (3.4.7) *holds if and only if*

$$\int \psi(t,y)I_{\overline{A}(t)}(y)\,\nu(dy) \to 0, \qquad t \to \theta. \tag{3.4.8}$$

PROOF. By the Lebesgue theorem

$$\int \psi(t,y)I_{A(t)}(y)\,\nu(dy) \to \int \psi(\theta,y)\,\nu(dy), \qquad t \to \theta.$$

Since

$$\int \psi(t,y)\,\nu(dy) = \int \psi(t,y)I_{A(t)}(y)\,\nu(dy) + \int \psi(t,y)I_{\overline{A}(t)}(y)\,\nu(dy),$$

relation (3.4.7) is equivalent to relation (3.4.8). □

COROLLARY 3.4.1. *Let $T(x)$ be a real measurable bounded function, $T: X \to \mathbf{R}^1$, and let $f(x;\theta)$ be continuous with respect to θ for μ-almost all $x \in X$. Then the function $\mathsf{E}_\theta T(\xi)$ is continuous with respect to θ.*

PROOF. We apply Lemma 3.4.3 for $Y = X$, $\nu = \mu$, $\psi(t,x) = T(x)f(x;t)$, and $A(t) = \{x: f(x;t) \le 2f(x;\theta)\}$. It is obvious that all the assumptions of Lemma 3.4.3 are satisfied. Since $T(x) \equiv 1$ and thus $\mathsf{E}_\theta T(\xi) \equiv 1$ is continuous, Lemma 3.4.3 yields

$$\int f(x;t) I_{\overline{A(t)}}\, \mu(dx) \to 0$$

as $t \to \theta$. This together with Lemma 3.4.3 implies that $\mathsf{E}_\theta T(\xi)$ is continuous for any bounded function $T(x)$. □

REMARK 3.4.1. If one seeks a simple sufficient condition for (3.4.7) in the case of $\psi(t,y) \to \psi(\theta;y)$ as $t \to \theta$ and ν-almost surely, then an appropriate candidate is the uniform convergence of integrals in (3.4.7). The latter condition can be reformulated as follows: there exists a finite measure λ such that the inequality $\lambda(A) < \delta = \delta(\varepsilon)$ implies $\sup_t \int_A |\psi(t,y)|\, \nu(dy) < \varepsilon$ for a given $\varepsilon > 0$. Moreover if the function $\psi(y) = \sup_t |\psi(t,y)|$ is integrable, then one can take $\lambda(A) = \int_A \psi(y)\, \nu(dy)$.

Below we consider some corollaries of conditions (R) that we will use in the proof of the Cramér–Rao inequality.

LEMMA 3.4.4. *Let conditions (R) hold. Let $T = T(\xi)$ be an arbitrary real statistic such that $\mathsf{E}_\theta T^2 < c < \infty$ for all $\theta \in \Theta$. Then the function $a(\theta) = \mathsf{E}_\theta T$ is differentiable with respect to θ, and moreover*

$$a'(\theta) = \mathsf{E}_\theta T(\xi) S(\xi;\theta) = \int T(x) \frac{\partial}{\partial\theta} f(x;\theta)\, \mu(dx) \tag{3.4.9}$$

where $S(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$, $x \in X$, and the function $a'(\theta)$ is continuous.

PROOF. In the same manner as in the proof of Lemma 3.4.1 we derive from condition (ii)′ that

$$\mathsf{E}_\theta S(\xi;\theta) = \int \frac{\partial}{\partial\theta} f(x;\theta)\, \mu(dx) = 0, \qquad \theta \in \Theta. \tag{3.4.10}$$

Note that (3.4.10) also follows from (3.4.9) for $T(x) \equiv 1$. Then

$$\mathsf{D}_\theta S(\xi;\theta) = \mathsf{E}_\theta S^2(\xi;\theta) = I(\theta) = 4 \int \left(\frac{\partial}{\partial\theta} \sqrt{f(x;\theta)} \right)^2 \mu(dx). \tag{3.4.11}$$

The function $I(\theta)$ is continuous by conditions (R). We apply Lemma 3.4.3 for $Y = X$, $\nu = \mu$, and $\psi(t,x) = (\partial\sqrt{f(x;t)}/\partial\theta)^2$:

$$A(t) = A_1(\delta) = \left\{ x : \sup_{|v-\theta|<|\delta|} \sqrt{f(x;v)} < 2\sqrt{f(x;\theta)}, \right. \tag{3.4.12}$$
$$\left. \sup_{|v-\theta|<|\delta|} \left| \frac{\partial}{\partial v} \sqrt{f(x;v)} \right| \le 2 \left| \frac{\partial}{\partial\theta} \sqrt{f(x;\theta)} \right| \right\}$$

where $\delta = t - \theta$. Assumptions of Lemma 3.4.3 hold for $\psi(x) = 4\psi(\theta; x)$, since the functions $\sqrt{f(x;\theta)}$ and $\partial\sqrt{f(x;\theta)}/\partial\theta$ are continuous. Since $I(t)$ converges to $I(\theta)$ as $t \to \theta$, we prove from (3.4.8) that

$$\varepsilon(t) = \int \left(\frac{\partial}{\partial\theta}\sqrt{f(x;t)}\right)^2 I_{\overline{A}_1(\delta)}(x)\,\mu(dx) \to 0 \tag{3.4.13}$$

as $t \to \theta$.

Similarly to the proof of Corollary 3.4.1 we prove that $\int T(x)\partial f(x;\theta)/\partial\theta\,\mu(dx)$ is a continuous function. In the proof of this result we apply Lemma 3.4.3 for $Y = X$, $\nu = \mu$, $\psi(t,x) = T(x)\partial f(x;t)/\partial t$, and $A(t) = A_1(\delta)$. Since

$$\partial f(x,\theta)/\partial\theta = 2\sqrt{f(x,\theta)}\,\partial\sqrt{f(x,\theta)}/\partial\theta,$$

we get $\sup_t |\psi(t,x)|I_{A(t)}(x) \le \psi(x) = 4|\psi(\theta;x)|$. The Cauchy–Bunyakovskiĭ inequality implies that

$$\int_{\overline{A}_1(\delta)} |\psi(\theta,x)|\,\mu(dx) \le 2\left(\mathsf{E}_\theta T^2(\xi)\int_{\overline{A}_1(\delta)} (\partial\sqrt{f(x,\theta)}/\partial\theta)^2\,\mu(dx)\right)^{1/2}$$

This together with (3.4.13) implies relation (3.4.8), whence it follows that the function $\int T(x)\partial f(x;\theta)/\partial\theta\,\mu(dx)$ is continuous.

Now we turn to the proof of equality (3.4.9). Note that

$$\begin{aligned}
&\frac{1}{\delta}\left(\int T(x)f(x;\theta+\delta)\,\mu(dx) - \int T(x)f(x;\theta)\,\mu(dx)\right)\\
&\quad = \int\int_0^1 T(x)\frac{\partial}{\partial\theta}f(x;\theta+u\delta)\,du\,\mu(dx)\\
&\quad = \int\int_0^1 2T(x)\sqrt{f(x;\theta+u\delta)}\,\frac{\partial}{\partial\theta}\sqrt{f(x;\theta+u\delta)}\,du\,\mu(dx)
\end{aligned}$$

by condition (ii)′. We apply Lemma 3.4.3 again for $Y = R \times X$, $y = (u,x)$, $\nu = \lambda \times \mu$, $\psi(\delta,y) = T(x)\partial f(x;\theta+u\delta)/\partial\theta$, and $A(\delta) = A_1(\delta)$, where λ is the Lebesgue measure, $A_1(\delta)$ is defined by (3.4.12), and $\delta \to 0$. Since the functions $\sqrt{f(x;\theta)}$ and $\partial\sqrt{f(x;\theta)}/\partial\theta$ are continuous with respect to θ, assumptions 1) and 2) of Lemma 3.4.3 hold, whence

$$\begin{aligned}
&\psi(\delta,y)I_{A(\delta)}(x) \to T(x)\frac{\partial}{\partial\theta}f(x;\theta) = \psi(0;y), \qquad \delta \to 0,\\
&\sup_\delta |\psi(\delta,y)I_{A(\delta)}(x)| \le 4T(x)\left|\frac{\partial}{\partial\theta}f(x;\theta)\right|
\end{aligned}$$

and by the Cauchy–Bunyakovskiĭ inequality

$$\int T(x)\left|\frac{\partial}{\partial\theta}f(x;\theta)\right|\mu(dx) \le \left(\mathsf{E}_\theta T^2(\xi)\cdot I(\theta)\right)^{1/2} < \infty.$$

Using the Cauchy–Bunyakovskiĭ inequality again we obtain from relation (3.4.13) that

$$\begin{aligned}\left|\int_{\overline{A}_1(\delta)}\int_0^1 T(x)\sqrt{f(x;\theta+u\delta)}\,\frac{\partial}{\partial\theta}\sqrt{f(x;\theta+u\delta)}\,du\,\mu(dx)\right| \\ \le \left[\int_{\overline{A}_1(\delta)}\int_0^1 T^2(x)f(x;\theta+u\delta)\,du\,\mu(dx)\right. \\ \left.\times\int_{\overline{A}_1(\delta)}\int_0^1\left(\frac{\partial}{\partial\theta}\sqrt{f(x;\theta+u\delta)}\right)^2 du\,\mu(dx)\right]^{1/2} \\ \le \left(c\int_0^1\varepsilon(\theta+u\delta)\,du\right)^{1/2}\to 0\end{aligned}$$

as $\delta\to 0$, whence relation (3.4.8) follows. Thus we proved that the derivative $a'(\theta)$ exists and equality (3.4.9) holds. □

LEMMA 3.4.5. *If the set Θ is compact and the function $\sqrt{f(x;\theta)}$ is continuously differentiable with respect to θ for μ-almost all x, then $I(\theta)$ is continuous if and only if*

$$\lim_{N\to\infty}\sup_\theta \mathsf{E}_\theta S^2(\xi;\theta)I(|S(\xi;\theta)|>N)=0. \tag{3.4.14}$$

PROOF. Let the function $I(\theta)$ be continuous but let relation (3.4.14) not hold. Then there is a $\gamma>0$ and a sequence $t\to\theta\in\Theta$ such that $N_t\to\infty$ and

$$m(t)=\mathsf{E}_t S^2(\xi;t)I(|S(\xi;t)|>N_t)>\gamma \tag{3.4.15}$$

for all t belonging to this sequence.

Applying Lemma 3.4.3 for $Y=X$ and $\mu=\nu$ we get

$$\psi(t,x)=\left(\frac{\partial}{\partial t}\sqrt{f(x;t)}\right)^2=4^{-1}S^2(x;t)f(x;t),$$
$$A(t)=\left\{x:\left|\frac{\partial}{\partial t}\sqrt{f(x;t)}\right|\le 2\left|\frac{\partial}{\partial\theta}\sqrt{f(x;\theta)}\right|\right\}.$$

Since $\partial\sqrt{f(x;\theta)}/\partial\theta$ is continuous, assumptions 1) and 2) of Lemma 3.4.3 hold and the continuity of $I(t)$ implies that

$$m_1(t)=\int_{\overline{A}(t)}\left|\frac{\partial}{\partial t}\sqrt{f(x;t)}\right|^2\mu(dx)\to 0$$

as $t\to\theta$. Note that $m(t)\le m_1(t)+m_2(t)$ where

$$m_2(t)=\int_{B(t)\cap A(t)}\left(\frac{\partial}{\partial t}\sqrt{f(x;t)}\right)^2\mu(dx),$$
$$B(t)=\left\{x:2\left|\frac{\partial}{\partial t}\sqrt{f(x;t)}\right|>N_t\sqrt{f(x;t)}\right\}.$$

It follows from the definition of the set $A(t)$ that

$$m_2(t) \le 4 \int_{B(t)} \left| \frac{\partial}{\partial\theta} \sqrt{f(x;\theta)} \right|^2 \mu(dx).$$

Since $\partial\sqrt{f(x;t)}/\partial t \to \partial\sqrt{f(x;\theta)}/\partial\theta$ and $\sqrt{f(x;t)} \to \sqrt{f(x;\theta)}$ as $t \to \theta$, we prove that $B(t)$ converges to a set whose measure μ is zero. This means that $\mu(B(t)) \to 0$, $m_2(t) \to 0$, and $m(t) \to 0$ as $t \to \theta$. This contradicts (3.4.15) and thus relation (3.4.14) holds.

Now we prove the converse statement. Let condition (3.4.14) hold. According to Lemma 3.4.3, $I(t)$ is continuous if $m_1(t) \to 0$ as $t \to \theta$ on the set $A(t)$ defined as above. Further

$$m_1(t) \le \int_{|S(x;t)|>N} |S(x;t)|^2 f(x;t)\,\mu(dx) + N^2 \int_{\overline{A}(t)} f(x;t)\,\mu(dx)$$

where the first integral is small by (3.4.14) if N is sufficiently large. To estimate the second integral we note that $\mu(\overline{A}(t)) \to 0$ as $t \to \theta$ and

$$\int_{\overline{C}(t)} f(x;t)\,\mu(dx) \to 0, \qquad t \to \theta,$$

for $C(t) = \{x: f(x;t) \le 2f(x;\theta)\}$ (see the proof of Corollary 3.4.1). Thus as $t \to \theta$

$$\int_{\overline{A}(t)} f(x;t)\,\mu(dx) \le 2 \int_{\overline{A}(t)} f(x;\theta)\,\mu(dx) + \int_{\overline{C}(t)} f(x;t)\,\mu(dx). \qquad \square$$

REMARK 3.4.2. If the set Θ is compact and conditions (R) are satisfied, then, due to Lemma 3.4.5, the Fisher information $I(\theta)$ is continuous if and only if condition (3.4.14) holds. It is natural to call the latter condition the uniform convergence condition for the integral $I(\theta) = \mathsf{E}_\theta S^2(\xi;\theta)$.

The Cramér–Rao inequality under regularity conditions (CR). The following result contains the lower bound for the variance of an unbiased estimator under the Cramér–Rao conditions (CR).

THEOREM 3.4.1. *Let the distributions P_θ, $\theta \in \Theta$, satisfy the regularity conditions (CR). Assume that $g(\theta)$ is a differentiable real function, $\widehat{g} = \widehat{g}(x)$ is an unbiased estimator of the function $g(\theta)$ such that the variance of $\widehat{g}$ exists, and*

$$\int \left| \widehat{g}(x) \frac{\partial}{\partial\theta} f(x;\theta) \right| \mu(dx) < \infty \quad \text{for all } \theta \in \Theta.$$

Then

$$\mathsf{D}_\theta \widehat{g}(\xi) \ge \frac{(g'(\theta))^2}{I(\theta)} \quad \text{for all } \theta \in \Theta. \tag{3.4.16}$$

Inequality (3.4.16) *becomes an equality if and only if the density $f(x;\theta)$ is of the form*

$$f(x;\theta) = \exp\{\psi_1(\theta)\widehat{g}(x) + \psi_2(\theta) + h(x)\}, \qquad x \in X, \tag{3.4.17}$$

where $\psi_1'(\theta) \ne 0$.

PROOF. According to the Cauchy–Bunyakovskiĭ inequality

$$\mathsf{E}_\theta S(\xi;\theta)(\widehat{g}(\xi)-g(\theta)) \le (I(\theta)\mathsf{D}_\theta\widehat{g}(\xi))^{1/2}, \qquad \theta\in\Theta. \tag{3.4.18}$$

Lemma 3.4.1 implies that

$$\mathsf{E}_\theta S(\xi;\theta)(\widehat{g}(\xi)-g(\theta)) = \mathsf{E}_\theta S(\xi;\theta)\widehat{g}(\xi) = \int \widehat{g}(x)\frac{\partial}{\partial\theta}f(x;\theta)\,\mu(dx). \tag{3.4.19}$$

Since $\widehat{g}$ is an unbiased estimator, we have

$$g(\theta+\delta)-g(\theta) = \int \widehat{g}(x)\big(f(x;\theta+\delta)-f(x;\theta)\big)\,\mu(dx),$$

whence it follows by the regularity condition (iii) that

$$g'(\theta) = \int \widehat{g}(x)\frac{\partial}{\partial\theta}f(x;\theta)\,\mu(dx). \tag{3.4.20}$$

Thus (3.4.18)–(3.4.20) imply inequality (3.4.16).

It remains to consider the case of equality in (3.4.16). The inequality in (3.4.16) becomes an equality if and only if inequality (3.4.18) becomes an equality. In its turn (3.4.18) becomes an equality if and only if

$$S(x;\theta) = \widehat{g}(x)\psi(\theta) + \widetilde{\psi}(\theta)$$

for all $x\in X$ where $\psi(\theta)\ne 0$. This implies that

$$\ln f(x;\theta) = \widehat{g}(x)\psi_1(\theta) + \psi_2(\theta) + h(x)$$

for all $x\in X$ where $\psi_1'(\theta)\ne 0$. □

Results of the type (3.4.16) are called the *Cramér–Rao inequality*. This inequality gives a lower bound for the variance of an unbiased estimator of the function $g(\theta)$. If $\widehat{\theta}$ is an unbiased estimator of a parameter θ, then the Cramér–Rao inequality becomes of the form

$$\mathsf{D}_\theta\widehat{\theta} \ge \frac{1}{I(\theta)}, \qquad \theta\in\Theta. \tag{3.4.21}$$

Inequality (3.4.16) also gives a lower bound for the variance of estimators $\widehat{\theta}$ of a parameter θ that are not necessarily unbiased. Indeed let $g(\theta)=\theta+b(\theta)$ where $b(\theta)$ is the bias of the estimator $\widehat{\theta}$, that is, $b(\theta)=\mathsf{E}_\theta(\widehat{\theta}-\theta)$. Assuming that the function $b(\theta)$ is differentiable, we obtain from (3.4.16) that

$$\mathsf{E}_\theta(\widehat{\theta}-\theta)^2 \ge b^2(\theta) + \frac{(1+b'(\theta))^2}{I(\theta)}. \tag{3.4.22}$$

Let an observation be a sample $\xi^{(n)}=(\xi_1,\xi_2,\dots,\xi_n)$ and let the random variables $\xi_1,\xi_2,\dots,\xi_n$ have the density $f(x;\theta)$, $x\in\mathbf{R}^1$, where θ is a real unknown parameter, $\theta\in\Theta\subset\mathbf{R}^1$. Assume that the density $f(x;\theta)$ satisfies the Cramér–Rao conditions (CR). Denote by $f_n(x;\theta)$, $x\in\mathbf{R}^n$, the density of the vector $\xi^{(n)}$. Let $I_n(\theta)$ be the Fisher information evaluated with respect to the density $f_n(x;\theta)$, while $I(\theta)=I_1(\theta)$ is the Fisher information evaluated with respect to the density $f(x;\theta)$, that is,

$$I_n(\theta) = \mathsf{E}_\theta S_n^2\left(\xi^{(n)};\theta\right) \quad\text{and}\quad I(\theta)=\mathsf{E}_\theta S^2(\xi_1;\theta)$$

where

$$S_n(x;\theta) = \frac{\partial \ln f_n(x;\theta)}{\partial\theta} \quad \text{and} \quad S(x;\theta) = \frac{\partial \ln f(x;\theta)}{\partial\theta}.$$

LEMMA 3.4.6. *If the Cramér–Rao regularity conditions (CR) hold, then*

$$I_n(\theta) = nI(\theta). \tag{3.4.23}$$

PROOF. Since $f_n(x;\theta) = \prod_{k=1}^n f(x_k;\theta)$, $x = (x_1,\dots,x_n)$, and $\mathsf{E}_\theta S(\xi_1;\theta) = 0$ by Lemma 3.4.1, we have

$$\begin{aligned} I_n(\theta) &= \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln f_n\left(\xi^{(n)};\theta\right)\right)^2 \\ &= \sum_{k=1}^n \mathsf{E}_\theta \left(\frac{\partial \ln f(\xi_k;\theta)}{\partial\theta}\right)^2 + \sum_{k\neq j} \mathsf{E}_\theta \frac{\partial \ln f(\xi_k;\theta)}{\partial\theta} \mathsf{E}_\theta \frac{\partial \ln f(\xi_j;\theta)}{\partial\theta} \\ &= nI(\theta). \end{aligned}$$

□

COROLLARY 3.4.2. *Assume that all the assumptions of Theorem* 3.4.1 *hold. If* $\widehat{g}_n(x)$ *is an unbiased estimator of a function* $g(\theta)$*, then the Cramér–Rao inequality holds*:

$$\mathsf{D}_\theta \widehat{g}_n(\xi) \geq \frac{(g'(\theta))^2}{nI(\theta)}, \qquad \theta \in \Theta. \tag{3.4.24}$$

In particular, if $g(\theta) = \theta$ *and* $\widehat{\theta}_n$ *is an unbiased estimator of a parameter* θ*, then*

$$\mathsf{D}_\theta \widehat{\theta}_n \geq \frac{1}{nI(\theta)}, \qquad \theta \in \Theta. \tag{3.4.25}$$

The proof of Corollary 3.4.2 follows from Theorem 3.4.1 and equality (3.4.23).□

Not all of the regularity conditions (CR) are used in the proof of the Cramér–Rao inequality. In fact, this result holds under a weaker set of regularity conditions called in what follows the Cramér–Rao conditions $(CR)^*$:

(i) Θ is a finite or infinite interval in $\mathbf{R}^1$;

(ii)* the function $f(x;\theta)$ is differentiable with respect to θ for μ-almost all $x \in X$ and

$$0 < I(\theta) = \int \left(\frac{\partial}{\partial\theta} f(x;\theta)\right)^2 (f(x;\theta))^{-1}\, \mu(dx) < \infty;$$

(iii)* the following relations are satisfied:

$$\begin{aligned} &\lim_{\Delta\to 0} \frac{1}{\Delta^2} \int \frac{(f(x;\theta+\Delta) - f(x;\theta))^2}{f(x;\theta)}\, \mu(dx) \\ &\qquad = \int \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2 \frac{\mu(dx)}{f(x;\theta)}, \end{aligned} \tag{3.4.26}$$

$$\int \frac{\partial f(x;\theta)}{\partial\theta}\, \mu(dx) = \frac{\partial}{\partial\theta} \int f(x;\theta)\, \mu(dx). \tag{3.4.27}$$

In fact, conditions (3.4.26) and (3.4.27) are necessary to justify the interchange of integration and differentiation. It is easily seen that the regularity conditions $(CR)^*$ are less restrictive than conditions (CR). Nevertheless the Cramér–Rao inequality can also be proved under conditions $(CR)^*$.

THEOREM 3.4.2. *If regularity conditions $(CR)^*$ hold, then*

$$\mathsf{D}_\theta\widehat{\theta} \geq 1/I(\theta), \qquad \theta \in \Theta, \tag{3.4.28}$$

for any unbiased estimator $\widehat{\theta} = \widehat{\theta}(\xi)$ of a parameter θ.

PROOF. Since $\widehat{\theta}$ is an unbiased estimator, we have

$$\Delta = \int (\widehat{\theta}(x) - \theta)\big(f(x;\theta+\Delta) - f(x;\theta)\big)\,\mu(dx),$$

whence

$$\begin{aligned}\Delta^2 &= \left(\int (\widehat{\theta}(x) - \theta)\sqrt{f(x;\theta)}\,\frac{f(x;\theta+\Delta) - f(x;\theta)}{\sqrt{f(x;\theta)}}\,\mu(dx)\right)^2 \\ &\leq \mathsf{D}_\theta\widehat{\theta}\int \frac{(f(x;\theta+\Delta) - f(x;\theta))^2}{f(x;\theta)}\,\mu(dx)\end{aligned}$$

by the the Cauchy–Bunyakovskiĭ inequality. Thus

$$\mathsf{D}_\theta\widehat{\theta} \geq \left(\frac{1}{\Delta^2}\int \frac{(f(x;\theta+\Delta) - f(x;\theta))^2}{f(x;\theta)}\,\mu(dx)\right)^{-1}$$

for all Δ. Passing to the limit as $\Delta \to 0$ we obtain inequality (3.4.28) in view of condition (3.4.26). □

Let an observation be a sample $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ with the density

$$f_n(x;\theta) = \prod_{i=1}^n f(x_i;\theta), \qquad x = (x_1, \ldots, x_n) \in \mathbf{R}^n,\ t \in \Theta.$$

If the density $f(x;\theta)$ satisfies conditions (ii)* and (3.4.27), while the density $f_n(x;\theta)$ satisfies condition (3.4.26), then the Cramér–Rao inequality (3.4.25) holds for any unbiased estimator $\widehat{\theta}_n$ of the parameter θ. To prove this result we use equality (3.4.23) that follows from condition (3.4.27).

The Cramér–Rao inequality under the regularity conditions (R). The following result contains a lower bound for the variance of a biased, generally speaking, estimator of a parameter θ.

THEOREM 3.4.3. *Let regularity conditions (R) hold. Let $\widehat{\theta}$ be an estimator of a parameter θ such that $\mathsf{E}_\theta\widehat{\theta}^2 \leq c < \infty$ for all $\theta \in \Theta$. Then*

$$\mathsf{D}_\theta\widehat{\theta} \geq \frac{(1+b'(\theta))^2}{I(\theta)}, \qquad \theta \in \Theta, \tag{3.4.29}$$

where $b(\theta) = \mathsf{E}_\theta\widehat{\theta} - \theta$ is the bias of the estimator $\widehat{\theta}$.

If (3.4.9) *becomes an equality on some interval* $[\theta_1, \theta_2] \subset \Theta$ *and* $\mathsf{D}_\theta \widehat{\theta} > 0$ *on this interval, then*

$$f(x;\theta) = \exp\{A(\theta)\widehat{\theta}(x) + B(\theta)\}h(x), \qquad x \in X, \tag{3.4.30}$$

for $\theta \in [\theta_1, \theta_2]$ *where* $A(\theta)$ *and* $B(\theta)$ *do not depend on* x.

Conversely if either $\widehat{\theta}(x) \equiv \text{const}$ *or representation* (3.4.30) *holds, then inequality* (3.4.29) *becomes an equality.*

PROOF. Let $a(\theta) = \mathsf{E}_\theta \widehat{\theta}$. Putting $T(x) \equiv 1$ in Lemma 3.4.4 we get from (3.4.9) that

$$\mathsf{E}_\theta S(\xi;\theta) = 0, \qquad \mathsf{E}_\theta a(\theta) S(\xi;\theta) = 0. \tag{3.4.31}$$

Again using Lemma 3.4.4 for $T(x) = \widehat{\theta}(x)$ we obtain from (3.4.9) and (3.4.31) that

$$\mathsf{E}_\theta \widehat{\theta}(\xi) S(\xi;\theta) = a'(\theta), \qquad \mathsf{E}_\theta (\widehat{\theta}(\xi) - a(\theta)) S(\xi;\theta) = a'(\theta). \tag{3.4.32}$$

Using the Cauchy–Bunyakovskiĭ inequality and the second equality in (3.4.32) we get

$$(a'(\theta))^2 \le \mathsf{E}_\theta (\widehat{\theta}(\xi) - a(\theta))^2 \mathsf{E}_\theta S^2(\xi;\theta) \tag{3.4.33}$$

or, equivalently,

$$\mathsf{D}_\theta \widehat{\theta} \ge \frac{(a'(\theta))^2}{\mathsf{E}_\theta S^2(\xi;\theta)}, \qquad \theta \in \Theta. \tag{3.4.34}$$

Since $a(\theta) = \theta + b(\theta)$ and $\mathsf{E}_\theta S^2(\xi;\theta) = I(\theta)$, we obtain inequality (3.4.29) from (3.4.34).

Now we prove the second statement of Theorem 3.4.3. For the sake of simplicity we assume that Θ coincides with an interval $[\theta_1, \theta_2]$ and that the measure μ is concentrated on a union of supports of measures P_θ, $\theta \in \Theta$. The equality in (3.4.29) (or, equivalently, in (3.4.33)) means that

$$\begin{aligned} &\int (\widehat{\theta}(x) - a(\theta)) \frac{\partial f(x;\theta)}{\partial \theta} \mu(dx) \\ &\quad = \left(\int (\widehat{\theta}(x) - a(\theta))^2 f(x;\theta)\, \mu(dx) \int \frac{(\partial f(x;\theta)/\partial\theta)^2}{f(x;\theta)} \mu(dx) \right)^{1/2} \end{aligned} \tag{3.4.35}$$

for all $\theta \in \Theta$. Since the first integral on the right-hand side of (3.4.35) is positive by condition, an equality in (3.4.35) is only possible if

$$\frac{\partial f(x;\theta)/\partial\theta}{\sqrt{f(x;\theta)}} = c(\theta)(\widehat{\theta}(x) - a(\theta))\sqrt{f(x;\theta)} \qquad (\mu\text{-a.s.}). \tag{3.4.36}$$

Let A be the set of $x \in X$ for which (3.4.36) holds and $|\widehat{\theta}(x)| < \infty$. Then $\mu(\overline{A}) = 0$ (here $\overline{A} = X \setminus A$ is the complement of the set A). Fix $x \in A$. Since $f(x;\theta)$ is continuous with respect to θ, we have $f(x;t) > 0$ on some interval $(t_1, t_2) \subset \Theta$, and moreover

$$S(x;t) = c(t)(\widehat{\theta}(x) - a(t)), \qquad t \in (t_1, t_2), \tag{3.4.37}$$

on this interval by (3.4.36). If (3.4.29) becomes an equality, then (3.4.32) and (3.4.37) imply that

$$a'(\theta) = \mathsf{E}_\theta(\widehat{\theta}(\xi) - a(\theta))S(\xi;\theta) = c(\theta)\mathsf{D}_\theta\widehat{\theta},$$
$$\mathsf{D}_\theta\widehat{\theta} = \frac{(a'(\theta))^2}{I(\theta)}, \qquad |c(\theta)| = \left(\frac{I(\theta)}{\mathsf{D}_\theta\widehat{\theta}}\right)^{1/2}$$

This implies that the variance $\mathsf{D}_\theta\widehat{\theta}$ is continuous with respect to θ as well as $a'(\theta)$, and $I(\theta)$ is continuous with respect to θ while the functions $|c(\theta)|$ and $a(\theta)$ are uniformly bounded on $[\theta_1, \theta_2]$. The derivative $S(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$ in equality (3.4.37) possesses the same property. This means that the function $\ln f(x;\theta)$ is finite and $f(x;\theta) > 0$ for all $\theta \in \Theta = [\theta_1, \theta_2]$, whence (3.4.37) follows for all θ. Integrating equality (3.4.37) with respect to t from θ_1 to θ, we obtain

$$\ln f(x;\theta) = \widehat{\theta}(x)\int_{\theta_1}^{\theta} c(t)\,dt - \int_{\theta_1}^{\theta} c(t)a(t)\,dt + \ln f(x;\theta_1)$$

and this is equivalent to (3.4.30) for μ-almost all x. Since the values of $f(x;\theta)$ on a set whose measure μ is zero do not matter, representation (3.4.30) is proved.

Finally, we prove the latter statement of Theorem 3.4.3. If $\widehat{\theta}(x) \equiv \text{const}$, then $b'(\theta) = -1$ and both sides of equality (3.4.29) vanish. Now let representation (3.4.30) hold. Differentiating the function $\ln f(x;\theta)$ with respect to θ we get

$$S(x;\theta) = \widehat{\theta}(x)A'(\theta) + B'(\theta).$$

The first equality in (3.4.31) implies that

$$a(\theta)A'(\theta) + B'(\theta) = 0.$$

Thus

$$S(x;\theta) = A'(\theta)(\widehat{\theta}(x) - a(\theta))$$

and inequality (3.4.29) becomes an equality in view of (3.4.36). □

REMARK 3.4.3. If $\mathsf{E}_\theta\widehat{\theta}^2 = \infty$, then $\mathsf{D}_\theta\widehat{\theta} = \infty$ and inequality (3.4.29) is trivial. In view of (3.4.29), the condition $\mathsf{D}_\theta\widehat{\theta} > 0$ can be substituted by $1 + b'(\theta) \neq 0$.

COROLLARY 3.4.3. *If all the assumptions of Theorem* 3.4.3 *are satisfied, then*

$$\mathsf{E}_\theta(\widehat{\theta} - \theta)^2 \geq \frac{(1 + b'(\theta))^2}{I(\theta)} + b^2(\theta), \qquad \theta \in \Theta.$$

For every unbiased estimator $\widehat{\theta}$ of the parameter θ

$$\mathsf{E}_\theta(\widehat{\theta} - \theta)^2 \geq \frac{1}{I(\theta)}, \qquad \theta \in \Theta.$$

Analogs of Theorem 3.4.3 can be proved under other sets of conditions. Below is a set of conditions, called $(R)^*$, which also is sufficient for the the Cramér–Rao inequality:

(i) Θ is a finite or infinite interval in $\mathbf{R}^1$;

(ii)″ the function $\sqrt{f(x;\theta)}$ is absolutely continuous with respect to θ for μ-almost all $x \in X$;

(iii)′ $0 < \mathsf{E}_\theta S^2(\xi;\theta) = I(\theta) < \infty$ for all $\theta \in \Theta$ and the function $I(\theta)$ is continuous with respect to θ.

It is obvious that conditions $(R)^*$ are weaker than conditions (R). Nevertheless the Cramér–Rao inequality holds under conditions $(R)^*$, too.

THEOREM 3.4.4. *Let the regularity conditions $(R)^*$ hold. Let $\widehat{\theta}$ be an unbiased estimator of a parameter θ. Then*

$$\mathsf{E}_\theta(\widehat{\theta}-\theta)^2 \geq \frac{1}{I(\theta)} \tag{3.4.38}$$

for all points $\theta \in \Theta$ of continuity of $\mathsf{E}_\theta(\widehat{\theta}-\theta)^2$.

PROOF. Since $\widehat{\theta}$ is unbiased,

$$\Delta = \int (\widehat{\theta}(x)-\theta)\left(\sqrt{f(x;\theta+\Delta)}+\sqrt{f(x;\theta)}\right)\left(\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)}\right)\mu(dx).$$

Applying the Cauchy–Bunyakovskiĭ inequality and then the elementary inequality $(\sqrt{a}+\sqrt{b})^2 \leq 2(a+b)$ for $a>0$ and $b>0$, we get

$$\begin{aligned}\Delta^2 &\leq \int (\widehat{\theta}(x)-\theta)^2\left(\sqrt{f(x;\theta+\Delta)}+\sqrt{f(x;\theta)}\right)^2\mu(dx)\\ &\qquad\times\int(\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)})^2\,\mu(dx)\\ &\leq 2\int(\widehat{\theta}(x)-\theta)^2(f(x;\theta+\Delta)+f(x;\theta))\,\mu(dx)\\ &\qquad\times\int\left(\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)}\right)^2\mu(dx).\end{aligned} \tag{3.4.39}$$

Then

$$\int(\widehat{\theta}(x)-\theta)^2 f(x;\theta)\,\mu(dx) = \mathsf{E}_\theta(\widehat{\theta}-\theta)^2, \tag{3.4.40}$$

$$\int(\widehat{\theta}(x)-\theta)^2 f(x;\theta+\Delta)\,\mu(dx) = \mathsf{E}_{\theta+\Delta}(\widehat{\theta}-\theta-\Delta)^2+\Delta^2. \tag{3.4.41}$$

It follows from (3.4.39)–(3.4.41) that

$$\begin{aligned}&\frac{\mathsf{E}_\theta(\widehat{\theta}-\theta)^2+\mathsf{E}_{\theta+\Delta}(\widehat{\theta}-\theta-\Delta)^2+\Delta^2}{2}\\ &\qquad\geq\left(\frac{4}{\Delta^2}\int\left(\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)}\right)^2\mu(dx)\right)^{-1}\end{aligned} \tag{3.4.42}$$

Condition (ii)″ implies that

$$\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)} = \int_\theta^{\theta+\Delta}\frac{\partial}{\partial u}\sqrt{f(x;u)}\,du.$$

Applying the Cauchy–Bunyakovskiĭ inequality and Fubini theorem we obtain

$$
\begin{aligned}
&\frac{4}{\Delta^2}\int\left(\sqrt{f(x;\theta+\Delta)}-\sqrt{f(x;\theta)}\right)^2\mu(dx)\\
&\quad=\frac{4}{\Delta^2}\int\left(\int_\theta^{\theta+\Delta}\frac{\partial\sqrt{f(x;u)}}{\partial u}\,du\right)^2\mu(dx)\\
&\quad\le\frac{4}{\Delta}\int\int_\theta^{\theta+\Delta}\left(\frac{\partial\sqrt{f(x;u)}}{\partial u}\right)^2 du\,\mu(dx) \qquad (3.4.43)\\
&\quad=\frac{1}{\Delta}\int_\theta^{\theta+\Delta}\left(4\int\left(\frac{\partial\sqrt{f(x;u)}}{\partial u}\right)^2\mu(dx)\right)du\\
&\quad=\frac{1}{\Delta}\int_\theta^{\theta+\Delta}I(u)\,du.
\end{aligned}
$$

Substituting (3.4.43) into (3.4.42) and passing to the limit as $\Delta\to 0$, we complete the proof of (3.4.38) in view of the continuity of the function $I(u)$. □

The above proof of inequality (3.4.42) does not require any condition posed on the density $f(x;\theta)$ or on an unbiased estimator $\widehat{\theta}$. Thus this proof can be used to obtain lower bounds of the variance of an estimator $\widehat{\theta}$ even in the case where regularity conditions are not satisfied for $f(x;\theta)$. Below we provide a result of this kind for biased estimators $\widehat{\theta}$.

THEOREM 3.4.5. *Let $\theta\in\Theta$ and $\theta+\Delta\in\Theta$ for some $\Delta\neq 0$. Then for all estimators $\widehat{\theta}$ of a parameter θ one has*

$$
\mathsf{D}_\theta\widehat{\theta}\ge(\Delta a(\theta))^2\left(\int\frac{(\Delta f(x;\theta))^2}{f(x;\theta)}\,\mu(dx)\right)^{-1} \qquad (3.4.44)
$$

where

$$
a(\theta)=\mathsf{E}_\theta\widehat{\theta},\qquad \Delta a(\theta)=a(\theta+\Delta)-a(\theta),
$$
$$
\Delta f(x;\theta)=f(x;\theta+\Delta)-f(x;\theta).
$$

In particular,

$$
\mathsf{D}_\theta\widehat{\theta}\ge\left(\frac{1}{\Delta^2}\int\frac{(\Delta f(x;\theta))^2}{f(x;\theta)}\,\mu(dx)\right)^{-1} \qquad (3.4.45)
$$

if the estimator $\widehat{\theta}$ is unbiased.

PROOF. First let the measure $\mathsf{P}_{\theta+\Delta}$ be not absolutely continuous with respect to the measure P_θ. Denote by N_θ the support of the measure P_θ in X and let $N_\theta=\{x\colon f(x;\theta)\neq 0\}$. Then there is a set $A\subset N_{\theta+\Delta}$ such that

$$
\mathsf{P}_{\theta+\Delta}(A)>0
$$

and $f(x;\theta)=0$ for all $x\in A$. Thus the integral in (3.4.44) is infinite and inequality (3.4.44) is trivial.

Now let the measure $\mathsf{P}_{\theta+\Delta}$ be absolutely continuous with respect to the measure P_θ. Then $N_{\theta+\Delta} \subset N_\theta$. Since $f(x;\theta)$ and $f(x;\theta+\Delta)$ are the densities of measures P_θ and $\mathsf{P}_{\theta+\Delta}$, respectively, with respect to μ, we have

$$\int \Delta f(x;\theta)\,\mu(dx) = 0.$$

Moreover

$$\int \widehat{\theta}(x)\Delta f(x;\theta)\,\mu(dx) = \Delta a(\theta).$$

This implies that

$$\int_{N_\theta} (\widehat{\theta}(x) - a(\theta))\Delta f(x;\theta)\,\mu(dx) = \Delta a(\theta). \tag{3.4.46}$$

The integrand in (3.4.46) can be represented on the set N_θ as

$$(\widehat{\theta}(x) - a(\theta))\sqrt{f(x;\theta)} \cdot \frac{\Delta f(x;\theta)}{\sqrt{f(x;\theta)}}.$$

Applying the Cauchy–Bunyakovskiĭ inequality we obtain

$$(\Delta a(\theta))^2 \le \int_{N_\theta} (\widehat{\theta}(x) - a(\theta))^2 f(x;\theta)\,\mu(dx) \cdot \int_{N_\theta} \frac{(\Delta f(x;\theta))^2}{f(x;\theta)}\,\mu(dx),$$

whence inequality (3.4.44) follows. Inequality (3.4.45) follows from (3.4.44). □

COROLLARY 3.4.4. *Assume that an arbitrary set of regularity conditions holds and*

$$\lim_{\Delta\to 0} \frac{1}{\Delta^2} \int \frac{(\Delta f(x;\theta))^2}{f(x;\theta)}\,\mu(dx) = I(\theta).$$

Then

$$\mathsf{D}_\theta \widehat{\theta} \ge (a'_+(\theta))^2 / I(\theta) \tag{3.4.47}$$

for all estimators $\widehat{\theta}$ of the parameter θ where

$$a'_+(\theta) = \limsup_{\Delta\to 0} \frac{\Delta a(\theta)}{\Delta}.$$

PROOF. To prove (3.4.47) we pass to the limit in (3.4.44) along a subsequence $\Delta \to 0$ such that $\Delta a(\theta)/\Delta \to a'_+(\theta)$. □

Inequality (3.4.44) is called the *Chapman–Robbins inequality*. Another name for it is the *difference Cramér–Rao type inequality*.

REMARK 3.4.4. If an observation ξ is a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, $f(x;\theta)$ is the density of ξ_1, and $f_n(x;\theta)$ is the density of the vector $\xi^{(n)}$, then one can obtain analogs of all the above results. If conditions (R) or $(R)^*$ hold for the density $f(x;\theta)$, then all the above inequalities hold for estimators $\widehat{\theta}_n$ of the parameter θ. The only exception is that the Fisher information $nI(\theta)$ substitutes the Fisher information $I(\theta)$ where $I(\theta)$ is evaluated with respect to the density $f(x;\theta)$. In particular, inequality (3.4.29) becomes in this case of the form

$$\mathsf{D}_\theta \widehat{\theta}_n \ge \frac{(1 + b'_n(\theta))^2}{nI(\theta)} \tag{3.4.48}$$

where $b_n(\theta) = \mathsf{E}_\theta \widehat{\theta}_n - \theta$ is the bias of the estimator $\widehat{\theta}_n$. Further, inequality (3.4.44) can be rewritten in this case as

$$\mathsf{D}_\theta \widehat{\theta}_n \geq (\Delta a_n(\theta))^2 \left(\int \frac{(\Delta f_n(x;\theta))^2}{f_n(x;\theta)} \mu(dx) \right)^{-1} \tag{3.4.49}$$

where $a_n(\theta) = \mathsf{E}_\theta \widehat{\theta}_n$.

Efficient and asymptotically efficient estimators of parameters. Let one of the sets of regularity conditions (CR), $(CR)^*$, (R), or $(R)^*$ hold. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$. Let the distribution of ξ belong to a family of probability measures

$$\{\mathsf{P}_\theta, \theta \in \Theta\}, \qquad \Theta \subset \mathbf{R}^1.$$

Let K_b^g be the class of estimators $\widehat{g} = \widehat{g}(\xi)$ of a function $g(\theta)$ with a bias $b(\theta)$. Let K_b be the class of estimators $\widehat{\theta} = \widehat{\theta}(\xi)$ of a parameter θ with a bias $b(\theta)$, that is,

$$K_b^g = \{\widehat{g} \colon \mathsf{E}_\theta \widehat{g} = g(\theta) + b(\theta)\}, \qquad K_b = \{\widehat{\theta} \colon \mathsf{E}_\theta \widehat{\theta} = \theta + b(\theta)\}.$$

We also consider the class K^g of unbiased estimators $\widehat{g}$ of a function $g(\theta)$ and the class K of unbiased estimators $\widehat{\theta}$ of a parameter θ, that is,

$$K^g = K_0^g = \{\widehat{g} \colon \mathsf{E}_\theta \widehat{g} = g(\theta)\}, \qquad K = K_0 = \{\widehat{\theta} \colon \mathsf{E}_\theta \widehat{\theta} = \theta\}.$$

Note that $K^g = U_g$ and $K = U_0$ where U_g and U_0 are the classes of estimators introduced in Section 3.1.

We say that $g^* \in K_b^g$ is an *efficient estimator of a function $g(\theta)$ in the class K_b^g* if the Cramér–Rao inequality for it becomes an equality, that is,

$$\mathsf{D}_\theta g^* = \frac{(g'(\theta) + b'(\theta))^2}{I(\theta)}, \qquad \theta \in \Theta. \tag{3.4.50}$$

Similarly, $g^* \in K^g$ is called an *efficient estimator of a function $g(\theta)$ in the class K^g* (or, an *efficient estimator of a function $g(\theta)$*) if

$$\mathsf{D}_\theta g^* = \frac{(g'(\theta))^2}{I(\theta)}, \qquad \theta \in \Theta. \tag{3.4.51}$$

Conditions (3.4.50) and (3.4.51) can be rewritten in the following equivalent form:

$$\mathsf{E}_\theta (g^* - g(\theta))^2 = \frac{(g'(\theta) + b'(\theta))^2}{I(\theta)} + b^2(\theta), \qquad \theta \in \Theta, \tag{3.4.52}$$

$$\mathsf{E}_\theta (g^* - g(\theta))^2 = \frac{(g'(\theta))^2}{I(\theta)}, \qquad \theta \in \Theta. \tag{3.4.53}$$

Efficient estimators θ^* of a parameter θ in the classes K_b and K can be introduced similarly to (3.4.50)–(3.4.53) if $g(\theta) = \theta$. In particular, $\theta^* \in K_b$ is called an *efficient estimator of a parameter θ in the class K_b* if the corresponding Cramér–Rao inequality becomes an equality, that is,

$$\mathsf{E}_\theta (\theta^* - \theta)^2 = \frac{(1 + b'(\theta))^2}{I(\theta)} + b^2(\theta), \qquad \theta \in \Theta. \tag{3.4.54}$$

We say that $\theta^* \in K$ is an *efficient estimator of a parameter θ in the class K* if

$$\mathsf{E}_\theta(\theta^* - \theta)^2 = \frac{1}{I(\theta)}, \qquad \theta \in \Theta. \tag{3.4.55}$$

Conditions (3.4.50)–(3.4.55) suggest a general definition: an *estimator is called efficient in the corresponding class if the Cramér–Rao inequality becomes an equality*.

Efficient estimators exist only in exceptional cases. In other cases one can construct the so-called asymptotically efficient estimators if the size of a sample increases.

Let an observation be a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ with the density

$$f_n(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \qquad x = (x_1, \dots, x_n),\ \theta \in \Theta.$$

A sequence of estimators g_n^*, $n = 1, 2, \dots$, is called an *asymptotically efficient estimator of a function $g(\theta)$* if

$$\mathsf{E}_\theta(g_n^* - g(\theta))^2 = \frac{(g'(\theta))^2}{nI(\theta)} + o\left(\frac{1}{n}\right), \qquad \theta \in \Theta,$$

as $n \to \infty$. For the sake of brevity we say that g_n^* is an asymptotically efficient estimator of a function $g(\theta)$. Similarly, θ_n^* is called an *asymptotically efficient estimator of a parameter θ* if

$$\mathsf{E}_\theta(\theta_n^* - \theta)^2 = \frac{1}{nI(\theta)} + o\left(\frac{1}{n}\right), \qquad \theta \in \Theta,$$

as $n \to \infty$.

Another name for asymptotically efficient estimators is *asymptotically efficient estimators in the strong sense*. If a sequence of estimators g_n^*, $n = 1, 2, \dots$, is asymptotically $\mathcal{N}\big(g(\theta), (g'(\theta))^2/(nI(\theta))\big)$ normal, then we say that g_n^* is an *asymptotically efficient estimator of a function $g(\theta)$ in the weak sense*. If a sequence of estimators θ_n^*, $n = 1, 2, \dots$, is asymptotically $\mathcal{N}\big(\theta, 1/(nI(\theta))\big)$ normal, then the estimator θ_n^* is called an *asymptotically efficient estimator of a parameter θ in the weak sense*.

EXAMPLE 3.4.1. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the normal distribution $\mathcal{N}(\theta, \sigma^2)$. In this case

$$I(\theta) = \mathsf{E}_\theta S^2(\xi_1;\theta) = \mathsf{E}_\theta\left(\frac{\xi_1 - \theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^2}.$$

Then $\widehat{\theta}_n = n^{-1}\sum_{i=1}^{n} \xi_i$ is an unbiased estimator of the parameter θ and

$$\mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 = \frac{\sigma^2}{n} = \frac{1}{nI(\theta)}.$$

Thus $\widehat{\theta}_n$ is an efficient estimator of the parameter θ.

EXAMPLE 3.4.2. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the $\mathcal{N}(\alpha, \theta)$ distribution. It is clear that

$$I(\theta) = \mathsf{E}_\theta S^2(\xi_1; \theta) = \mathsf{E}_\theta \left(\frac{(\xi_1 - \alpha)^2}{2\theta^2} - \frac{1}{2\theta} \right)^2$$
$$= \frac{1}{4\theta^4} \mathsf{E}_\theta (\xi_1 - \alpha)^4 - \frac{1}{2\theta^3} \mathsf{E}_\theta (\xi_1 - \alpha)^2 + \frac{1}{4\theta^2} = \frac{1}{2\theta^2},$$

since $\mathsf{E}_\theta (\xi_1 - \alpha)^4 = 3\theta^2$. It is clear that $\widehat{\theta}_n = n^{-1} \sum_{i=1}^n (\xi_i - \alpha)^2$ is an unbiased estimator of the parameter θ and

$$\mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 = \mathsf{E}_\theta \left(\frac{1}{n} \sum_{i=1}^n \left[(\xi_i - \alpha)^2 - \theta \right] \right)^2$$
$$= \frac{1}{n^2} \sum_{i=1}^n \mathsf{E}_\theta \left[(\xi_i - \alpha)^2 - \theta \right]^2$$
$$+ \frac{1}{n^2} \sum_{j \neq i} \mathsf{E}_\theta \left[(\xi_i - \alpha)^2 - \theta \right] \mathsf{E}_\theta \left[(\xi_j - \alpha)^2 - \theta \right]$$
$$= \frac{1}{n} \mathsf{E}_\theta \left[(\xi_1 - \alpha)^2 - \theta \right]^2 = \frac{1}{n} \left[\mathsf{E}_\theta (\xi_1 - \alpha)^4 - 2\theta \mathsf{E}_\theta (\xi_1 - \alpha)^2 + \theta^2 \right]$$
$$= \frac{2\theta^2}{n} = \frac{1}{nI(\theta)},$$

whence it follows that $\widehat{\theta}_n$ is an efficient estimator of the parameter θ.

Consider the estimator $\theta_n = (n-1)^{-1} \sum_{i=1}^n (\xi_i - \overline{\xi})^2$, where $\overline{\xi} = n^{-1} \sum_{i=1}^n \xi_i$. It is clear that $\mathsf{E}_\theta \widetilde{\theta}_n = \theta$, that is, $\widetilde{\theta}_n$ is an unbiased estimator of the parameter θ (see Example 3.1.3). According to (3.2.9) we have

$$\mathsf{E}_\theta (\widetilde{\theta}_n - \theta)^2 = \frac{2\theta^2}{n-1} = \frac{1}{(n-1)I(\theta)} > \frac{1}{nI(\theta)},$$

that is, $\widetilde{\theta}_n$ is not an efficient estimator of the parameter θ. Nevertheless $\widetilde{\theta}_n$ is an asymptotically efficient estimator of the parameter in the strong sense.

EXAMPLE 3.4.3. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the Gamma distribution with the density

$$f(x; \theta) = \frac{1}{\Gamma(\theta)} x^{\theta - 1} e^{-x} I_{(0,\infty)}(x)$$

where $\theta \in \Theta = (0, \infty)$ and $\Gamma(\theta)$ is the Gamma function. It is obvious that regularity conditions hold in this case. By Lemma 3.4.2 we have

$$I(\theta) = -\mathsf{E}_\theta \left(\frac{\partial^2 \ln f(\xi_1; \theta)}{\partial \theta^2} \right) = \frac{d^2 \ln \Gamma(\theta)}{d\theta^2}. \tag{3.4.56}$$

Consider the estimator $\widehat{\theta}_n = n^{-1} \sum_{i=1}^n \xi_i$. It is clear that $\mathsf{E}_\theta \widehat{\theta}_n = \theta$, that is, $\widehat{\theta}_n$ is an unbiased estimator of the parameter θ. Moreover,

$$\mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 = \frac{\theta}{n} = \frac{1}{nI(\theta)} \cdot \kappa(\widehat{\theta}_n; \theta) \tag{3.4.57}$$

by (3.4.56) where

$$\kappa(\widehat{\theta}_n;\theta) = \theta\,\frac{d^2\ln\Gamma(\theta)}{d\theta^2}. \tag{3.4.58}$$

By the Stirling formula (see [**9**], relation (12.5.3)) we have

$$\ln\Gamma(\theta) = \left(\theta-\frac{1}{2}\right)\ln\theta - \theta + \frac{1}{2}\ln 2\pi + \int_0^\infty \frac{P_1(x)}{\theta+x}\,dx$$

where $P_1(x)$ is the periodic function with period 1 such that $P_1(x) = -x + \frac{1}{2}$ for $x \in (0,1)$. This implies

$$\frac{d^2\ln\Gamma(\theta)}{d\theta^2} = \frac{1}{\theta} + \frac{1}{2\theta^2} + 2\int_0^\infty \frac{P_1(x)}{(\theta+x)^3}\,dx.$$

Thus for all $\theta \in (0,\infty)$

$$\kappa(\widehat{\theta}_n;\theta) = 1 + \frac{1}{2\theta} + 2\theta\int_0^\infty \frac{P_1(x)}{(\theta+x)^3}\,dx > 1. \tag{3.4.59}$$

The coefficient $\kappa(\widehat{\theta}_n;\theta)$ can be made as large as we want by choosing a sufficiently large θ. Thus (3.4.57)–(3.4.59) imply that the estimator $\widehat{\theta}_n$ is not asymptotically efficient whatever the parameter θ is.

EXAMPLE 3.4.4. Let $\xi^{(n)} = (\xi_1,\dots,\xi_n)$ be a sample from the exponential distribution with the density

$$f(x;\theta) = e^{-x+\theta}I_{[\theta,\infty)}(x).$$

Regularity conditions do not hold in this case, since the function $f(x;\theta)$ is discontinuous with respect to θ. Consider the estimator

$$\widehat{\theta}_n = \min_{1\le i\le n}\xi_i - \frac{1}{n}.$$

We learned from Example 3.2.2 that $\widehat{\theta}_n$ is the Pitman estimator of the parameter θ and moreover

$$\mathsf{E}_\theta\widehat{\theta}_n = \theta, \qquad \mathsf{E}_\theta(\widehat{\theta}_n-\theta)^2 = \frac{2}{n^2}.$$

This implies that the mean square error $\mathsf{E}_\theta(\widehat{\theta}_n-\theta)^2$ is of order n^{-2} for large n. This is a higher rate of decay as compared to the one given by the Cramér–Rao lower bound. This phenomenon occurs, since the regularity conditions fail in this case. Other examples of higher rates of decay of $\mathsf{E}_\theta(\widehat{\theta}_n-\theta)^2$ can be obtained by using the lower bound in the Chapman–Robbins inequality in some other cases where the regularity conditions fail.

REMARK 3.4.5. Further information about the regularity conditions and Cramér–Rao inequalities can be found in [**36**] and [**13**].

REMARK 3.4.6. The Cramér–Rao inequalities belong to the family of results, called *information inequalities*, which provide lower bounds for the risk functions or risks of estimators of parameters. See [**22**], Chapter 5, about the relation between the Cramér–Rao inequalities and for other information about inequalities.

3.5. The Cramér–Rao inequality for a multidimensional parameter

In this section we consider analogs of the Cramér–Rao inequalities for the case of a multidimensional parameter θ.

The Fisher information matrix. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$. Assume that its distribution belongs to a family of probability measures $\{\mathsf{P}_\theta,\ \theta \in \Theta\}$ where Θ is some subset of $\mathbf{R}^k$, $k \geq 1$. As in the case of a one-dimensional parameter we assume that for all $\theta \in \Theta$ the measure P_θ is absolutely continuous with respect to some σ-finite measure μ on $(X, \mathcal{B})$ and that there exists the density $f(x;\theta)$ of the measure P_θ with respect to the measure μ.

Let the derivatives $S_i(x;\theta) = \partial \ln f(x;\theta)/\partial\theta_i$, $i = 1, \ldots, k$, exist for μ-almost all $x \in X$. The matrix $I(\theta)$ with the entries $I_{ij}(\theta) = \mathsf{E}_\theta S_i(\xi;\theta) S_j(\xi;\theta)$, $i, j = 1, \ldots, k$, is called the *Fisher information matrix*. In the case $k = 1$, $I(\theta)$ is the Fisher information.

It is easy to see that the matrix $I(\theta)$ is nonnegative definite. Indeed, for all $\lambda = (\lambda_1, \ldots, \lambda_k)' \in \mathbf{R}^k$,

$$\begin{aligned} \lambda' I(\theta)\lambda &= \sum_{i,j=1}^{k} I_{ij}(\theta)\lambda_i\lambda_j = \mathsf{E}_\theta \sum_{i,j=1}^{k} S_i(\xi;\theta) S_j(\xi;\theta)\lambda_i\lambda_j \\ &= \mathsf{E}_\theta \left(\sum_{i=1}^{k} S_i(\xi;\theta)\lambda_i \right)^2 = \mathsf{E}_\theta \left(\lambda' S(\xi;\theta)\right)^2 \geq 0 \end{aligned} \tag{3.5.1}$$

where $S(\xi;\theta)$ is the vector defined by

$$S(\xi;\theta) = (S_1(\xi;\theta), S_2(\xi;\theta), \ldots, S_k(\xi;\theta))'. \tag{3.5.2}$$

Inequality (3.5.1) implies that the Fisher information matrix $I(\theta)$ is nonnegative definite and this explains why we write $I(\theta) \geq 0$ in this case. If the matrix $I(\theta)$ is positive definite, then $\lambda' I(\theta)\lambda > 0$ for all vectors $\lambda \neq 0$. We write $I(\theta) > 0$ in the latter case. We write $A \geq B$ for matrices A and B if $A - B \geq 0$, that is, if the matrix $A - B$ is nonnegative definite.

The Cramér–Rao inequality for unbiased estimators. Let $\widehat{\theta}$ be an estimator of a parameter θ constructed from an observation ξ where $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k)'$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$. Denote by $R(\widehat{\theta};\theta)$ the matrix with entries

$$\mathsf{E}_\theta(\widehat{\theta}_i - \theta_i)(\widehat{\theta}_j - \theta_j), \qquad i, j = 1, 2, \ldots, k.$$

In other words, $R(\widehat{\theta};\theta) = \mathsf{E}_\theta(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)'$ is the matrix of mixed moments of the vector $\widehat{\theta} - \theta$. It is easy to show that for all vectors $\lambda \in \mathbf{R}^k$

$$\mathsf{E}_\theta((\widehat{\theta} - \theta)'\lambda)^2 = \lambda' R(\widehat{\theta};\theta)\lambda, \tag{3.5.3}$$

that is, the matrix $R(\widehat{\theta};\theta)$ is nonnegative definite.

If the matrix $I(\theta)$ is nondegenerate, then we will show that $R(\widehat{\theta};\theta) \geq I^{-1}(\theta)$ under some conditions on $f(x;\theta)$, that is, we will show for all vectors $\lambda \in \mathbf{R}^k$ that

$$\lambda' R(\widehat{\theta};\theta)\lambda \geq \lambda' I^{-1}(\theta)\lambda. \tag{3.5.4}$$

The latter is a *matrix analog of the Cramér–Rao inequality*. Inequality (3.5.4) for an unbiased estimator $\widehat{\theta}$ means in view of relations (3.5.1) and (3.5.3) that the variance of the projection of the vector $\widehat{\theta}$ on an arbitrary direction λ in $\mathbf{R}^k$ is greater than or equal to the variance of the projection of the vector $S(\xi;\theta)$ on the same direction.

Let $\mathbf{C}_1^0(\Theta)$ be the class of real functions $\phi(\theta)$ defined on $\Theta \subset \mathbf{R}^k$ that are differentiable in θ almost everywhere with respect to the Lebesgue measure, and such that the function $\phi(\theta + t\lambda)$ is absolutely continuous in t for all θ and λ for which $\theta + t\lambda \in \Theta$ and $0 \le t \le 1$. If $\phi \in \mathbf{C}_1^0(\Theta)$ and $\theta + t\lambda \in \Theta$ for $0 \le t \le \Delta$, then

$$\phi(\theta+\Delta\lambda) - \phi(\theta) = \int_0^\Delta \frac{d\phi(\theta+u\lambda)}{du}\,du = \int_0^\Delta \lambda' \frac{\partial\phi(\theta+u\lambda)}{\partial\theta}\,du \tag{3.5.5}$$

where

$$\frac{\partial\phi(\theta)}{\partial\theta} = \left(\frac{\partial\phi(\theta)}{\partial\theta_1}, \frac{\partial\phi(\theta)}{\partial\theta_2}, \dots, \frac{\partial\phi(\theta)}{\partial\theta_k}\right)'.$$

The following result contains sufficient regularity conditions posed on the density $f(x;\theta)$ under which the Cramér–Rao inequality (3.5.4) holds.

THEOREM 3.5.1. *Let $\sqrt{f(x;\theta)} \in \mathbf{C}_1^0(\Theta)$ for μ-almost all $x \in X$. Assume that the matrix $I(\theta)$ is continuous in θ and nondegenerate. If $\widehat{\theta}$ is an unbiased estimator of the parameter θ, then*

$$R(\widehat{\theta};\theta) \ge I^{-1}(\theta) \tag{3.5.6}$$

for all points θ of continuity of the matrix $R(\widehat{\theta};\theta)$.

PROOF. Let $\theta \in \Theta$, $\lambda \in \mathbf{R}^k$, and $|\lambda| = 1$. Then $\theta + t\lambda \in \Theta$ for all sufficiently small $\Delta > 0$ and for all $t \in [0,\Delta]$. Since $\widehat{\theta}$ is an unbiased estimator,

$$\mathsf{E}_\theta\widehat{\theta} = \theta, \qquad \mathsf{E}_{\theta+\Delta\lambda}\widehat{\theta} = \theta + \Delta\lambda,$$

whence we obtain

$$\int (\widehat{\theta}(x) - \theta)[f(x;\theta+\Delta\lambda) - f(x;\theta)]\,\mu(dx) = \Delta\lambda.$$

Multiplying this equality on the left by u' and applying the Cauchy–Bunyakovskiĭ inequality we get

$$\begin{aligned}\Delta^2(u'\lambda)^2 &\le \int \left(u'(\widehat{\theta}(x) - \theta)\right)^2 \left(\sqrt{f(x;\theta+\Delta\lambda)} + \sqrt{f(x;\theta)}\right)^2 \mu(dx) \\ &\qquad \times \int \left(\sqrt{f(x;\theta+\Delta\lambda)} - \sqrt{f(x;\theta)}\right)^2 \mu(dx) \\ &\le 2\int \left(u'(\widehat{\theta}(x) - \theta)\right)^2 (f(x;\theta+\Delta\lambda) + f(x;\theta))\,\mu(dx) \\ &\qquad \times \int \left(\sqrt{f(x;\theta+\Delta\lambda)} - \sqrt{f(x;\theta)}\right)^2 \mu(dx).\end{aligned} \tag{3.5.7}$$

Relations (3.5.3) and (3.5.7) imply that

$$\begin{aligned}\Delta^2(u'\lambda)^2 &\le 2\left[u'R(\widehat{\theta};\theta)u + u'R(\widehat{\theta};\theta+\Delta\lambda)u + \Delta^2(u'\lambda)^2\right] \\ &\qquad \times \int \left(\sqrt{f(x;\theta+\Delta\lambda)} - \sqrt{f(x;\theta)}\right)^2 \mu(dx).\end{aligned} \tag{3.5.8}$$

Since $\sqrt{f(x;\theta)} \in \mathbf{C}_1^0(\Theta)$, we obtain from (3.5.5) that

$$\begin{aligned}
&\int \left(\sqrt{f(x;\theta+\Delta\lambda)} - \sqrt{f(x;\theta)}\right)^2 \mu(dx) \\
&\quad = \int \left(\int_0^\Delta \frac{\lambda' \partial f(x;\theta+t\lambda)/\partial\theta}{2\sqrt{f(x;\theta+t\lambda)}}\, dt\right)^2 \mu(dx) \qquad (3.5.9) \\
&\quad \le \frac{\Delta}{4}\int_0^\Delta \lambda' I(\theta+t\lambda)\lambda\, dt.
\end{aligned}$$

Let θ be a point of continuity of the matrix $R(\widehat{\theta};\theta)$. Substituting (3.5.9) into (3.5.8) and passing to the limit as $\Delta \to 0$ we get

$$\big(u' R(\widehat{\theta};\theta)u\big)(\lambda' I(\theta)\lambda) \ge (u'\lambda)^2. \qquad (3.5.10)$$

Putting $\lambda = I^{-1}(\theta)u$ we derive from (3.5.10) that

$$u' R(\widehat{\theta};\theta)u \ge u' I^{-1}(\theta)u$$

for all vectors $u \in \mathbf{R}^k$. The latter inequality is equivalent to (3.5.6). □

Let an observation be a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ and let the density of the sample be $f_n(x;\theta) = \prod_{i=1}^n f(x_i;\theta)$ where $x = (x_1, \dots, x_n)$. Let $I_n(\theta)$ and $I(\theta)$ be the information matrices constructed from the densities $f_n(x;\theta)$ and $f(x;\theta)$, respectively. If the regularity conditions of Theorem 3.5.1 hold, then

$$I_n(\theta) = nI(\theta).$$

If $\widehat{\theta}_n$ is an unbiased estimator of the parameter θ, then under the conditions of Theorem 3.5.1 we get the following matrix analog of the Cramér–Rao inequality:

$$R(\widehat{\theta}_n;\theta) \ge \frac{1}{n} I^{-1}(\theta). \qquad (3.5.11)$$

The Cramér–Rao inequality for biased estimators. Let $\widehat{\theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)'$ be an estimator of a parameter $\theta = (\theta_1, \dots, \theta_k)'$ constructed from an observation ξ. Put

$$a(\theta) = \mathsf{E}_\theta \widehat{\theta} = \theta + b(\theta), \qquad b(\theta) = (b_1(\theta), \dots, b_k(\theta))'.$$

Here $b(\theta)$ is the *bias vector of the estimator* $\widehat{\theta}$ of the parameter θ.

Consider the *multivariate analog of the regularity conditions* (R) introduced in Section 3.4:

(R) the function $\sqrt{f(x;\theta)}$ is continuously differentiable in θ for μ-almost all x; the matrix $I(\theta)$ is nondegenerate and continuous in θ.

In what follows we need the following *Cauchy–Bunyakovskiĭ inequality for matrices*.

LEMMA 3.5.1. *Let η and ζ be two random matrices of the same size (they are not necessarily square matrices). Assume that the inverse matrix of $\mathsf{E}\eta\eta'$ exists. Then*

$$\mathsf{E}\zeta\zeta' \ge \mathsf{E}\zeta\eta'(\mathsf{E}\eta\eta')^{-1}\mathsf{E}\eta\zeta'. \qquad (3.5.12)$$

This inequality becomes an inequality if and only if $\zeta = z\eta$ where $z = \mathsf{E}\zeta\eta'(\mathsf{E}\eta\eta')^{-1}$.

PROOF. Since $AA' \geq 0$ for an arbitrary matrix A (that is, the matrix AA' is nonnegative definite),

$$0 \leq \mathsf{E}(\zeta - z\eta)(\zeta - z\eta)' = \mathsf{E}\zeta\zeta' - z\mathsf{E}\eta\zeta' - \mathsf{E}\zeta\eta'z' + z\mathsf{E}\eta\eta'z'$$

for a nonrandom square matrix z. Putting $z = \mathsf{E}\zeta\eta'(\mathsf{E}\eta\eta')^{-1}$ we obtain inequality (3.5.12). The statement concerning the case of an equality in (3.5.12) is obvious. □

The following result contains the Cramér–Rao inequality for a biased estimator $\widehat{\theta}$ under the regularity conditions (R).

THEOREM 3.5.2. *Let conditions* (R) *hold. Let* $D(\widehat{\theta};\theta) = \mathsf{E}_\theta(\widehat{\theta} - a(\theta))(\widehat{\theta} - a(\theta))'$ *be the matrix of mixed central moments of second order of an estimator* $\widehat{\theta}$ *of a parameter* θ. *Then*

$$D(\widehat{\theta};\theta) \geq (I_k + B(\theta))I^{-1}(\theta)(I_k + B(\theta))' \tag{3.5.13}$$

where I_k *is the unit matrix,* $B(\theta) = \|b_{ij}(\theta)\|$, *and* $b_{ij}(\theta) = \partial b_i(\theta)/\partial\theta_j$.

Let $\det(D(\widehat{\theta};\theta)) > 0$ (*or* $\det(I_k + B(\theta)) > 0$) *for all* θ. *Then inequality* (3.5.13) *becomes an equality if and only if the density of the distribution is such that*

$$f(x;\theta) = \exp\{A(\theta)'\widehat{\theta}(x) + C(\theta)\}h(x), \qquad x \in X, \tag{3.5.14}$$

for some scalar functions $C(\theta)$ *and* $h(x)$ *where*

$$\|A_{ij}(\theta)\| = \left\|\frac{\partial A_i(\theta)}{\partial\theta_j}\right\| = \big((I_k + B(\theta))^{-1}\big)'I(\theta) \tag{3.5.15}$$

is the matrix of derivatives of the vector $A(\theta) = (A_1(\theta), \ldots, A_k(\theta))'$.

If $\widehat{\theta}$ *is an unbiased estimator, then*

$$D(\widehat{\theta};\theta) \geq I^{-1}(\theta). \tag{3.5.16}$$

Inequality (3.5.16) *becomes an equality if and only if relation* (3.5.14) *holds where* $\|A_{ij}(\theta)\| = I(\theta)$.

PROOF. As in the proof for the one-dimensional case we use the regularity conditions (R) to prove that

$$\mathsf{E}_\theta S_i(\xi;\theta) = 0, \qquad \mathsf{E}_\theta\widehat{\theta}_i S_j(\xi;\theta) = \delta_{ij} + b_{ij}(\theta), \qquad i,j = 1,2,\ldots,k$$

(see Lemma 3.4.4), where δ_{ij} is the Kronecker symbol and the functions $b_{ij}(\theta)$ are continuous. The latter condition can be written in matrix form as follows:

$$\mathsf{E}_\theta S(\xi;\theta) = 0, \tag{3.5.17}$$

$$\mathsf{E}_\theta\widehat{\theta}S(\xi;\theta)' = I_k + B(\theta) \tag{3.5.18}$$

where the matrix $B(\theta)$ is continuous in θ and the vector $S(\xi;\theta)$ is of the form (3.5.2). Equalities (3.5.17) and (3.5.18) imply that

$$\mathsf{E}_\theta(\widehat{\theta} - a(\theta))S(\xi;\theta)' = I_k + B(\theta). \tag{3.5.19}$$

Now we apply the Cauchy–Bunyakovskiĭ inequality for matrices. Put $\zeta = \widehat{\theta} - a(\theta)$ and $\eta = S(\xi;\theta)$ in (3.5.12). Then

$$\mathsf{E}_\theta \zeta\zeta' = \mathsf{E}_\theta(\widehat{\theta} - a(\theta))(\widehat{\theta} - a(\theta))' = D(\widehat{\theta};\theta),$$
$$\mathsf{E}_\theta \eta\eta' = \mathsf{E}_\theta S(\xi;\theta)S(\xi;\theta)' = I(\theta).$$

It follows from (3.5.19) that

$$\mathsf{E}_\theta \zeta\eta' = \mathsf{E}_\theta(\widehat{\theta} - a(\theta))S(\xi;\theta)' = I_k + B(\theta).$$

This together with (3.5.12) implies inequality (3.5.13).

According to Lemma 3.5.1, inequality (3.5.13) becomes an equality if and only if

$$(\widehat{\theta}(x) - a(\theta)) = (I_k + B(\theta))I^{-1}(\theta)S(x;\theta)$$

for points $(x;\theta)$ such that $f(x;\theta) > 0$. The latter condition is equivalent to

$$S(x;\theta) = I(\theta)(I_k + B(\theta))^{-1}(\widehat{\theta}(x) - a(\theta)). \tag{3.5.20}$$

If inequality (3.5.13) becomes an equality, then

$$\det(I_k + B(\theta))^2 = \det D(\widehat{\theta};\theta) \det I(\theta).$$

If $\det \mathsf{D}(\widehat{\theta};\theta)$ is far away from zero, then so is $\det(I_k + B(\theta))$, whence it follows that the inverse matrix $(I_k + B(\theta))^{-1}$ exists and is bounded. Thus the derivative $S(x;\theta)$ in (3.5.20) is bounded, $f(x;\theta) > 0$ everywhere on Θ, and equality (3.5.20) holds everywhere on Θ. Let $\theta_0, \theta \in \Theta$ and $\theta_0 + s(\theta - \theta_0) \in \Theta$ for all $s \in [0,1]$. Then

$$\ln f(x;\theta) = \ln f(x;\theta_0) + \int_0^1 (\theta - \theta_0)' S(x;\theta_0 + s(\theta - \theta_0))\, ds$$

in view of conditions (R). Thus

$$\ln f(x;\theta) = A(\theta)'\widehat{\theta}(x) + C(\theta) + H(x) \tag{3.5.21}$$

according to (3.5.20) where $C(\theta)$ and $H(x)$ are some scalar functions, and

$$A(\theta) = (A_1(\theta), \ldots, A_k(\theta))'$$

is a column-vector depending only on θ. This means that representation (3.5.14) holds.

If relation (3.5.21) is satisfied, then

$$S(x;\theta) = \|A_{ij}(\theta)\|'\widehat{\theta}(x) + \partial B(\theta)/\partial\theta \tag{3.5.22}$$

and

$$\partial B(\theta)/\partial\theta = -\|A_{ij}(\theta)\|'a(\theta), \tag{3.5.23}$$

since $\mathsf{E}_\theta S(\xi;\theta) = 0$. It follows from (3.5.22) and (3.5.23) that

$$S(x;\theta) = \|A_{ij}\|'(\widehat{\theta}(x) - a(\theta)).$$

Multiplying this equality on the right by $(\widehat{\theta}(x) - a(\theta))'$, we obtain from (3.5.19) that condition (3.5.20) (which is equivalent to the case of equality in (3.5.13)) follows from (3.5.15).

Inequality (3.5.16) follows from (3.5.13), since the matrix $B(\theta)$ is zero if $\widehat{\theta}$ is an unbiased estimator. □

All the remarks concerning the Cramér–Rao inequality that we made in Section 3.4 for the regularity conditions (R) in the one-dimensional case remain true in the multidimensional case, too.

One can prove the Cramér–Rao inequality for estimators $\widehat{g}$ of a function $g(\theta)$ of a parameter θ in the same way as in the one-dimensional case.

Note that if an observation is a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, then as above

$$I_n(\theta) = nI(\theta)$$

where $I_n(\theta)$ and $I(\theta)$ are the Fisher information matrices constructed from the distribution of the sample $\xi^{(n)}$ and from the distribution of the component ξ_1, respectively. If the regularity conditions (R) hold, then

$$D(\widehat{\theta}_n; \theta) \geq \frac{1}{n}(I_k + B_n(\theta))I^{-1}(\theta)(I_k + B_n(\theta))' \tag{3.5.24}$$

for all estimators $\widehat{\theta}_n$ constructed from the sample $\xi^{(n)}$ where

$$D(\widehat{\theta}_n; \theta) = \mathsf{E}_\theta(\widehat{\theta}_n - a_n(\theta))(\widehat{\theta}_n - a_n(\theta))', \tag{3.5.25}$$

$$a_n(\theta) = \mathsf{E}_\theta\widehat{\theta}_n = \theta + b_n(\theta), \qquad b_n(\theta) = \big(b_n^1(\theta), \dots, b_n^k(\theta)\big)', \tag{3.5.26}$$

$$B_n(\theta) = \|b_n^{ij}(\theta)\|, \qquad b_n^{ij}(\theta) = \partial b_n^i/\partial\theta_j. \tag{3.5.27}$$

The case of equality in (3.5.24) can be considered by applying Theorem 3.5.2.

Efficient and asymptotically efficient estimators. The definitions of efficient and asymptotically efficient estimators in the case of a multidimensional parameter are similar to those in the case of a one-dimensional parameter. An estimator θ^* is called an *efficient estimator* of a parameter θ if the Cramér–Rao inequality for this estimator becomes an equality. If θ^* is an unbiased estimator of a parameter θ, then θ^* is efficient if (3.5.6) becomes an equality, that is, if

$$R(\theta^*; \theta) = I^{-1}(\theta), \qquad \theta \in \Theta. \tag{3.5.28}$$

If θ^* is a biased estimator of a parameter θ, then it is efficient if inequality (3.5.13) becomes an equality, that is, if

$$D(\theta^*; \theta) = (I_k + B(\theta))I^{-1}(\theta)(I_k + B(\theta))', \qquad \theta \in \Theta. \tag{3.5.29}$$

If an observation is a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, then conditions (3.5.28) and (3.5.29) become of the form

$$R(\theta_n^*; \theta) = \frac{1}{n}I^{-1}(\theta), \qquad \theta \in \Theta, \tag{3.5.30}$$

$$D(\theta_n^*; \theta) = \frac{1}{n}(I_k + B_n(\theta))I^{-1}(\theta)(I_k + B_n(\theta))', \qquad \theta \in \Theta, \tag{3.5.31}$$

where $I(\theta)$ is the Fisher information matrix constructed from the distribution of ξ_1 and $B_n(\theta)$ is the matrix defined by (3.5.24)–(3.5.27) for $\widehat{\theta}_n = \theta_n^*$.

Equalities (3.5.30)–(3.5.31) hold, that is, θ_n^* is an efficient estimator, only in exceptional cases. However there exist the so-called asymptotically efficient estimators and conditions for their existence are quite general. An estimator θ_n^* of a parameter θ constructed from a sample $\xi^{(n)}$ is called *asymptotically efficient* if

$$R(\theta_n^*; \theta) = \frac{1 + o(1)}{n}I^{-1}(\theta), \qquad \theta \in \Theta, \tag{3.5.32}$$

where $R(\theta_n^*;\theta) = \mathsf{E}_\theta(\theta_n^* - \theta)(\theta_n^* - \theta)'$. The estimator θ_n^* is, generally speaking, biased.

EXAMPLE 3.5.1. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the normal distribution $\mathcal{N}(\theta_1, \theta_2)$. Thus $\theta = (\theta_1, \theta_2)'$ and the density $f(x;\theta)$ is of the form

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\theta_2} + \frac{x\theta_1}{\theta_2} - \frac{\theta_1^2}{2\theta_2} - \frac{1}{2}\ln\theta_2\right\}. \tag{3.5.33}$$

Consider the estimator $\widehat{\theta}_n = (\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n})$ such that

$$\widehat{\theta}_{1,n} = \overline{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i, \qquad \widehat{\theta}_{2,n} = \frac{1}{n-1}\sum_{i=1}^n (\xi_i - \overline{\xi})^2.$$

It is clear that $\widehat{\theta}_n$ is an unbiased estimator. It follows from (3.5.33) that representation (3.5.14) does not hold for the density $f_n(x;\theta)$, since

$$\begin{aligned} f_n(x;\theta) &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^n x_i^2 + \frac{\theta_1}{\theta_2}\sum_{i=1}^n x_i - \frac{n\theta_1^2}{2\theta_2} - \frac{n}{2}\ln\theta_2\right\} \\ &= (2\pi)^{-n/2} \exp\left\{\frac{n\theta_1}{\theta_2}\widehat{\theta}_{1,n} - \frac{n-1}{2\theta_2}\widehat{\theta}_{2,n} - \frac{n}{2\theta_2}\widehat{\theta}_{1,n}^2 - \frac{n\theta_1^2}{2\theta_2} - \frac{n}{2}\ln\theta_2\right\}. \end{aligned}$$

This means that the lower bound in the multivariate Cramér–Rao inequality is not attained and therefore $\widehat{\theta}_n$ is not an efficient estimator. Nevertheless $\widehat{\theta}_n$ is an asymptotically efficient estimator. Below we prove this result.

First we evaluate the matrix $I(\theta)$. We have

$$S_1(x;\theta) = \frac{x - \theta_1}{\theta_2}, \qquad S_2(x;\theta) = \frac{(x-\theta_1)^2}{2\theta_2^2} - \frac{1}{2\theta_2}$$

where $S_i(x;\theta) = \partial \ln f(x;\theta)/\partial\theta_i$, $i = 1, 2$. Thus

$$\begin{aligned} I_{11}(\theta) &= \mathsf{E}_\theta \frac{(\xi_1 - \theta_1)^2}{\theta_2^2} = \frac{1}{\theta_2}, \\ I_{12}(\theta) = I_{21}(\theta) &= \mathsf{E}_\theta\left(\frac{(\xi_1-\theta_1)^3}{2\theta_2^3} - \frac{\xi_1 - \theta_1}{2\theta_2^2}\right) = 0, \\ I_{22}(\theta) &= \mathsf{E}_\theta \frac{\left((\xi_1-\theta_1)^2 - \theta_2\right)^2}{4\theta_2^4} = \frac{1}{2\theta_2^2} \end{aligned}$$

and the lower bound in the Cramér–Rao inequality is given by

$$\frac{1}{n} I^{-1}(\theta) = \begin{pmatrix} \theta_2/n & 0 \\ 0 & 2\theta_2^2/n \end{pmatrix}. \tag{3.5.34}$$

Now we evaluate the matrix $R(\widehat{\theta}_n;\theta)$. We have

$$\begin{aligned} \mathsf{E}_\theta(\widehat{\theta}_{1,n} - \theta_1)^2 &= \mathsf{E}_\theta(\overline{\xi} - \theta_1)^2 = \frac{\theta_2}{n}, \\ \mathsf{E}_\theta(\widehat{\theta}_{2,n} - \theta_2)^2 &= \frac{2\theta_2^2}{n-1}, \qquad \mathsf{E}_\theta(\widehat{\theta}_{1,n} - \theta_1)(\widehat{\theta}_{2,n} - \theta_2) = 0. \end{aligned}$$

The first of the latter equalities is obvious. The last of them follows from the independence of $\widehat{\theta}_{1,n}$ and $\widehat{\theta}_{2,n}$ (see Theorem 1.4.2). To prove the second equality we note that

$$\mathcal{L}\left(\sum_{i=1}^{n}\frac{(\xi_i-\bar{\xi})^2}{\theta_2}\,\Big|\,\mathsf{P}_\theta\right)=\chi^2(n-1)$$

by Theorem 1.4.2. Since $\mathsf{D}\chi^2_{n-1}=2(n-1)$, we get the desired equality. Thus the matrix $R(\widehat{\theta}_n;\theta)$ is given by

$$R(\widehat{\theta}_n;\theta)=\begin{pmatrix}\theta_2/n & 0\\ 0 & 2\theta_2^2/(n-1)\end{pmatrix}. \tag{3.5.35}$$

Finally we apply (3.5.34) and (3.5.35) and obtain (3.5.32). This shows that $\widehat{\theta}_n$ is an asymptotically efficient estimator.

Other results related to the Cramér–Rao inequality. Inequality (3.5.12) allows one to obtain some other results corresponding to other matrices ζ and η. We restrict our consideration to the case of a one-dimensional parameter θ. Assume that the density $f(x;\theta)$ satisfies a stronger condition as compared to the regularity condition (R), namely let

(i_1) the density $f(x;\theta)$ be continuously differentiable $m\geq 1$ times in θ;

(i_2) the integrals

$$K_j(\theta)=\int_{N_\theta}\frac{\left|\partial^j f(x;\theta)/\partial\theta^j\right|^2}{f(x;\theta)}\,\mu(dx),\qquad j=1,2,\dots,m,$$

converge for all $\theta\in\Theta$ and, as functions of θ, be continuous on Θ where $N_\theta=\{x\colon f(x;\theta)\neq 0\}$.

Using the same method as that in the proof of Lemma 3.4.4 we show that conditions (i_1)–(i_2) imply that the function $a(\theta)=\mathsf{E}_\theta\widehat{\theta}$ has m continuous derivatives for an arbitrary estimator $\widehat{\theta}$ if its second moment $\mathsf{E}_\theta\widehat{\theta}^2$ is locally bounded.

Let $c=(c_1,c_2,\dots,c_m)$ be some vector of $\mathbf{R}^m$. Put $\zeta=\widehat{\theta}(\xi)-a(\theta)$ and

$$\eta=\sum_{j=1}^{m}c_j\frac{\partial^j f(\xi;\theta)/\partial\theta^j}{f(\xi;\theta)}I_{N_\theta}(\xi).$$

Then it follows from (3.5.12) that

$$\mathsf{D}_\theta\widehat{\theta}\geq\sup_c\frac{\left(c_1+\sum_{j=1}^{m}c_j\left(\partial^j b(\theta)/\partial\theta^j\right)\right)^2}{\sum_{i,j=1}^{m}c_ic_j\int_{N_\theta}\left(\partial^i f(x;\theta)/\partial\theta^i\right)\left(\partial^j f(x;\theta)/\partial\theta^j\right)f^{-1}(x;\theta)\,\mu(dx)} \tag{3.5.36}$$

where $b(\theta)=\mathsf{E}_\theta\widehat{\theta}-\theta$ is the bias of the estimator $\widehat{\theta}$. Inequality (3.5.36) is called the *Bhattacharyya inequality*. More details about the Bhattacharyya inequality can be found in [**36**].

Assume that the set N_θ does not depend on θ and let $N_\theta=N$ for all $\theta\in\Theta$. Below we avoid the regularity conditions posed on the density $f(x;\theta)$. Denote by M the set of charges m on Θ such that

$$\int_\Theta\left(1+f(x;u)\right)|m(du)|<\infty.$$

Put $\zeta = \widehat{\theta} - a(\theta)$ and

$$\eta = \frac{1}{f(\xi;\theta)} \int_\Theta f(\xi;u)\, m(du) I_N(\xi) - m(\Theta)$$

in (3.5.12). Inequality (3.5.12) becomes of the form
(3.5.37)

$$\mathsf{D}_\theta \widehat{\theta} \geq \sup_{m \in M} \frac{\left(\int_\Theta (a(u) - a(\theta))\, m(du)\right)^2}{\int_\Theta \int_\Theta m(du_1)\, m(du_2) \int_N f(x;u_1) f(x;u_2)/f(x;\theta)\, \mu(dx) - m^2(\Theta)}$$

and is called the *Barankin–Kiefer inequality* in this case. If the upper bound in the Barankin–Kiefer inequality is evaluated only with respect to δ-measures on Θ instead of charges $m \in M$, then (3.5.37) becomes of the form

$$\mathsf{D}_\theta \widehat{\theta} \geq \sup_{u \in \Theta} \frac{(a(u) - a(\theta))^2}{\int_N (f(x;u) - f(x;\theta))^2 f^{-1}(x;\theta)\, \mu(dx)} \tag{3.5.38}$$

and is called the *Chapman–Robbins inequality* (cf. inequality (3.4.43)).

3.6. Integral inequalities of Cramér–Rao type

We follow the Bayes approach and obtain lower estimates for risks of estimators in this section. The corresponding inequalities can be called integral inequalities of Cramér–Rao type, since they involve risk functions integrated with respect to the a priori measure.

Efficient and superefficient estimators. Throughout this section we assume that an observation is a sample $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ from a distribution for which there exists the density $f(x;\theta)$ with respect to a measure μ where $\theta = (\theta_1, \ldots, \theta_k)'$ is an unknown parameter whose value belongs to a certain set $\Theta \subset \mathbf{R}^k$, $k \geq 1$. First we consider the case $k = 1$, that is, the case of a scalar parameter θ. Let an arbitrary set of regularity conditions given in Section 3.4 hold. Then the Cramér–Rao inequality

$$\mathsf{E}_\theta (\widehat{\theta}_n - \theta)^2 \geq \frac{1}{nI(\theta)}, \qquad \theta \in \Theta, \tag{3.6.1}$$

holds for all unbiased estimators $\widehat{\theta}_n$ of the parameter θ where $I(\theta)$ is the Fisher information evaluated with respect to the density $f(x;\theta)$, that is,

$$I(\theta) = \mathsf{E}_\theta S^2(\xi_1;\theta), \qquad S(x;\theta) = \partial \ln f(x;\theta)/\partial\theta.$$

The right-hand side of inequality (3.6.1) is sometimes called the *Cramér–Rao bound*. This bound is attained if an estimator is efficient. The question is whether one can improve this result for biased estimators. In other words, the question is how precise is the Cramér–Rao bound for biased estimators.

It is quite obvious that the expectation $\mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2$ at a fixed point θ_0 can be smaller than the Cramér–Rao bound. Indeed, this is true for $\widehat{\theta}_n \equiv \theta_0$, for example. However the latter estimator is very bad at any other point.

Below is another example of this kind. Let $\mathcal{L}(\xi_1 \mid \mathsf{P}_\theta) = \mathcal{N}(\theta, 1)$ where

$$\theta \in \Theta = [0, \infty).$$

It is clear that the estimator $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n \xi_i$ is efficient. Nevertheless the estimator $\theta_n^* = 0 \vee \widehat{\theta}_n$ is even better, since it decreases the mean square deviation by substituting 0 for negative values of $\widehat{\theta}_n$ that are meaningless in view of the restriction that $\theta \in [0, \infty)$. On the other hand, θ_n^* is a biased estimator, since $\mathsf{E}_\theta \theta_n^* > \mathsf{E}_\theta \widehat{\theta}_n = \theta$. We have $I(\theta) = 1$ for all $\theta \in \Theta$. At the point $\theta = 0$ we get

$$\mathsf{E}_0 \widehat{\theta}_n^2 = \frac{1}{n} = \frac{1}{nI(0)}, \qquad \mathsf{E}_0 (\theta_n^*)^2 = \frac{1}{2n} < \frac{1}{nI(0)}.$$

The improvement of the Cramér–Rao bound is explained in this example by a restriction of the domain of the estimator $\widehat{\theta}_n$ to the set Θ.

Another example is due to Hodges. The improvement of the Cramér–Rao bound in the Hodges example is not due to the restriction of the set Θ and is explained by other circumstances.

Again let $\mathcal{L}(\xi_1 \mid \mathsf{P}_\theta) = \mathcal{N}(\theta, 1)$ where $\theta \in \Theta = (-\infty, \infty)$. Along with an efficient estimator $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n \xi_i$ we consider another estimator

$$\theta_n^* = \begin{cases} \widehat{\theta}_n, & \text{if } |\widehat{\theta}_n| \geq n^{-1/4}, \\ \beta \widehat{\theta}_n, & \text{if } |\widehat{\theta}_n| < n^{-1/4}, \end{cases}$$

where $|\beta| < 1$. It is easy to see for $\theta > 0$ that

$$\mathsf{P}_\theta \left(|\widehat{\theta}_n| < n^{-1/4} \right) \leq \mathsf{P}_\theta \left((\widehat{\theta}_n - \theta)\sqrt{n} < n^{1/4} - \theta\sqrt{n} \right) = \Phi \left(n^{1/4} - \theta\sqrt{n} \right) \to 0$$

as $n \to \infty$ where $\Phi(x)$ is the standard $\mathcal{N}(0,1)$ distribution function. A similar result holds for the case $\theta < 0$, too. If $\theta \neq 0$, then the estimator θ_n^* coincides with $\widehat{\theta}_n$ on an event whose probability approaches 1 as $n \to \infty$. Thus

$$\mathcal{L}\big((\theta_n^* - \theta)\sqrt{n} \mid \mathsf{P}_\theta\big) \to \mathcal{N}(0,1)$$

as $n \to \infty$ if $\theta \neq 0$.

If $\theta = 0$ and as $n \to \infty$

$$\mathsf{P}_0 \left(|\widehat{\theta}_n| < n^{-1/4} \right) = \mathsf{P}_0 \left(|\widehat{\theta}_n \sqrt{n}| < n^{1/4} \right) = 1 - 2\Phi \left(-n^{1/4} \right) \to 1.$$

If $\theta = 0$, then the estimator θ_n^* coincides with $\beta\widehat{\theta}_n$ on an event whose probability approaches 1 as $n \to \infty$. Hence

$$\mathcal{L}\left(\theta_n^* \sqrt{n} \mid \mathsf{P}_0\right) \to \mathcal{N}\left(0, \beta^2\right)$$

as $n \to \infty$.

Therefore the estimator θ_n^* is asymptotically $\mathcal{N}(\theta, \sigma^2(\theta) n^{-1})$ normal for all θ where

$$\sigma^2(\theta) = \begin{cases} 1, & \text{if } \theta \neq 0, \\ \beta^2, & \text{if } \theta = 0. \end{cases}$$

Note that $\beta^2 < 1$. Thus the asymptotic variance of the estimator θ_n^* at the point $\theta = 0$ is equal to $n^{-1}\beta^2$ which is less than the lower Cramér–Rao bound

$$(nI(0))^{-1} = n^{-1}.$$

Asymptotically normal estimators whose asymptotic variance $\sigma^2(\theta)$ is such that $\sigma^2(\theta)/n \leq 1/(nI(\theta))$ and is less than $(nI(\theta))^{-1}$ for some θ are sometimes called

superefficient. The points θ for which $\sigma^2(\theta)n^{-1} < (nI(\theta))^{-1}$ are called the *points of superefficiency*.

The examples of superefficient estimators do not change our conclusion that efficient or asymptotically efficient estimators are the best. Namely Le Cam (1953) proved that an improvement of an efficient estimator can be made only at a set of points of superefficiency whose Lebesgue measure is small.

We show in this section that $\inf_{\widehat{\theta}_n} \mathsf{E}_t(\widehat{\theta}_n - t)^2 = 0$ for all t and that there is a lower bound of the integral of $\mathsf{E}_t(\widehat{\theta}_n - t)^2$ that does not depend on $\widehat{\theta}_n$ but still is closely related to the function $(nI(t))^{-1}$. More precisely, we obtain a lower bound for

$$\inf_{\widehat{\theta}_n} \int_\Theta \mathsf{E}_t(\widehat{\theta}_n; t)^2 q(t)\, dt \tag{3.6.2}$$

for an arbitrary nonnegative weight function $q(t)$ such that $\int_\Theta q(t)\, dt = 1$. This lower bound is close to J/n where

$$J = \int_\Theta \frac{q(t)}{I(t)}\, dt. \tag{3.6.3}$$

Note that the integral in (3.6.2) can be treated as the unconditional mathematical expectation $\mathsf{E}(\widehat{\theta}_n - \theta)^2$ for the Bayes approach where the a priori measure

$$\mathbf{Q}(A) = \int_A q(t)\, dt$$

is a probability distribution of the parameter θ and the density $q(t)$ of $\mathbf{Q}$ with respect to the Lebesgue measure exists. Relation (3.6.3) in this case can be rewritten as $J = \mathsf{E}I^{-1}(\theta)$.

Integral inequalities. Let $f_n(x;t)$ be the density of the sample $\xi^{(n)}$ with respect to the measure μ if $\theta = t$. Then $f_n(x;t)q(t) = p_n(x,t)$ is the density of the joint distribution of the vector $(\xi^{(n)}, \theta)$. Denote by $N_h \subset \Theta$ the support of a function h defined on Θ. In other words $N_h = \{t: h(t) \neq 0\}$. By N we denote the support of the function $p_n(x;t)$ in $X \times \Theta$, that is, N is the support of the distribution of the vector $(\xi^{(n)}; \theta)$.

THEOREM 3.6.1. *Let the function $f_n(x;t)$ be differentiable with respect to t, while $\sqrt{I(t)}$ is integrable on every finite interval. Then*

$$\begin{aligned} \mathsf{E}(\widehat{\theta}_n - \theta)^2 &\geq \frac{[\mathsf{E}(h(\theta)/q(\theta))]^2}{n\mathsf{E}\big(I(\theta)[h(\theta)/q(\theta)]^2\big) + \mathsf{E}(h'(\theta)/q(\theta))^2} \\ &= \frac{(\int h(t)\, dt)^2}{n\int I(t)h^2(t)/q(t)\, dt + \int (h'(t))^2/q(t)\, dt} \end{aligned} \tag{3.6.4}$$

for all differentiable functions $h(t)$ with bounded support such that $N_h \subset N_q$ and for all estimators $\widehat{\theta}_n$ of the parameter θ.

PROOF. Since the support of the function $h(t)$ is bounded, we get

$$\int (f_n(x;t)h(t))'\, dt = \int d(f_n(x;t)h(t)) = 0,$$

$$\int t(f_n(x;t)h(t))'\, dt = -\int f_n(x;t)h(t)\, dt.$$

Thus

$$\int_X \int_\Theta (\widehat{\theta}_n(x) - t)(f_n(x;t)h(t))'\, dt\, \mu(dx) = \int_X \int_\Theta f_n(x;t)h(t)\, dt\, \mu(dx) = \int_\Theta h(t)\, dt \tag{3.6.5}$$

for an arbitrary estimator $\widehat{\theta}_n$. Since $N_h \subset N_q$, equality (3.6.5) holds for integrals over the set N. Multiplying and dividing (3.6.5) by $p_n(x;t)$ we obtain

$$\mathsf{E}\left[(\widehat{\theta}_n - \theta)\frac{\left(f_n\left(\xi^{(n)};\theta\right)h(\theta)\right)'}{f_n\left(\xi^{(n)};\theta\right)q(\theta)}\right] = \int_{N_q} h(t)\, dt = \mathsf{E}\frac{h(\theta)}{q(\theta)}.$$

By the Cauchy–Bunyakovskiĭ inequality

$$\mathsf{E}(\widehat{\theta}_n - \theta)^2 \geq \frac{\left[\mathsf{E}\left(h(\theta)/q(\theta)\right)\right]^2}{\mathsf{E}\left[\left(f_n\left(\xi^{(n)};\theta\right)h(\theta)\right)' / \left(f_n\left(\xi^{(n)};\theta\right)q(\theta)\right)\right]^2}. \tag{3.6.6}$$

It remains to rewrite the latter result in the form of (3.6.4). Note that

$$\mathsf{E}_t\left|S_n\left(\xi^{(n)};t\right)\right| \leq n\sqrt{I(t)} \tag{3.6.7}$$

and for almost all t

$$\mathsf{E}_t S_n\left(\xi^{(n)};t\right) = 0 \tag{3.6.8}$$

where $S_n(x;t) = \partial \ln f_n(x;t)/\partial t$. Estimate (3.6.7) follows from

$$\mathsf{E}_t\left|S_n\left(\xi^{(n)};t\right)\right| \leq n\mathsf{E}_t|S(\xi_1;t)| \leq n\left(\mathsf{E}_t S^2(\xi_1;t)\right)^{1/2} = n\sqrt{I(t)},$$

since

$$S_n\left(\xi^{(n)};t\right) = \sum_{i=1}^n S(\xi_i;t) \quad \text{and} \quad S(x;t) = \partial \ln f(x;t)/\partial t.$$

To prove equality (3.6.8) we consider an arbitrary function $g(t)$ whose support is bounded and whose derivative $g'(t)$ is continuous everywhere. Then

$$\int g(t)\frac{\partial f_n(x;t)}{\partial t}\, dt = -\int g'(t)f_n(x;t)\, dt.$$

Moreover

$$\int |g(t)|\mathsf{E}_t\left|S_n\left(\xi^{(n)};t\right)\right|\, dt \leq n\int |g(t)|\sqrt{I(t)}\, dt < \infty.$$

This implies that one can interchange the integrals:

$$\begin{aligned}\int g(t)\mathsf{E}_t S_n\left(\xi^{(n)};t\right)\, dt &= \int_X \int_\Theta g(t)\frac{\partial f_n(x;t)}{\partial t}\, dt\, \mu(dx)\\ &= -\int_X \int_\Theta g'(t)f_n(x;t)\, dt\, \mu(dx)\\ &= -\int_\Theta g'(t)\, dt = -\int_\Theta dg(t) = 0.\end{aligned}$$

Since this equality holds for all g, we prove that (3.6.8) holds for almost all t.

Now we transform the right-hand side of (3.6.6):

$$\begin{aligned}
&\mathsf{E}\left(\frac{\left(f_n\left(\xi^{(n)};\theta\right)h(\theta)\right)'}{f_n\left(\xi^{(n)};\theta\right)q(\theta)}\right)^2 \\
&\quad = \mathsf{E}\left(S_n\left(\xi^{(n)};\theta\right)\frac{h(\theta)}{q(\theta)} + \frac{h'(\theta)}{q(\theta)}\right)^2 \\
&\quad = \mathsf{E}\left[\left(\frac{h(\theta)}{q(\theta)}\right)^2 \mathsf{E}_\theta S_n^2\left(\xi^{(n)};\theta\right)\right] + 2\mathsf{E}\left[\frac{h'(\theta)h(\theta)}{q^2(\theta)}\mathsf{E}_\theta S_n\left(\xi^{(n)};\theta\right)\right] + \mathsf{E}\left(\frac{h'(\theta)}{q(\theta)}\right)^2 \\
&\quad = n\mathsf{E}\left(\frac{h(\theta)}{q(\theta)}\right)^2 I(\theta) + \mathsf{E}\left(\frac{h'(\theta)}{q(\theta)}\right)^2.
\end{aligned}$$

Here we used equalities $\mathsf{E}_t S_n^2(\xi^{(n)};t) = nI(t)$ and

$$\mathsf{E}\left(\frac{h'(\theta)h(\theta)}{q^2(\theta)}\mathsf{E}_\theta S_n\left(\xi^{(n)};\theta\right)\right) = \int_{N_q}\frac{h'(t)h(t)}{q(t)}\mathsf{E}_t S_n\left(\xi^{(n)};t\right)dt = 0$$

following from (3.6.8). Thus relation (3.6.4) is proved. □

THEOREM 3.6.2. *Let the function $f_n(x;t)$ satisfy the conditions of Theorem 3.6.1. Assume that the function $h(t) = h_0(t) \equiv q(t)/I(t)$ has finite support and is differentiable. Then for all estimators $\widehat{\theta}_n$ one has*

$$\mathsf{E}(\widehat{\theta}_n - \theta)^2 \geq \frac{J}{n}\left(1 + \frac{H}{nJ}\right)^{-1} \geq \frac{J}{n} - \frac{H}{n^2} \tag{3.6.9}$$

where

$$H = \int\left[\left(\frac{q(t)}{I(t)}\right)'\right]^2\frac{dt}{q(t)}.$$

PROOF. Theorem 3.6.2 follows directly from Theorem 3.6.1 since the right-hand side of (3.6.4) is equal to $J^2/(nJ+H)$ for $h(t) = q(t)/I(t)$. □

REMARK 3.6.1. Inequalities (3.6.4) and (3.6.9) are *integral* in the sense that they provide the lower bounds for integrals of $\mathsf{E}_t(\widehat{\theta}_n - t)^2$.

We see from inequalities (3.6.4) and (3.6.9) that the lower bound of $\mathsf{E}(\widehat{\theta}_n - \theta)^2$ differs slightly from

$$\frac{J}{n} = \int\frac{1}{nI(t)}q(t)\,dt$$

for large n. The latter integral is equal to $\mathsf{E}(\theta_n^* - \theta)^2$ if θ_n^* is an efficient estimator. This indicates that one should use efficient estimators, since $\mathsf{E}(\widehat{\theta}_n - \theta)^2$ attains its minimum at efficient estimators whatever function $q(t)$ is.

The following example shows that the lower bounds (3.6.4) and (3.6.9) cannot be improved in general.

EXAMPLE 3.6.1. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a sample from the distribution $\mathcal{N}(\theta, 1)$. In this case $I(\theta) = 1$. Let the parameter θ be a random variable with a smooth density $q(t)$, $t \in (-\infty, \infty)$. Then the lower bound in (3.6.9) is of the form $(n + H)^{-1}$ where

$$H = \int \frac{q'(t)}{q(t)}\, dt = \mathsf{E}\big[(\ln q(\theta))'\big]^2.$$

Let θ_n^Q be the Bayes estimator of the parameter θ corresponding to the a priori probability measure $\mathbf{Q}$ whose density is $q(t)$ and let the loss function be quadratic. The estimator θ_n^Q minimizes the risk $\mathsf{E}(\widehat{\theta}_n - \theta)^2$ and coincides with the a posteriori mean $\theta_n^Q(x) = \mathsf{E}\{\theta \,/\, \xi^{(n)} = x\}$. Thus

$$\begin{aligned} \theta_n^Q(x) &= \frac{\int tq(t) f_n(x;t)\, dt}{\int q(t) f_n(x;t)\, dt} = \frac{\int tq(t) \exp\left(n\overline{x}t - nt^2/2\right)\, dt}{\int q(t) \exp\left(n\overline{x}t - nt^2/2\right)\, dt} \\ &= \frac{\int tq(t) \exp\left(-n(t-\overline{x})^2/2\right)\, dt}{\int q(t) \exp\left(-n(t-\overline{x})^2/2\right)\, dt} \end{aligned} \tag{3.6.10}$$

where $\overline{x} = n^{-1} \sum_{i=1}^n x_i$ and $x = (x_1, \dots, x_n)$. If the function $q(t)$ is sufficiently smooth, then (3.6.10) implies that

$$\theta_n^Q(x) = \overline{x} + \frac{q'(\overline{x})}{nq(x)} + O\left(\frac{1}{n^2}\right),$$

$$\mathsf{E}\left(\theta_n^Q - \theta\right)^2 = \frac{1}{n} - \frac{H}{n^2} + O\left(\frac{1}{n^3}\right)$$

as $n \to \infty$. In particular, let $q(t) = (2\pi)^{-1/2} \exp(-t^2/2)$. Then $H = 1$ and the lower bound in (3.6.9) is $(n+1)^{-1}$. On the other hand, we learned in Example 3.1.5 that

$$\mathsf{E}\left(\theta_n^Q - \theta\right)^2 = \frac{1}{n+1}.$$

This proves that the lower bounds in (3.6.4) and (3.6.9) cannot be improved, indeed.

The following result follows from Theorem 3.6.1. It allows one to make some conclusions concerning the points of superefficiency.

THEOREM 3.6.3. *Let the density $f_n(x;t)$ satisfy the assumptions of Theorem* 3.6.1. *If the interval $(a - \varepsilon, a + \varepsilon)$ belongs to Θ, then*

$$\max_{t \in (a-\varepsilon, a+\varepsilon)} \mathsf{E}_t(\widehat{\theta}_n - t)^2 \geq \left(n \max_{t \in (a-\varepsilon, a+\varepsilon)} I(t) + \pi^2 \varepsilon^{-2}\right)^{-1} \tag{3.6.11}$$

for an arbitrary estimator $\widehat{\theta}_n$.

PROOF. Let $q(t) = 0$ for $t \notin (a - \varepsilon, a + \varepsilon)$. Then

$$\max_{t \in (a-\varepsilon, a+\varepsilon)} \mathsf{E}(\widehat{\theta} - t)^2 \geq \int_{a-\varepsilon}^{a+\varepsilon} \mathsf{E}_t(\widehat{\theta}_n - t)^2 q(t)\, dt = \mathsf{E}(\widehat{\theta}_n - \theta)^2. \tag{3.6.12}$$

Put

$$h(t) = q(t) = \frac{1}{\varepsilon} \cos^2 \frac{\pi(t-a)}{2\varepsilon}, \qquad |t - a| \leq \varepsilon,$$

in Theorem 3.6.1. Then inequality (3.6.4) implies that

$$\mathsf{E}(\widehat{\theta}-\theta)^2 \geq \left(n\int_{a-\varepsilon}^{a+\varepsilon} I(t)q(t)\,dt + \int_{a-\varepsilon}^{a+\varepsilon} (q'(t))^2/q(t)\,dt\right)^{-1} \tag{3.6.13}$$

where

$$\begin{aligned}\int_{a-\varepsilon}^{a+\varepsilon} \frac{(q'(t))^2}{q(t)}\,dt &= \int_{-\varepsilon}^{\varepsilon}\left(\frac{\pi}{2\varepsilon^2}2\cos\frac{\pi t}{2\varepsilon}\sin\frac{\pi t}{2\varepsilon}\right)^2 \varepsilon\cos^{-2}\frac{\pi t}{2\varepsilon}\,dt \\ &= \frac{1}{\varepsilon^2}\int_{-1}^{1}\pi^2\sin^2\frac{\pi t}{2}\,dt = \frac{\pi^2}{\varepsilon^2}.\end{aligned} \tag{3.6.14}$$

Now relations (3.6.12)–(3.6.14) yield inequality (3.6.11). □

REMARK 3.6.2. It is not hard to show that the minimum of the functional $\int_{-1}^{1}(q'(t))^2q^{-1}(t)\,dt$ in the class of all differentiable densities $q(t)$ whose support belongs to $[-1,1]$ is attained for the density $q(t)=\cos^2(\pi t/2)$.

REMARK 3.6.3. Theorem 3.6.3 implies that the length of the interval of values of θ for which $\widehat{\theta}_n$ is superefficient does not exceed $O\left(n^{-1/2}\right)$.

Integral inequalities for nondifferentiable functions $q(t)/I(t)$. If the function $h(t)=q(t)/I(t)$ does not satisfy the assumptions of Theorem 3.6.1, then one can estimate the asymptotic behavior of $\mathsf{E}(\widehat{\theta}_n-\theta)^2$ by using the following result.

THEOREM 3.6.4. *Let the function $f_n(x;t)$ satisfy the assumptions of Theorem* 3.6.1. *Let the functions $h_\varepsilon(t)$ depend on a positive parameter ε, and satisfy the assumptions of Theorem* 3.6.1 *and*

1) $h_\varepsilon(t)\leq h_0(t)=q(t)/I(t)$ *for all* $\varepsilon>0$,
2) $H(\varepsilon)=\int(h_\varepsilon'(t))^2/q(t)\,dt<\infty$ *for all* $\varepsilon>0$.

Then for all $\varepsilon>0$

$$\mathsf{E}(\widehat{\theta}_n-\theta)^2 \geq \frac{(\int h_\varepsilon(t)\,dt)^2}{nJ+H(\varepsilon)}.$$

PROOF. It is necessary to put $h(t)=h_\varepsilon(t)$ in Theorem 3.6.1. □

Theorem 3.6.4 implies the following useful result.

THEOREM 3.6.5. *Let the function $f_n(x;t)$ satisfy the assumptions of Theorem* 3.6.1. *If the function $q(t)$ is Riemann integrable and $J<\infty$, then*

$$\mathsf{E}(\widehat{\theta}_n-\theta)^2 \geq \frac{J}{n}(1+o(1)) \tag{3.6.15}$$

as $n\to\infty$.

PROOF. Consider the following functions:

$$\widehat{q}_\varepsilon(t)=\min_{|u|\leq\varepsilon} q(t+u), \qquad q_\varepsilon(t)=\widehat{q}_\varepsilon(t)I(\widehat{q}_\varepsilon(t)\geq\varepsilon),$$

$$I_\varepsilon(t)=\max(\varepsilon,I(t)), \qquad h_\varepsilon(t)=\frac{1}{2\varepsilon}\int_{t-\varepsilon}^{t+\varepsilon}\frac{q_\varepsilon(s)}{I_\varepsilon(s)}\,ds.$$

It is clear that the support of the function $h_\varepsilon(t)$ is bounded, $h_\varepsilon(t)$ is differentiable for all $\varepsilon>0$, and $h_\varepsilon(t)\leq h_0(t)=q(t)/I(t)$.

Since the function $q(t)$ is Riemann integrable, we obtain $q_\varepsilon(t) \nearrow q(t)$ almost everywhere as $\varepsilon \to 0$. This result follows from

$$\int_a^b [q(t) - q_\varepsilon(t)]\, dt \downarrow 0 \tag{3.6.16}$$

as $\varepsilon \to 0$ for all $a, b \in \Theta$, $-\infty < a < b < \infty$. Moreover

$$\sum_k q_\delta(2k\delta)2\delta \uparrow \int_a^b q(t)\, dt \quad \text{and} \quad \sum_k q_\delta((2k+1)\delta)2\delta \uparrow \int_a^b q(t)\, dt$$

as $\delta \to 0$. Therefore, as $\varepsilon \to 0$

$$\begin{aligned}\int_a^b q_\varepsilon(t)\, dt &\geq \sum_k q_{2\varepsilon}(2k\varepsilon)2\varepsilon = \frac{1}{2}\left(\sum_k q_{2\varepsilon}(4k\varepsilon)4\varepsilon + \sum_k q_{2\varepsilon}((4k+2)\varepsilon)4\varepsilon\right) \\ &\to \int_a^b q(t)\, dt.\end{aligned}$$

Thus relation (3.6.16) is proved, whence the convergence $q_\varepsilon(t) \uparrow q(t)$ follows.

The convergence $q_\varepsilon(t) \uparrow q(t)$ implies that

$$\begin{gathered}\frac{q_\varepsilon(t)}{I_\varepsilon(t)} \uparrow \frac{q(t)}{I(t)} = h_0(t), \\ \int h_\varepsilon(t)\, dt = \int \frac{dt}{2\varepsilon} \int_{-e}^{\varepsilon} \frac{q_\varepsilon(t+s)}{I_\varepsilon(t+s)} ds = \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} ds \int \frac{q_\varepsilon(t)}{I_\varepsilon(t)}\, dt \\ = \int \frac{q_\varepsilon(t)}{I_\varepsilon(t)}\, dt \uparrow \int \frac{q(t)}{I(t)}\, dt = J\end{gathered}$$

as $\varepsilon \to 0$. Moreover

$$\begin{gathered}|h'_\varepsilon(t)| = \frac{1}{2\varepsilon}\left|\frac{q_\varepsilon(t+\varepsilon)}{I_\varepsilon(t+\varepsilon)} - \frac{q_\varepsilon(t-\varepsilon)}{I_\varepsilon(t-\varepsilon)}\right| \leq \frac{q(t)}{\varepsilon^2}, \\ H(\varepsilon) = \int \frac{(h'_\varepsilon(t))^2}{q(t)}\, dt \leq \int \frac{q(t)}{\varepsilon^4}\, dt = \frac{1}{\varepsilon^4}.\end{gathered}$$

Putting $\varepsilon = n^{-1/5}$ in Theorem 3.6.4 we obtain

$$\mathsf{E}(\widehat{\theta}_n - \theta)^2 \geq \frac{(\int h_\varepsilon(t)\, dt)^2}{nJ + H(\varepsilon)} \geq \frac{J^2 + o(1)}{nJ + n^{4/5}}$$

as $n \to \infty$, whence the desired estimate (3.6.15) follows. □

Asymptotically Bayes and asymptotically minimax estimators. One of the main results following from the above integral inequalities can be stated as follows. If an efficient or, at least, an asymptotically efficient estimator exists, then any other estimator is asymptotically "worse". Below we introduce the notions of the asymptotically Bayes and the asymptotically minimax estimators. We consider the quadratic loss function and the a priori measure $\mathbf{Q}$ and assume without loss of generality that $\mathbf{Q}$ is a probability measure for which the density $q(t)$ exists.

An estimator θ_n^* of a parameter θ is called *asymptotically Bayes* (with respect to the quadratic loss function and the a priori measure $\mathbf{Q}$) if

$$\limsup_{n\to\infty}\left[\mathsf{E}n(\theta_n^*-\theta)^2-\mathsf{E}n(\widehat{\theta}_n-\theta)^2\right]\le 0 \tag{3.6.17}$$

for an arbitrary estimator $\widehat{\theta}_n$. An estimator θ_n^* is called *asymptotically R-Bayes* if

$$\mathsf{E}n(\theta_n^*-\theta)^2=J+o(1) \tag{3.6.18}$$

as $n\to\infty$. In other words, an estimator is asymptotically R-Bayes if the lower bound for the mean square deviation given by Theorems 3.6.2 and 3.6.5 is attained at this estimator. Another name for this estimator is *asymptotically R-efficient in the mean square sense.*

The following result contains a relationship between asymptotically Bayes and asymptotically R-Bayes estimators.

THEOREM 3.6.6. *Let all the assumptions of Theorem* 3.6.1 *hold. If the function $q(t)$ is Riemann integrable, then every R-Bayes estimator is an asymptotically Bayes estimator.*

PROOF. Let θ_n^* be an asymptotically R-Bayes estimator, that is, (3.6.18) holds. According to Theorem 3.6.5

$$\liminf_{n\to\infty}\mathsf{E}n(\widehat{\theta}_n-\theta)^2\ge J$$

for an arbitrary estimator $\widehat{\theta}_n$. This together with (3.6.18) implies (3.6.17) for the estimator θ_n^*, that is, θ_n^* is an asymptotically Bayes estimator. □

It is clear that if an asymptotically R-Bayes estimator exists, then every asymptotically Bayes estimator is an asymptotically R-Bayes estimator.

The equivalence of all asymptotically R-Bayes estimators is established in the following result.

THEOREM 3.6.7. *Let all the assumptions of Theorem* 3.6.1 *hold. Assume that the function $q(t)$ is Riemann integrable. If θ_n^* and θ_n^{**} are two asymptotically R-Bayes estimators, then they are asymptotically equivalent in the following sense:*

$$\mathsf{E}n(\theta_n^*-\theta_n^{**})^2\to 0,\qquad (\theta_n^*-\theta_n^{**})\sqrt{n}\to 0 \tag{3.6.19}$$

as $n\to\infty$ where the second relation means the convergence in probability with respect to the joint distribution of $\xi^{(n)}$ and θ.

PROOF. It follows from (3.6.18) that

$$\lim_{n\to\infty}\mathsf{E}n(\theta_n^*-\theta)^2=\lim_{n\to\infty}\mathsf{E}n(\theta_n^{**}-\theta)^2=J. \tag{3.6.20}$$

Let $\widehat{\theta}_n=(\theta_n^*+\theta_n^{**})/2$. Relation (3.6.20) implies that

$$\lim_{n\to\infty}\mathsf{E}n(\widehat{\theta}_n-\theta)^2=J. \tag{3.6.21}$$

It is easy to show that

$$(\widehat{\theta}_n-\theta)^2+\left(\frac{\theta_n^*-\theta_n^{**}}{2}\right)^2=\frac{(\theta_n^*-\theta)^2+(\theta_n^{**}-\theta)^2}{2}.$$

This equality together with (3.6.20) and (3.6.21) yields

$$\lim_{n\to\infty} \mathsf{E}n(\theta_n^* - \theta_n^{**})^2 = 0$$

and the first relation in (3.6.19) is proved. The second relation in (3.6.19) obviously follows from the first one. □

Further we consider the asymptotically minimax approach. An estimator θ_n^* is called *asymptotically minimax* if

$$\limsup_{n\to\infty} \sup_{t\in\Theta} \mathsf{E}_t n(\theta_n^* - t)^2 \le \liminf_{n\to\infty} \sup_{t\in\Theta} \mathsf{E}_t n(\widehat{\theta}_n - t)^2 \tag{3.6.22}$$

for all estimators $\widehat{\theta}_n$.

The following result contains sufficient conditions that an estimator is asymptotically minimax.

THEOREM 3.6.8. *Let the Fisher information $I(\theta)$ exist and be continuous. If*

$$\limsup_{n\to\infty} \sup_{t\in\Theta} \mathsf{E}_t n(\theta_n^* - t)^2 \le \sup_{t\in\Theta} I^{-1}(t), \tag{3.6.23}$$

then θ_n^* is an asymptotically minimax estimator.

PROOF. It is sufficient to show that

$$\liminf_{n\to\infty} \sup_{t\in\Theta} \mathsf{E}_t n(\widehat{\theta}_n - t)^2 \ge \sup_{t\in\Theta} I^{-1}(t) \tag{3.6.24}$$

for an arbitrary estimator $\widehat{\theta}_n$. Let $\mathbf{Q}$ be an arbitrary probability measure on Θ whose density $q(t)$ is smooth. It is obvious that

$$\sup_{t\in\Theta} \mathsf{E}_t n(\widehat{\theta}_n - t)^2 \ge \int \mathsf{E}_t n(\widehat{\theta}_n - t)^2 q(t)\, dt. \tag{3.6.25}$$

The right-hand side of (3.6.25) is greater than or equal to $J - H/n$ according to Theorem 3.6.2. Thus (3.6.25) implies that

$$\liminf_{n\to\infty} \sup_{t\in\Theta} \mathsf{E}_t n(\widehat{\theta}_n - t) \ge \int I^{-1}(t) q(t)\, dt. \tag{3.6.26}$$

Since $q(t)$ is arbitrary, it can be specified such that

$$\int I^{-1}(t) q(t)\, dt \ge \sup_{t\in\Theta} I^{-1}(t) - \varepsilon \tag{3.6.27}$$

for a given $\varepsilon > 0$. The number ε is arbitrary and thus (3.6.26) and (3.6.27) imply relation (3.6.24). Taking into account (3.6.22), we obtain from (3.6.24) and (3.6.23) that θ_n^* is an asymptotically minimax estimator. □

Remarks concerning the multidimensional case. All the results of this section can be proved for the case of a multidimensional parameter $\theta \in \Theta \subset \mathbf{R}^k$, $k \geq 1$.

In particular, Theorem 3.6.5, one of the main results of this section, is of the following form in the multidimensional case:

$$\mathsf{E}(\theta_n^* - \theta)(\theta_n^* - \theta)' \geq \frac{1 + o(1)}{n} \mathsf{E} I^{-1}(\theta)$$

where $I(t)$ is the Fisher information matrix.

The results for asymptotic Bayes and asymptotic minimax estimators are also valid in the multidimensional case if the quality of an estimator is measured by

$$v(\theta_n^*) = \mathsf{E}(\theta_n^* - t)' V (\theta_n^* - \theta)$$

where V is a certain nonnegative definite matrix.

Bayes and minimax (or asymptotically Bayes and asymptotically minimax) estimators can be defined in the multidimensional case as estimators whose qualities satisfy the corresponding inequalities for all nonnegative definite matrices V.

REMARK 3.6.4. Other approaches to integral inequalities of Cramér–Rao type can be found in Chapter 5 of [**22**] where the estimates of the Shannon information contained in an observation $\xi^{(n)}$ and in an estimator $\widehat{\theta}_n$ with respect to a random parameter θ are used. The corresponding results are called *information inequalities* in [**22**].

CHAPTER 4

Sufficient Statistics

In the preceding section we discussed the problem on how to construct different kinds of optimal estimators, namely Bayes, minimax, efficient, asymptotically Bayes, asymptotically minimax, asymptotically efficient, and others. In this section, we introduce the so-called sufficient estimators that allow one to construct optimal estimators by using a sufficient statistic instead of an observation. Sufficient statistics play an important role in mathematical statistics in general and in the theory of estimation in particular.

4.1. Sufficient statistics and a theorem on factorization

Conditional expectations, conditional probabilities, and sufficient statistics. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, let ξ be a nonnegative random variable, and let $\mathcal{G}$ be a σ-algebra, $\mathcal{G} \subset \mathcal{F}$. A generalized nonnegative random variable $\mathsf{E}(\xi/\mathcal{G})$ (the extended form of this notation is $\mathsf{E}(\xi/\mathcal{G})(\omega)$, $\omega \in \Omega$) is called the *conditional expectation of the random variable ξ with respect to the σ-algebra $\mathcal{G}$* if $\mathsf{E}(\xi/\mathcal{G})$ is $\mathcal{G}$-measurable and for all $A \in \mathcal{G}$

$$\int_A \xi(\omega)\,\mathsf{P}(d\omega) = \int_A \mathsf{E}(\xi/\mathcal{G})(\omega)\,\mathsf{P}(d\omega) \tag{4.1.1}$$

or, equivalently,

$$\mathsf{E} I_A \xi = \mathsf{E} I_A \mathsf{E}(\xi/\mathcal{G}) \quad \text{for all } A \in \mathcal{G} \tag{4.1.2}$$

where $I_A = I_A(\omega)$ is the indicator of the set A. The conditional expectation $\mathsf{E}(\xi/\mathcal{G})$ of a random variable ξ with respect to a σ-algebra $\mathcal{G}$ is well defined if

$$\min\bigl(\mathsf{E}(\xi^+/\mathcal{G}),\, \mathsf{E}(\xi^-/\mathcal{G})\bigr) < \infty \qquad (\mathsf{P}\text{-a.s.}). \tag{4.1.3}$$

In this case

$$\mathsf{E}(\xi/\mathcal{G}) = \mathsf{E}(\xi^+/\mathcal{G}) - \mathsf{E}(\xi^-/\mathcal{G}) \qquad (\mathsf{P}\text{-a.s.}). \tag{4.1.4}$$

Here $\xi^+ = 0 \vee \xi$ and $\xi^- = -(0 \wedge \xi)$. Note that the conditional expectation $\mathsf{E}(\xi/\mathcal{G})$ exists if $\xi \geq 0$. Indeed, let $\mathbf{Q}(A) = \mathsf{E} I_A \xi$, $A \in \mathcal{G}$. Then $\mathbf{Q}$ is a measure on $(\Omega, \mathcal{G})$ and it is absolutely continuous with respect to the measure P. According to the Radon–Nikodym theorem, there exists a generalized nonnegative $\mathcal{G}$-measurable random variable $\mathsf{E}(\xi/\mathcal{G})$ such that

$$\mathbf{Q}(A) = \int_A \mathsf{E}(\xi/\mathcal{G})(\omega)\,\mathsf{P}(d\omega) \quad \text{for all } A \in \mathcal{G}.$$

Thus $\mathsf{E}(\xi/\mathcal{G})(\omega) = d\mathbf{Q}/d\mathsf{P}(\omega)$ (P-a.s.) is the Radon–Nikodym derivative of the measure $\mathbf{Q}$ with respect to the measure P; both measures $\mathbf{Q}$ and P are considered on the space $(\Omega, \mathcal{G})$.

Let $B \in \mathcal{F}$. Then the conditional expectation $\mathsf{E}(I_B/\mathcal{G})$ is called the *conditional probability of an event B with respect to a σ-algebra $\mathcal{G}$, $\mathcal{G} \subset \mathcal{F}$*. The conditional probability is denoted by $\mathsf{P}(B/\mathcal{G})$. Therefore the conditional probability of an event $B \in \mathcal{F}$ with respect to the σ-algebra $\mathcal{G}$ is a $\mathcal{G}$-measurable random variable $\mathsf{P}(B/\mathcal{G})$ such that

$$\mathsf{P}(A \cap B) = \mathsf{E} I_A \mathsf{P}(B/\mathcal{G}) = \int_A \mathsf{P}(B/\mathcal{G})\, d\mathsf{P} \quad \text{for all } A \in \mathcal{G}. \tag{4.1.5}$$

Let ξ be a random variable and let $\mathcal{G}_\eta$ be the σ-algebra generated by some random element η. Then the conditional expectation $\mathsf{E}(\xi/\mathcal{G}_\eta)$, if it exists, is denoted by $\mathsf{E}(\xi/\eta)$ or by $\mathsf{E}(\xi/\eta)(\omega)$ and is called the *conditional expectation of ξ with respect to η*. The conditional probability $\mathsf{P}(B/\mathcal{G}_\eta)$ is denoted by $\mathsf{P}(B/\eta)$ or by $\mathsf{P}(B/\eta)(\omega)$ and is called the *conditional probability of an event $B \in \mathcal{F}$ with respect to η*. Let $\eta = \eta(\omega)$ be a random element assuming values in a measurable space $(Y, \mathcal{S})$. Since $\mathsf{E}(\xi/\eta)$ is a $\mathcal{G}_\eta$-measurable function, there exists a real Borel function $m = m(y)$ defined on $(Y, \mathcal{S})$, assuming values in $\overline{\mathbf{R}} = [-\infty, \infty]$, and such that for all $\omega \in \Omega$

$$m(\eta(\omega)) = \mathsf{E}(\xi/\eta)(\omega).$$

This function $m(y)$ is denoted by $\mathsf{E}(\xi/y) = \mathsf{E}(\xi/\eta = y)$ and is called the *conditional expectation of a random variable ξ with respect to an event $\{\eta = y\}$* or *conditional expectation of ξ given $\eta = y$*.

According to definitions (4.1.1)–(4.1.4) we have

$$\mathsf{E} I_A \xi = \mathsf{E} I_A \mathsf{E}(\xi/\eta) = \mathsf{E} I_A m(\eta) \quad \text{for all } A \in \mathcal{G}_\eta. \tag{4.1.6}$$

Changing the variables in the integral we obtain

$$\mathsf{E} I_{\{\eta \in B\}} m(\eta) = \int_{\{\omega : \eta(\omega) \in B\}} m(\eta)\, d\mathsf{P} = \int_B m(y)\, \mathsf{P}_\eta(dy) \quad \text{for all } B \in \mathcal{S} \tag{4.1.7}$$

where $\{\omega : \eta(\omega) \in B\} \in \mathcal{G}_\eta$ for all $B \in \mathcal{S}$ and P_η is the probability distribution of the random element η on $(Y, \mathcal{S})$. Thus equalities (4.1.6) and (4.1.7) imply that $m = m(y)$ is a Borel function defined on $(Y, \mathcal{S})$ and such that for all $B \in \mathcal{S}$

$$\mathsf{E} I_{\{\eta \in B\}} \xi = \int_{\{\omega : \eta(\omega) \in B\}} \xi(\omega)\, \mathsf{P}(d\omega) = \int_B m(y)\, \mathsf{P}_\eta(dy). \tag{4.1.8}$$

Relation (4.1.8) can be used as an alternative definition of the conditional expectation $\mathsf{E}(\xi/\eta = y) = \mathsf{E}(\xi/y) = m(y)$.

The conditional expectation $\mathsf{E}(I_A/\eta = y)$ is called the *conditional probability of an event $A \in \mathcal{F}$ given $\eta = y$*; this expectation is denoted by

$$\mathsf{P}(A/\eta = y) = \mathsf{P}(A/y).$$

It is clear that the conditional probability $\mathsf{P}(A/\eta = y)$ can be defined as a measurable function defined on $(Y, \mathcal{S})$, assuming values in $([0,1], \mathcal{B}([0,1]))$, and such that for all $B \in \mathcal{S}$

$$\mathsf{P}(A \cap \{\eta \in B\}) = \int_B \mathsf{P}(A/\eta = y)\, \mathsf{P}_\eta(dy) \tag{4.1.9}$$

(see (4.1.5)). Note that the conditional expectation $\mathsf{E}(\xi/\mathcal{G})$ can be defined in a similar way for rather general random elements ξ if the expectation $\mathsf{E}\xi$ exists. A

detailed treatment of this topic as well as a discussion of properties of conditional expectations can be found in [**30**].

Let ξ be a random element assuming values in a measurable space $(X, \mathcal{B})$; let the distribution of ξ be a probability measure belonging to a family $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ is an unknown parameter, $\theta \in \Theta \subset \mathbf{R}^k$, $k \geq 1$. We call ξ an *observation*. An arbitrary measurable function $T = T(x)$ mapping $(X, \mathcal{B})$ into some measurable space $(Y, \mathcal{S})$ is called a *statistic*.

For a fixed $\theta \in \Theta$ consider the probability space $(X, \mathcal{B}, \mathsf{P}_\theta)$. Let $\mathcal{B}_T$ be the σ-algebra in $(X, \mathcal{B})$ generated by the statistic $T = T(x)$. It is clear that

$$\mathcal{B}_T = T^{-1}(\mathcal{S}) \subset \mathcal{B}$$

where $T^{-1}(\mathcal{S})$ is the minimal σ-algebra generated by the family of events

$$\{x: T(x) \in B\}, \qquad B \in \mathcal{S}.$$

According to definition (4.1.5), a $\mathcal{B}_T$-measurable function

$$\mathsf{P}_\theta(A/\mathcal{B}_T) = \mathsf{P}_\theta(A/\mathcal{B}_T)(x)$$

such that

$$\mathsf{P}_\theta(A \cap B) = \int_B \mathsf{P}_\theta(A/\mathcal{B}_T)(x)\, \mathsf{P}_\theta(dx) \quad \text{for all } B \in \mathcal{B}_T \tag{4.1.10}$$

is called the conditional probability measure P_θ of the set $A \in \mathcal{B}$ with respect to the σ-algebra $\mathcal{B}_T$. By definition (4.1.9), the conditional probability measure P_θ of the set $A \in \mathcal{B}$ given $T = y$ is a measurable function

$$\mathsf{P}_\theta(A/\eta = y) = \mathsf{P}_\theta(A/y)$$

defined on $(Y, \mathcal{S})$, assuming values in $([0,1], \mathcal{B}([0,1]))$, and such that for all $B \in \mathcal{S}$

$$\mathsf{P}_\theta(A \cap \{x: T(x) \in B\}) = \int_B \mathsf{P}_\theta(A/y)\, \mathsf{P}_{\theta,T}(dy) \tag{4.1.11}$$

where $\mathsf{P}_{\theta,T}$ is the distribution of the statistic T defined by

$$\mathsf{P}_{\theta,T}(B) = \mathsf{P}_\theta\{x: T(x) \in B\}, \qquad B \in \mathcal{S}.$$

Relations (4.1.10) and (4.1.11) imply that

$$\mathsf{P}_\theta(A/T(x)) = \mathsf{P}_\theta(A/\mathcal{B}_T)(x) \qquad (\mathsf{P}_\theta\text{-a.s.}) \quad \text{for all } A \in \mathcal{B}. \tag{4.1.12}$$

A statistic $T = T(x)$ is called a *sufficient statistic for a family* $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ (or *for a parameter* θ) if for any $A \in \mathcal{B}$ there exists a measurable function

$$\psi_A = \psi_A(y)$$

defined on $(Y, \mathcal{S})$, that depends on A and y and does not depend on θ, and such that

$$\mathsf{P}_\theta(A/T(x)) = \psi_A(T(x)) \qquad (\mathsf{P}_\theta\text{-a.s.}). \tag{4.1.13}$$

This property means that the conditional distribution of the observation ξ given a fixed value of the statistic T does not depend on the parameter θ. This means that the fact that a sampling point $x \in X$ lies on the surface $T(x) = y$ gives no additional information about the parameter θ. In other words, the statistic T exhausts all the information about θ that is contained in the sample. This explains the name for

a *sufficient statistic*: knowledge of $T(x)$ is *sufficient* to construct an estimator of the parameter θ, while the other information included into the observed point x is useless.

EXAMPLE 4.1.1. Let an observation be a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ from a Poisson distribution with parameter θ. Consider a statistic $T_n = \sum_{i=1}^n \xi_i$. This statistic obviously has the Poisson distribution with parameter $n\theta$. Note that

$$\left\{\xi^{(n)} = x, \sum_{i=1}^n \xi_i = y\right\} = \begin{cases} \{\xi^{(n)} = x\}, & \text{if } \sum_{i=1}^n x_i = y, \\ \varnothing, & \text{if } \sum_{i=1}^n x_i \neq y \end{cases}$$

where $x = (x_1, \dots, x_n)$ and $x_i, y \in \{0, 1, \dots\}$ for all $i = 1, \dots, n$. Then

$$\mathsf{P}_\theta\left\{\xi^{(n)} = x / T_n = y\right\} = \begin{cases} \frac{\mathsf{P}_\theta\{\xi^{(n)} = x\}}{\mathsf{P}_\theta\{T_n = y\}}, & \text{if } \sum_{i=1}^n x_i = y, \\ 0, & \text{if } \sum_{i=1}^n x_i \neq y. \end{cases} \tag{4.1.14}$$

If $\sum_{i=1}^n x_i = y$, then

$$\frac{\mathsf{P}_\theta\{\xi^{(n)} = x\}}{\mathsf{P}_\theta\{T_n = y\}} = \left(e^{-n\theta}\frac{(n\theta)^y}{y!}\right)^{-1} \prod_{i=1}^n e^{-\theta}\frac{\theta^{x_i}}{x_i!} = \frac{y!}{n^y}\prod_{i=1}^n \frac{1}{x_i!}.$$

Thus relation (4.1.14) implies that the conditional probability

$$\mathsf{P}_\theta\left(\xi^{(n)} = x / T_n = y\right)$$

does not depend on the parameter θ. This means that the statistic T_n is sufficient according to definition (4.1.13).

Relation (4.1.12) suggests the following definition of a sufficient σ-algebra. We say $\mathcal{B}_T$ is a *sufficient σ-algebra* if the statistic T is sufficient. The notion of a sufficient statistic is important in many problems, however the notion of a σ-algebra is more convenient, at least from the point of view of the theory, than that of a sufficient statistic. Note that there are examples of sufficient σ-algebras that are not generated by any sufficient statistics assuming values in a given measurable space [3].

Dominated families of distributions. Let $\mathcal{P}$ be a family of probability measures defined on a measurable space $(X, \mathcal{B})$ and let μ be some σ-finite measure on $(X, \mathcal{B})$. We say that a family $\mathcal{P}$ is *dominated by the measure μ* if every measure $\mathsf{P} \in \mathcal{P}$ is absolutely continuous with respect to μ. A family $\mathcal{P}$ is called *dominated* if there is a σ-finite measure dominating the family $\mathcal{P}$. Note that if a family $\mathcal{P}$ is dominated, then there exists a finite dominating measure. Indeed, let a family $\mathcal{P}$ be dominated by a σ-finite measure μ and let $X = \bigcup_{i=1}^\infty A_i$ where $\mu(A_i) < \infty$ for all $i = 1, 2, \dots$ and $A_i \cap A_j = \varnothing$ for $i \neq j$. Then the measure ν defined by

$$\nu(A) = \sum_{i=1}^\infty 2^{-i}\mu(A \cap A_i)/\mu(A_i), \qquad A \in \mathcal{B},$$

also dominates the family $\mathcal{P}$; moreover the measure ν is finite.

Let two families of measure $\mathcal{M}$ and $\mathcal{N}$ be given. We say that a *family $\mathcal{M}$ is dominated by a family $\mathcal{N}$* if every measure of the family $\mathcal{N}$ dominates the family $\mathcal{M}$. The families of measures $\mathcal{M}$ and $\mathcal{N}$ are called *equivalent* if the family $\mathcal{M}$ is dominated by the family $\mathcal{N}$, and the family $\mathcal{N}$ is dominated by the family $\mathcal{M}$.

THEOREM 4.1.1. *A family of probability measures $\mathcal{P}$ is dominated a σ-finite measure if and only if the family $\mathcal{P}$ contains a countable equivalent subfamily.*

PROOF. First we assume that the family $\mathcal{P}$ contains a countable equivalent subfamily $\{\mathsf{P}_1, \mathsf{P}_2, \dots\}$. Then the family $\mathcal{P}$ is dominated by the measure

$$\mu = \sum_{n=1}^{\infty} 2^{-n} \mathsf{P}_n.$$

Conversely let the family $\mathcal{P}$ be dominated by a σ-finite measure μ. Without loss of generality we assume that the measure μ is finite. Let $\mathcal{K}$ be the class of all probability measures $\mathbf{Q}$ of the form $\sum c_i \mathsf{P}_i$ where $\mathsf{P}_i \in \mathcal{P}$, all numbers c_i are positive, and $\sum c_i = 1$. The class $\mathcal{K}$ is dominated by the measure μ. For $\mathbf{Q} \in \mathcal{K}$ we denote by $q(x) = d\mathbf{Q}/d\mu(x)$ the density of the measure $\mathbf{Q}$ with respect to μ.

Our current goal is to prove the following assertion (which is equivalent to the statement of the theorem): *there exists a measure $\mathbf{Q}_0 \in \mathcal{K}$ such that the equality $\mathbf{Q}_0(A) = 0$ implies $\mathbf{Q}(A) = 0$ for all $\mathbf{Q} \in \mathcal{K}$.*

Consider the class S of sets $C \in \mathcal{B}$ for which there exists a measure $\mathbf{Q} \in \mathcal{K}$ such that $q(x) > 0$ almost surely with respect to the measure μ on a set C with $\mathbf{Q}(C) > 0$. Let $\mu(C_i) \to \sup_{C \in S} \mu(C)$ as $i \to \infty$ where $C_i \in S$ and $q_i(x) > 0$ almost surely with respect to the measure μ on a set C_i ($q_i(x)$ corresponds to $\mathbf{Q}_i$ which in turn corresponds to C_i). Let $C_0 = \bigcup_{i=1}^{\infty} C_i$. Then $q_0^*(x) = \sum_{i=1}^{\infty} c_i q_i(x)$ coincides μ-almost surely with the density $d\mathbf{Q}_0/d\mu(x)$ where $\mathbf{Q}_0 = \sum_{i=1}^{\infty} c_i \mathbf{Q}_i$. It is clear that $q_0^*(x) > 0$ almost surely with respect to the measure μ on the set C_0, whence $C_0 \in S$.

Assume that $\mathbf{Q}_0(A) = 0$. Let $\mathbf{Q}$ be another measure of the class $\mathcal{K}$ and let $C = \{x: q(x) > 0\}$ and $q(x) = d\mathbf{Q}/d\mu(x)$. Then $\mathbf{Q}_0(A \cap C_0) = 0$, whence $\mu(A \cap C_0) = 0$ and $\mathbf{Q}(A \cap C_0) = 0$. We also have $\mathbf{Q}(A \cap \overline{C}_0 \cap \overline{C}) = 0$ where $\overline{C} = X \setminus C$ and $\overline{C}_0 = X \setminus C_0$. Now we prove that $\mathbf{Q}(A \cap \overline{C}_0 \cap C) = 0$. Assume the converse, namely let $\mathbf{Q}(A \cap \overline{C}_0 \cap C) > 0$. This implies

$$\mu(C_0 \cup (A \cap \overline{C}_0 \cap C)) = \mu(C_0) + \mu(A \cap \overline{C}_0 \cap C) > \mu(C_0), \tag{4.1.15}$$

since $\mu(A \cap \overline{C}_0 \cap C) > 0$ in view of the inequality $\mathbf{Q}(A \cap \overline{C}_0 \cap C) > 0$ and $\mathbf{Q}$ is absolutely continuous with respect to μ. Inequality (4.1.15) contradicts the equality $\mu(C_0) = \sup_{C \in S} \mu(C)$. Thus $\mathbf{Q}(A \cap \overline{C}_0 \cap C) = 0$ and

$$\mathbf{Q}(A) = \mathbf{Q}(A \cap C_0) + \mathbf{Q}(A \cap \overline{C}_0 \cap C) + \mathbf{Q}(A \cap \overline{C}_0 \cap \overline{C}) = 0.$$

Therefore the equality $\mathbf{Q}_0(A) = 0$ implies that $\mathbf{Q}(A) = 0$ for all $\mathbf{Q} \in \mathcal{K}$. □

Theorem on the factorization. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$ and whose distribution belongs to a family of probability measures $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ where $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$ is an unknown parameter $\theta \in \Theta \subset \mathbf{R}^k$, $k \geq 1$.

The following result, known as the *Neyman–Fisher factorization criterion*, contains a necessary and sufficient condition for a statistic to be sufficient for a dominated family $\mathcal{P}$. The short name of this result is the *factorization criterion*. The first result of this type is obtained by Fisher (1920) and rediscovered by Neyman (1935). It is proved for general dominated families by Halmos and Savage (1949). Further generalization is due to Bahadur (1954). The result below is closer to the theorem of Bahadur.

THEOREM 4.1.2. *Let a family $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ be dominated by a measure μ and let $T = T(x)$ be a statistic mapping $(X, \mathcal{B})$ into a measurable space $(Y, \mathcal{S})$. The statistic T is sufficient for a family $\mathcal{P}$ if and only if the density*

$$f(x;\theta) = d\mathsf{P}_\theta/d\mu(x)$$

admits the factorization

$$f(x;\theta) = g(T(x);\theta)r(x) \qquad (\mu\text{-a.s.}) \quad \text{for all } \theta \in \Theta \tag{4.1.16}$$

where $g(y;\theta)$ is a nonnegative $\mathcal{S}$-measurable function for all $\theta \in \Theta$ and $r(x)$ is a nonnegative $\mathcal{B}$-measurable function.

PROOF. Since the family $\mathcal{P}$ is dominated, Theorem 4.1.1 implies that there is a countable subfamily $\{\mathsf{P}_{\theta_1}, \mathsf{P}_{\theta_2}, \dots\} \subset \mathcal{P}$ that is equivalent to the family $\mathcal{P}$. Consider the probability measure $\lambda = \sum_i c_i \mathsf{P}_{\theta_i}$ where $c_i > 0$ for all $i = 1, 2, \dots$ and $\sum_i c_i = 1$. It is obvious that the measure λ dominates the family $\mathcal{P}$. Denote by $p(x;\theta) = d\mathsf{P}_\theta/d\lambda(x)$ the density of the measure P_θ with respect to the measure λ.

Necessity. Let T be a sufficient statistic for the family $\mathcal{P}$. According to definition (4.1.13), for every $A \in \mathcal{B}$ there exists an $\mathcal{S}$-measurable function $\psi_A(y)$ defined on $(Y, \mathcal{S})$ and such that $\mathsf{P}_\theta(A/T(x)) = \psi_A(T(x))$ almost surely with respect to the measure P_θ for all $\theta \in \Theta$. Then $\lambda(A/T(x)) = \psi_A(T(x))$ almost surely with respect to the measure λ. Indeed, the definition of the conditional probability implies for all $A \in \mathcal{B}$ and $B \in \mathcal{B}_T = T^{-1}(\mathcal{S})$ that

$$\begin{aligned}\lambda(A \cap B) &= \sum_i c_i \mathsf{P}_{\theta_i}(A \cap B) = \sum_i c_i \int_B \mathsf{P}_{\theta_i}(A/T(x))\, \mathsf{P}_{\theta_i}(dx) \\ &= \sum_i c_i \int_B \psi_A(T(x))\, \mathsf{P}_{\theta_i}(dx) = \int_B \psi_A(T(x))\, \lambda(dx)\end{aligned}$$

which together with (4.1.10)–(4.1.12) implies that $\lambda(A/T(x)) = \psi_A(T(x))$ almost surely with respect to the measure λ for all $A \in \mathcal{B}$. Further, for all $A \in \mathcal{B}$ we have

$$\begin{aligned}\mathsf{E}_\lambda I_A(\xi)p(\xi;\theta) &= \mathsf{P}_\theta(A) = \mathsf{E}_\theta \mathsf{P}_\theta(A/T(\xi)) \\ &= \mathsf{E}_\theta \psi_A(T(\xi)) = \mathsf{E}_\lambda p(\xi;\theta)\psi_A(T(\xi)) \\ &= \mathsf{E}_\lambda p(\xi;\theta)\lambda(A/T(\xi)) = \mathsf{E}_\lambda \lambda(A/T(\xi))\mathsf{E}_\lambda(p(\xi;\theta)/T(\xi)) \\ &= \mathsf{E}_\lambda \mathsf{E}_\lambda(I_A(\xi)/T(\xi))\mathsf{E}_\lambda(p(\xi;\theta)/T(\xi)) \\ &= \mathsf{E}_\lambda \mathsf{E}_\lambda\big(I_A(\xi)\mathsf{E}_\lambda(p(\xi;\theta)/T(\xi))/T(\xi)\big) \\ &= \mathsf{E}_\lambda I_A(\xi)\mathsf{E}_\lambda(p(\xi;\theta)/T(\xi))\end{aligned} \tag{4.1.17}$$

where E_λ and E_θ are the expectations with respect to the measures λ and P_θ, respectively. This implies that $p(x;\theta) = \mathsf{E}_\lambda(p(\xi;\theta)/T(x))$ almost surely with respect to the measure λ, that is, the density $p(x;\theta)$ is $\mathcal{B}_T$-measurable or, in other words, $p(x;\theta) = g(T(x);\theta)$ almost surely with respect to the measure λ. Since $\lambda \ll \mu$, equality (4.1.17) implies that for all $A \in \mathcal{B}$

$$\begin{aligned}\int_A f(x;\theta)\, \mu(dx) &= \mathsf{P}_\theta(A) = \mathsf{E}_\lambda I_A(\xi) g(T(\xi);\theta) \\ &= \int_A g(T(x);\theta) r(x)\, \mu(dx)\end{aligned} \tag{4.1.18}$$

where $r(x) = d\lambda/d\mu(x)$. Since the set $A \in \mathcal{B}$ is arbitrary, equality (4.1.18) implies the required relation (4.1.16).

Sufficiency. Let relation (4.1.16) hold. Then

$$\begin{aligned}\frac{d\lambda}{d\mu}(x) &= \sum_i c_i f(x;\theta_i) = r(x)\sum_i c_i g(T(x);\theta_i) \\ &= r(x)G(T(x)) \qquad (\mu\text{-a.s.})\end{aligned} \tag{4.1.19}$$

where

$$G(T(x)) = \sum_i c_i g(T(x);\theta_i), \qquad x \in X.$$

Consider the function

$$\widetilde{p}(x;\theta) = \begin{cases} \frac{g(T(x);\theta)}{G(T(x))}, & \text{if } G(T(x)) > 0, \\ 0, & \text{if } G(T(x)) = 0. \end{cases} \tag{4.1.20}$$

It is clear that the function $\widetilde{p}(x;\theta)$ is $\mathcal{B}_T$-measurable. Moreover

$$\frac{d\mathsf{P}_\theta}{d\mu}(x) = \frac{d\mathsf{P}_\theta}{d\lambda}(x) \cdot \frac{d\lambda}{d\mu}(x) \qquad (\mu\text{-a.s.}),$$

since $\mathsf{P}_\theta \ll \lambda \ll \mu$.

Therefore

$$\frac{d\mathsf{P}_\theta}{d\lambda}(x) = \frac{d\mathsf{P}_\theta}{d\mu}(x) \Big/ \frac{d\lambda}{d\mu}(x) = \frac{g(T(x);\theta)}{G(T(x))} \qquad (\mu\text{-a.s.}) \tag{4.1.21}$$

by (4.1.16) and (4.1.19). Equality (4.1.19) implies that $\lambda\{x\colon G(T(x)) = 0\} = 0$. Thus $p(x;\theta) = \widetilde{p}(x;\theta)$ (λ-a.s.) by (4.1.20) and (4.1.21), since λ is absolutely continuous with respect to μ. The function $\widetilde{p}(x;\theta)$ is $\mathcal{B}_T$-measurable, whence for all $A \in \mathcal{B}$ we have

$$\begin{aligned}\mathsf{E}_\theta \mathsf{P}_\theta(A/T(\xi)) &= \mathsf{P}_\theta(A) = \mathsf{E}_\lambda I_A(\xi) p(\xi;\theta) \\ &= \mathsf{E}_\lambda \mathsf{E}_\lambda\big(I_A(\xi)p(\xi;\theta)/T(\xi)\big) = \mathsf{E}_\lambda p(\xi;\theta)\mathsf{E}_\lambda(I_A(\xi)/T(\xi)) \\ &= \mathsf{E}_\theta \mathsf{E}_\lambda(I_A(\xi)/T(\xi)) = \mathsf{E}_\theta \lambda(A/T(\xi)).\end{aligned}$$

If A is replaced with $A \cap B$ in this equality and if $A \in \mathcal{B}$ and $B \in \mathcal{B}_T$, then

$$\mathsf{E}_\theta I_B(\xi)\mathsf{P}_\theta(A/T(\xi)) = \mathsf{E}_\theta I_B(\xi)\lambda(A/T(\xi)).$$

Since $B \in \mathcal{B}_T$ is arbitrary, it follows that for all $A \in \mathcal{B}$

$$\mathsf{P}_\theta(A/T(x)) = \lambda(A/T(x)) \qquad (\mathsf{P}_\theta\text{-a.s.}) \quad \text{for all } \theta \in \Theta.$$

This means that the statistic $T(x)$ is sufficient. □

Below are two corollaries of Theorem 4.1.2.

COROLLARY 4.1.1. *Let a family $\mathcal{P}$ be dominated. If a measurable function of some statistic is sufficient for the family $\mathcal{P}$, then the statistic itself is sufficient for this family, too.*

PROOF. Let T and $\widetilde{T}$ be two statistics such that $T = \phi(\widetilde{T})$ where ϕ is a measurable function. Assume that T is a sufficient statistic for the family $\mathcal{P}$. According to Theorem 4.1.2 we have

$$f(x;\theta) = g(T(x);\theta)r(x) = \tilde{g}\big(\widetilde{T}(x);\theta\big)r(x) \qquad (\mu\text{-a.s.})$$

(see relation (4.1.16)), that is, the statistic $\widetilde{T}$ is also sufficient for the family $\mathcal{P}$. □

COROLLARY 4.1.2. *Let a family $\mathcal{P}$ be dominated. If T is a sufficient statistic for the family $\mathcal{P}$ and a function ϕ is such that $v = \phi(y)$ is a measurable one-to-one mapping, then the statistic $\widetilde{T} = \phi(T)$ is also sufficient for the family $\mathcal{P}$.*

PROOF. Since ϕ is a one-to-one mapping, we have $T = \phi^{-1}(\widetilde{T})$. Now Corollary 4.1.2 follows from Corollary 4.1.1. □

Applying the factorization criterion one can find sufficient statistics for dominated families. Below are some applications of the factorization criterion for obtaining sufficient statistics.

Examples of sufficient statistics. According to the factorization criterion the statistic $T(x) = x$, called the *trivial sufficient statistic*, is sufficient for every dominated family of probability measures $\mathcal{P}$. In the examples below we find nontrivial sufficient statistics.

EXAMPLE 4.1.2. Let $\Theta = \{\theta_1, \theta_2, \dots, \theta_s\}$. Then any family $\mathcal{P}$ is finite, that is, $\mathcal{P} = \{\mathsf{P}_{\theta_1}, \mathsf{P}_{\theta_2}, \dots, \mathsf{P}_{\theta_s}\}$. A dominating measure μ exists in this case. In particular, one can put $\mu = \sum_{i=1}^{s} c_i \mathsf{P}_{\theta_i}$ for $c_i > 0$, $i = 1, 2, \dots, s$. Consider the statistic

$$T(x) = (T_1(x), T_2(x), \dots, T_{s-1}(x))$$

where $T_j(x) = \ln(f(x;\theta_{j+1})/f(x;\theta_1))$, $j = 1, 2, \dots, s-1$. Then

$$f(x;\theta_j) = \exp(T_{j-1}(x))r(x), \qquad j = 2, 3, \dots, s,$$

where $r(x) = f(x;\theta_1)$. Thus

$$f(x;\theta) = g(T(x);\theta)r(x) \quad \text{for all } \theta \in \Theta$$

where

$$g(T(x);\theta) = \begin{cases} \exp(T_{j-1}(x)), & \text{if } \theta = \theta_j,\ j = 2, 3, \dots, s, \\ 1, & \text{if } \theta = \theta_1. \end{cases}$$

Therefore $T(x) = (T_1(x), T_2(x), \dots, T_{s-1}(x))$ is a sufficient statistic.

EXAMPLE 4.1.3. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the normal distribution $\mathcal{N}(\theta_1, \theta_2^2)$. Then the density $f(x;\theta)$, $\theta = (\theta_1, \theta_2)$, is represented in the form

$$\begin{aligned} f(x;\theta) &= \frac{1}{(\sqrt{2\pi}\theta_2)^n} \exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2\right\} \\ &= \frac{1}{(\sqrt{2\pi}\theta_2)^n} \exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n} x_i^2 + \frac{\theta_1}{\theta_2^2}\sum_{i=1}^{n} x_i - \frac{n\theta_1^2}{2\theta_2^2}\right\} \end{aligned}$$

where $x = (x_1, \ldots, x_n)$. This means that the statistic $T(x) = (T_1(x), T_2(x))$, where

$$T_1(x) = \sum_{i=1}^{n} x_i, \qquad T_2(x) = \sum_{i=1}^{n} x_i^2$$

is sufficient. Note that according to Corollary 4.1.1, the statistic $\widetilde{T}(x) = (\widetilde{T}_1(x), \widetilde{T}_2(x))$, where

$$\widetilde{T}_1(x) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widetilde{T}_2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

is also sufficient because $T(x) = \phi(\widetilde{T}(x))$.

If $\xi^{(n)}$ is a sample from the distribution $\mathcal{N}(\theta_1, \sigma^2)$ and the variance σ^2 is known, then $T(x) = \sum_{i=1}^{n} x_i$ is a sufficient statistic, while if $\xi^{(n)}$ is a sample from the distribution $\mathcal{N}(\tilde{a}, \theta_2^2)$ and the expectation $\tilde{a}$ is known, then $T(x) = \sum_{i=1}^{n}(x_i - \tilde{a})^2$ is a sufficient statistic. It is clear that the statistic $T(x) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$ is also sufficient in both cases.

EXAMPLE 4.1.4. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ where $\xi_1, \ldots, \xi_n$ are independent random variables. Assume that ξ_i has the normal $\mathcal{N}(\theta_0, \theta_i^2)$ distribution, $i = 1, \ldots, n$. Then the density $\xi^{(n)}$ is of the form

$$f(x;\theta) = \prod_{i=1}^{n} \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i}{\theta_i}\right)^2 + \theta_0 \sum_{i=1}^{n}\frac{x_i}{\theta_i^2} - \frac{\theta_0^2}{2}\sum_{i=1}^{n}\frac{1}{\theta_i^2}\right\}}{\sqrt{2\pi}\theta_i}$$

where $\theta = (\theta_0, \ldots, \theta_n)$ is an unknown vector parameter. It is clear that the "best" sufficient statistic in this case is $T(x) = x$. If the random variable ξ_i has the normal $\mathcal{N}(\theta, \sigma_i^2)$ distribution and the variances σ_i^2 are known, then $T(x) = \sum_{i=1}^{n} x_i/\sigma_i^2$ is a sufficient statistic.

EXAMPLE 4.1.5. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample from an uniform distribution on the interval $[0, \theta]$ where $\theta > 0$ is an unknown parameter. Then the density of the sample $\xi^{(n)}$ is of the form

$$f(x;\theta) = \theta^{-n} I_{[0,\infty)}(x_{n,1}) I_{(-\infty,\theta]}(x_{n,n})$$

where

$$x_{n,1} = \min_{1 \le i \le n} x_i, \qquad x_{n,n} = \max_{1 \le i \le n} x_i, \qquad (x_1, \ldots, x_n) = x.$$

Thus relation (4.1.16) holds with $g(t;\theta) = \theta^{-n} I_{(-\infty,\theta]}(t)$, $r(x) = I_{[0,\infty)}(x_{n,1})$, and $T(x) = x_{n,n}$, whence it follows that $T(x) = x_{n,n}$ is a sufficient statistic.

If $\xi^{(n)}$ is a sample from the uniform distribution on the interval $[\theta, \theta + 1]$ and $\theta \in \mathbf{R}$ is an unknown parameter, then we can proceed in the same way as above to show that $T(x) = (x_{n,1}, x_{n,n})$ is a sufficient statistic of the parameter θ. The same statistic is sufficient for the two-dimensional parameter $\theta = (\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$, in the case of a sample from the uniform distribution on the interval $[\theta_1, \theta_2]$.

EXAMPLE 4.1.6. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the Pearson type III distribution whose density is given by

$$\frac{\theta_2^{\theta_3}}{\Gamma(\theta_3)}(x-\theta_1)^{\theta_3-1}e^{-\theta_2(x-\theta_1)}I_{[\theta_1,\infty)}(x), \qquad x \in (-\infty, \infty),$$

where $(\theta_1, \theta_2, \theta_3) = \theta$ is an unknown parameter. Then the distribution density of the sample $\xi^{(n)}$ is

$$f(x;\theta) = \frac{\theta_2^{n\theta_3}}{\Gamma^n(\theta_3)}\prod_{i=1}^{n}(x_i-\theta_1)^{\theta_3-1}\exp\left\{-\theta_2\sum_{i=1}^{n}(x_i-\theta_1)\right\}I_{[\theta_1,\infty)}(x_{n,1})$$

where $x = (x_1, \dots, x_n)$. If $\theta_3 = 1$ and an unknown parameter is $\theta = (\theta_1, \theta_2)$, then the density is of the form

$$f(x;\theta) = \theta_2^n \exp\left\{-\theta_2\sum_{i=1}^{n}(x_i-\theta_1)\right\}I_{[\theta_1,\infty)}(x_{n,1}).$$

Thus the statistic $T(x) = (\sum_{i=1}^n x_i, x_{n,1})$ is sufficient. If $\theta_1 = 0$ and $\theta = (\theta_2, \theta_3)$ is an unknown parameter, then $T(x) = (\sum_{i=1}^n x_i, \prod_{i=1}^n x_i)$ is a sufficient statistic. If $\theta_3 \neq 1$ is either a known or unknown parameter, while θ_1 is an unknown parameter, then the "simplest" sufficient statistic is $T(x) = (x_1, x_2, \dots, x_n)$.

Note that the density of the Pearson type III distribution belongs to the system of Pearson curves (its description can be found in [**9**], §19.4, or in [**24**], §5.6).

Sufficient statistics in the Bayes approach. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$. Let its distribution depend on an unknown parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$ which is random with the distribution $\mathbf{Q}$ concentrated on a Borel set $\Theta \subset \mathbf{R}^k$, $k \geq 1$. Let P_t, $t \in \Theta$, be probability measures corresponding to the conditional distribution of the observation ξ given $\theta = t$, that is, $\mathsf{P}_t(A) = \mathsf{P}\{\xi \in A/\theta = t\}$ for all $A \in \mathcal{B}$ and $t \in \Theta$. Thus we are given a family of probability measures $\mathcal{P} = (\mathsf{P}_t, t \in \Theta)$ which we assume to be dominated by some measure μ. We also assume that the measure $\mathbf{Q}$ possesses the density $q(t)$ with respect to some measure λ.

The following result contains necessary and sufficient conditions that a statistic is sufficient for a family $\mathcal{P} = (\mathsf{P}_t, t \in \Theta)$ expressed in terms of a posteriori density.

THEOREM 4.1.3. *A statistic $T = T(x)$ mapping $(X, \mathcal{B})$ into some measurable space $(Y, \mathcal{S})$ is sufficient for a family $\mathcal{P} = (\mathsf{P}_t, t \in \Theta)$ if and only if for any a priori distribution $\mathbf{Q}$ of the parameter θ, the a posteriori distribution $\mathbf{Q}_x$ depends on x through $T(x)$. Here $\mathbf{Q}_x(A) = \mathsf{P}\{\theta \in A/\xi = x\}$, $x \in X$, $A \in \mathcal{B}^k$.*

PROOF. Let T be a sufficient statistic for a family $\mathcal{P}$ and let $q(t)$ be the density of the measure $\mathbf{Q}$ with respect to the measure λ. The density $q(t/x)$ of the a posteriori measure $\mathbf{Q}_x$ with respect to the measure λ exists for all $x \in X$. Applying relation (4.1.16) we obtain from the Bayes theorem that

$$q(t/x) = \frac{f(x;t)q(t)}{\int f(x;u)q(u)\,\lambda(du)} = \frac{g(T(x);t)q(t)}{\int g(T(x);u)q(u)\,\lambda(du)},$$

that is, $q(t/x)$ depends on x through $T(x)$.

Now we prove the converse. Consider an a priori distribution such that $q(t) > 0$ everywhere on Θ and for all t

$$f(x;t) = \frac{q(t/x)f(x)}{q(t)}, \qquad f(x) = \int f(x;u)q(u)\,\lambda(du).$$

Put $q(t/x) = \psi(t, T(x))$. Setting $g(y;t) = \psi(t,y)/q(t)$ and $r(x) = f(x)$ we get relation (4.1.16), that is, $T(x)$ is a sufficient statistic. □

COROLLARY 4.1.3. *If $T = T(x)$ is a sufficient statistic for a family*

$$\mathcal{P} = (\mathsf{P}_t, t \in \Theta),$$

then any Bayes estimator, as well as any minimax estimator, of the parameter θ with respect to the quadratic loss function defined as in Theorem 3.1.3 *depend on the statistic T.*

PROOF. Note that the Bayes estimator of the parameter θ with respect to the quadratic loss function is the a posteriori expectation

$$\mathsf{E}(\theta/\xi = x) = \int tq(t/x)\,\lambda(dt).$$

Now we apply Theorem 4.1.3 to complete the proof. □

REMARK 4.1.1. Using Theorem 4.1.3 one can provide the following equivalent definition, called the *Bayes definition of a sufficient statistic* (see [**36**]). A statistic $T = T(x)$ is sufficient for a family $\mathcal{P} = (\mathsf{P}_t, t \in \Theta)$ if for any a priori distribution $\mathbf{Q}$ of the parameter θ, the a posteriori distribution $\mathbf{Q}_x$ depends on x through $T(x)$ almost surely with respect to μ.

Fisher information and sufficient statistics. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$. Let its distribution belong to a family $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ where θ is a nonrandom scalar unknown parameter. We assume that the family $\mathcal{P}$ is dominated by some σ-finite measure μ whose density $f(x;\theta) = d\mathsf{P}_\theta/d\mu(x)$ satisfies an arbitrary regularity condition under which the Fisher information $I(\theta)$ is well defined. For the sake of definiteness we assume the regularity conditions (CR) (see Section 3.4). Then $I(\theta) = \mathsf{E}_\theta S^2(\xi;\theta)$ where $S(x;\theta) = \partial \ln f(x;\theta)/\partial\theta$. Note that conditions (CR) imply that

$$\mathsf{E}_\theta S(\xi;\theta) = 0$$

(see Lemma 3.4.1), whence it follows that $I(\theta) = \mathsf{D}_\theta S(\xi;\theta)$ is the variance of the random variable $S(\xi;\theta)$.

Let $T = T(x)$ be some statistic mapping $(X, \mathcal{B})$ into some measurable space $(Y, \mathcal{S})$. Denote by μ^T and P_θ^T, $\theta \in \Theta$, the images of measures μ and P_θ, $\theta \in \Theta$, respectively, under the mapping $T\colon (X, \mathcal{B}) \to (Y, \mathcal{S})$, that is,

$$\mathsf{P}_\theta^T(B) = \mathsf{P}_\theta\big(T^{-1}(B)\big) \quad \text{and} \quad \mu^T(B) = \mu\big(T^{-1}(B)\big) \quad \text{for all } B \in \mathcal{S}$$

where $T^{-1}(B) = \{x\colon T(x) \in B\}$ is the preimage of the set B under the mapping T. It is clear that the family of measures $\mathcal{P}^T = (\mathsf{P}_\theta^T, \theta \in \Theta)$ is dominated by the measure μ^T. Denote by $g(y;\theta) = d\mathsf{P}_\theta^T/d\mu^T(y)$, $y \in Y$, the density of the measure P_θ^T with respect to the measure μ^T. If the density $g(y;\theta)$ also satisfies conditions (CR), then the Fisher information $I^T(\theta) = \mathsf{E}_\theta(\partial \ln g(T(\xi);\theta)/\partial\theta)^2$ is well defined.

The following result provides a relation between the Fisher information $I^T(\theta)$ generated by the distribution of the statistic $T(\xi)$ and the Fisher information $I(\theta)$ generated by the distribution of the observation ξ.

THEOREM 4.1.4. *Let a family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ be dominated by a measure μ and satisfy the regularity conditions (CR). Let $T = T(x)$ be a statistic mapping $(X, \mathcal{B})$ into a measurable space $(Y, \mathcal{S})$. Assume that the family of probability measures $\mathcal{P}^T = (\mathsf{P}_\theta^T, \theta \in \Theta)$ is generated by T on $(Y, \mathcal{S})$ and satisfies the regularity conditions (CR). Then*

$$I^T(\theta) \le I(\theta) \quad \text{for all } \theta \in \Theta. \tag{4.1.22}$$

Moreover inequality (4.1.22) *becomes an equality if and only if the statistic T is sufficient for the family $\mathcal{P}$.*

PROOF. Let C be an arbitrary set of $\mathcal{S}$. According to the definition of the conditional expectation and in view of condition (iii) in Section 3.4 we obtain from (CR) that for all $\theta \in \Theta$

$$\int_{T^{-1}(C)} S(x;\theta)\, \mathsf{P}_\theta(dx) = \int_C \mathsf{E}_\theta\{S(\xi;\theta)/T(\xi) = y\}\, \mathsf{P}_\theta^T(dy). \tag{4.1.23}$$

Since conditions (CR) holds for both families $\mathcal{P}$ and $\mathcal{P}^T$, we have for any P_θ^T-nonzero set $C \in \mathcal{S}$ that

$$\begin{aligned} \int_{T^{-1}(C)} S(x;\theta)\, \mathsf{P}_\theta(dx) &= \int_{T^{-1}(C)} \frac{\partial}{\partial\theta} f(x;\theta)\, \mu(dx) \\ &= \frac{\partial}{\partial\theta} \int_{T^{-1}(C)} f(x;\theta)\, \mu(dx) = \frac{\partial}{\partial\theta} \int_C g(y;\theta)\, \mu^T(dy) \\ &= \int_C \frac{\partial}{\partial\theta} g(y;\theta)\, \mu^T(dy) = \int_C \frac{\partial}{\partial\theta} \ln g(y;\theta)\, \mathsf{P}_\theta^T(dy). \end{aligned} \tag{4.1.24}$$

Relations (4.1.23) and (4.1.24) yield

$$\frac{\partial}{\partial\theta} \ln g(y;\theta) = \mathsf{E}_\theta\{S(\xi;\theta)/T(\xi) = y\} \qquad (\mathsf{P}_\theta^T\text{-a.s.}) \tag{4.1.25}$$

for all $\theta \in \Theta$. By definition we have $I^T(\theta) = \mathsf{E}_\theta(\partial \ln g(T(\xi);\theta)/\partial\theta)^2$. To prove the inequality $I^T(\theta) \le I(\theta)$, note that

$$\begin{aligned} 0 &\le \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln f(\xi;\theta) - \frac{\partial}{\partial\theta} \ln g(T(\xi);\theta) \right)^2 \\ &= E_\theta \left(\frac{\partial}{\partial\theta} \ln f(\xi;\theta) \right)^2 + \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln g(T(\xi);\theta) \right)^2 \\ &\quad - 2\mathsf{E}_\theta \frac{\partial}{\partial\theta} \ln f(\xi;\theta) \frac{\partial}{\partial\theta} \ln g(T(\xi);\theta). \end{aligned} \tag{4.1.26}$$

It follows from (4.1.25) that

$$\begin{aligned} &\mathsf{E}_\theta \frac{\partial}{\partial\theta} \ln f(\xi;\theta) \frac{\partial}{\partial\theta} \ln g(T(\xi);\theta) \\ &\quad = \mathsf{E}_\theta \frac{\partial}{\partial\theta} \ln g(T(\xi);\theta) \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln f(\xi;\theta)/T(\xi) \right) \\ &\quad = \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln g(T(\xi);\theta) \right)^2 . \end{aligned} \tag{4.1.27}$$

Substituting (4.1.27) into (4.1.26) we prove inequality (4.1.22). It remains to show that the inequality (4.1.22) becomes an equality if and only if the statistic T is sufficient for the family $\mathcal{P}$.

If T is a sufficient statistic for $\mathcal{P}$, then by the factorization criterion

$$f(x;\theta) = g^*(T(x);\theta)h(x) \qquad (\mu\text{-a.s.}) \tag{4.1.28}$$

where $h(x) \geq 0$, $g^*(y;\theta) \geq 0$, the function $h(x)$ is $\mathcal{B}$-measurable, and $g^*(y;\theta)$ is $\mathcal{S}$-measurable for all $\theta \in \Theta$. Thus for all $\theta \in \Theta$

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) = \frac{\partial}{\partial\theta} \ln g^*(T(x);\theta) \qquad (\mathsf{P}_\theta\text{-a.s.}).$$

Therefore

$$I(\theta) = \mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln g^*(T(\xi);\theta) \right)^2 \quad \text{for all } \theta \in \Theta. \tag{4.1.29}$$

On the space $(Y, \mathcal{S})$ consider the measure

$$\lambda(B) = \int_{T^{-1}(B)} h(x)\, \mu(dx), \qquad B \in \mathcal{S}.$$

According to (4.1.28) we have for all $B \in \mathcal{S}$

$$\begin{aligned} \mathsf{P}_\theta^T(B) &= \int_{T^{-1}(B)} f(x;\theta)\, \mu(dx) \\ &= \int_{T^{-1}(B)} g^*(T(x);\theta)h(x)\, \mu(dx) = \int_B g^*(y;\theta)\, \lambda(dy). \end{aligned}$$

Thus the measure P_θ^T is absolutely continuous with respect to the measure λ and the density is $g^*(y;\theta)$. Therefore (4.1.29) implies that $I(\theta) = I^T(\theta)$ for all $\theta \in \Theta$.

Now we show that if $I(\theta) = I^T(\theta)$ for all $\theta \in \Theta$, then T is a sufficient statistic for $\mathcal{P}$. Indeed, $I(\theta)$ is the variance of $S(\xi;\theta)$, hence

$$I(\theta) = \mathsf{D}_\theta S(\xi;\theta) = \mathsf{E}_\theta \mathsf{D}_\theta (S(\xi;\theta)/T(\xi)) + \mathsf{D}_\theta \mathsf{E}(S(\xi;\theta)/T(\xi)) \tag{4.1.30}$$

where $\mathsf{E}_\theta(S(\xi;\theta)/y) = \mathsf{E}_\theta\{S(\xi;\theta)/T(\xi) = y\}$ and

$$\mathsf{D}_\theta(S(\xi;\theta)/y) = \mathsf{D}_\theta\{S(\xi;\theta)/T(\xi) = y\}$$

are the conditional expectation and variance, respectively. Equality (4.1.25) implies

$$\mathsf{E}_\theta(S(\xi;\theta)/y) = \frac{\partial}{\partial\theta} \ln g(y;\theta) \qquad (\mathsf{P}_\theta^T\text{-a.s.}).$$

The family $\mathcal{P}^T$ satisfies conditions (CR), thus

$$I^T(\theta) = \mathsf{D}_\theta \ln g(T(\xi);\theta) = \mathsf{D}_\theta \mathsf{E}_\theta (S(\xi;\theta)/T(\xi)), \qquad \theta \in \Theta. \tag{4.1.31}$$

Since $I(\theta) = I^T(\theta)$ for all $\theta \in \Theta$, relations (4.1.30) and (4.1.31) yield

$$\mathsf{E}_\theta \mathsf{D}_\theta(S(\xi;\theta)/T(\xi)) = 0 \quad \text{for all } \theta \in \Theta.$$

This implies that

$$\mathsf{D}_\theta(S(\xi;\theta)/T(x)) = 0 \qquad (\mathsf{P}_\theta\text{-a.s.}).$$

In other words, the function $S(x;\theta)$ is $T^{-1}(\mathcal{S})$-measurable. Therefore there exists a measurable function $k(y;\theta)$ on $(Y,\mathcal{S})$ such that $S(x;\theta) = k(T(x);\theta)$ for all $\theta \in \Theta$, whence

$$\ln f(x;\theta) = \int_{\theta_0}^{\theta} k(T(x);t)\,dt + a(x) \quad \text{for all } \theta \in \Theta$$

and some $\theta_0 \in \Theta$. Finally,

$$f(x;\theta) = g^*(T(x);\theta)h(x) \quad \text{for all } \theta \in \Theta.$$

Now the factorization criterion implies that $T(x)$ is a sufficient statistic for $\mathcal{P}$. □

It follows from inequality (4.1.22) that a sufficient statistic is the only statistic that compresses the sampling data without loss of information. The precise statement concerning the Fisher information is given by Theorem 4.1.4. Similar results can be given for other measures of the amount of information about a parameter contained in an observation, say for the Shannon information or Kullback information (see [**22**] and [**20**]).

REMARK 4.1.2. Theorem 4.1.4 remains true for the Fisher information matrix in the case of a multidimensional parameter θ under the multidimensional analogue of the regularity conditions (CR) (see [**36**], §4.3). Theorem 4.1.4 can be proved under other sets of regularity conditions, say $(CR)^*$, (R), or $(R)^*$, in both one-dimensional and multidimensional cases.

EXAMPLE 4.1.7. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample from the Bernoulli distribution with parameter $\theta \in \Theta = (0,1)$. The density of the distribution of ξ_i with respect to the counting measure is of the form

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} = \mathsf{P}_\theta\{\xi_i = x\}$$

where x is either 0 or 1. Thus the Fisher information contained in a single observation ξ_i is

$$I(\theta) = \frac{1}{\theta(1-\theta)}, \qquad \theta \in \Theta,$$

while the Fisher information contained in the whole sample $\xi^{(n)}$ is

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}, \qquad \theta \in \Theta.$$

Let ν_n be the number of "successes" in the sample $\xi^{(n)}$, that is, $\nu_n = \sum_{i=1}^n \xi_i$ is a statistic assuming values in the set $Y = \{0,1,2,\ldots,n\}$. The density of the distribution of the statistic ν_n with respect to the counting measure is given by

$$g(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \mathsf{P}_\theta\{\nu_n = y\}, \qquad y \in Y.$$

The Fisher information contained in the statistic ν_n is

$$I^{\nu_n}(\theta) = \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \ln g(\nu_n; \theta) \right)^2 = \sum_{y=0}^{n} \binom{n}{y} \theta^y (1-\theta)^{n-y} \left(\frac{y}{\theta} - \frac{n-y}{1-\theta} \right)^2$$

$$= \sum_{y=0}^{n} \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{(y-n\theta)^2}{(\theta(1-\theta))^2} = \frac{\mathsf{D}\nu_n}{(\theta(1-\theta))^2} = \frac{n}{\theta(1-\theta)}.$$

Thus $I^{\nu_n}(\theta) = I_n(\theta)$ for all $\theta \in \Theta$. By Theorem 4.1.4 the statistic ν_n is sufficient for the parameter θ. This conclusion can also be made from the factorization criterion.

4.2. Sufficient statistics and optimal estimators

Rao–Blackwell–Kolmogorov theorem. Let ξ be an observation assuming values in a measurable space $(X, \mathcal{B})$ and whose distribution belongs to a family

$$\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$$

where $\theta = (\theta_1, \ldots, \theta_k)'$ is an unknown parameter of a set $\Theta \subset \mathbf{R}^k, k \geq 1$. First we consider the case where θ is a scalar parameter, that is, the case $k = 1$. Let K_b be the class of estimators $\widehat{\theta}$ of the parameter θ with a bias $b(\theta)$ (see Section 3.4). In other words, $a(\theta) = \mathsf{E}_\theta \widehat{\theta} = \theta + b(\theta)$ for all $\theta \in \Theta$.

The following result, known as the *Rao–Blackwell–Kolmogorov theorem*, highlights the role of sufficient statistics in the theory of estimation.

THEOREM 4.2.1. *Let $T = T(x)$ be a sufficient statistic for a family*

$$\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$$

and let $\widehat{\theta} \in K_b$. Then the function $\widehat{\theta}_T = \mathsf{E}_\theta(\widehat{\theta}/T)$ is an estimator such that

1) $\widehat{\theta}_T \in K_b$;
2) *the estimator $\widehat{\theta}_T$ depends on x through $T(x)$;*
3) $\mathsf{E}_\theta(\widehat{\theta}_T - \theta)^2 \leq \mathsf{E}_\theta(\widehat{\theta} - \theta)^2$ *for all $\theta \in \Theta$, and moreover the inequality becomes an equality if and only if $\widehat{\theta} = \widehat{\theta}_T$ almost surely with respect to the measure P_θ.*

PROOF. Let T be a sufficient statistic. The conditional probability $\mathsf{P}_\theta(A/T)$ depends on A and T and does not depend on θ. Moreover $\mathsf{P}_\theta(A/T)$ is a measurable function of T. Thus the conditional expectation $\widehat{\theta}_T = \mathsf{E}_\theta(\widehat{\theta}/T)$ depends on T and does not depend on θ. Therefore the estimator $\widehat{\theta}_T$ satisfies condition 2) of the theorem.

Properties of the conditional expectation imply that

$$\mathsf{E}_\theta \widehat{\theta}_T = \mathsf{E}_\theta \mathsf{E}_\theta(\widehat{\theta}/T) = \mathsf{E}_\theta \widehat{\theta},$$

that is, $\widehat{\theta}_T \in K_b$, whence condition 1) of the theorem follows.

The inequality in statement 3) of the theorem is obvious if $\mathsf{E}_\theta(\widehat{\theta} - \theta)^2 = \infty$. Thus we consider the case when $\mathsf{E}_\theta(\widehat{\theta} - \theta)^2 < \infty$. We have

$$\begin{aligned} \mathsf{E}_\theta(\widehat{\theta} - \theta)^2 &= \mathsf{E}_\theta(\widehat{\theta} - \widehat{\theta}_T + \widehat{\theta}_T - \theta)^2 \\ &= \mathsf{E}_\theta(\widehat{\theta} - \widehat{\theta}_T)^2 + \mathsf{E}_\theta(\widehat{\theta}_T - \theta)^2 + 2\mathsf{E}_\theta(\widehat{\theta} - \widehat{\theta}_T)(\widehat{\theta}_T - \theta). \end{aligned} \tag{4.2.1}$$

By the properties of the conditional expectation

$$\begin{aligned} \mathsf{E}_\theta(\widehat{\theta}-\widehat{\theta}_T)(\widehat{\theta}_T-\theta) &= \mathsf{E}_\theta\mathsf{E}_\theta\big((\widehat{\theta}-\widehat{\theta}_T)(\widehat{\theta}_T-\theta)/T\big) \\ &= \mathsf{E}_\theta(\widehat{\theta}_T-\theta)\mathsf{E}_\theta(\widehat{\theta}-\widehat{\theta}_T/T)=0, \end{aligned} \tag{4.2.2}$$

since $\mathsf{E}_\theta(\widehat{\theta}-\widehat{\theta}_T/T)=\mathsf{E}_\theta(\widehat{\theta}/T)-\widehat{\theta}_T=0$.

Equalities (4.2.1) and (4.2.2) imply

$$\mathsf{E}_\theta(\widehat{\theta}-\theta)^2=\mathsf{E}_\theta(\widehat{\theta}-\widehat{\theta}_T)^2+\mathsf{E}_\theta(\widehat{\theta}_T-\theta)^2,$$

whence statement 3) of the theorem follows. □

Theorem 4.2.1 shows that if T is a sufficient statistic, then one can improve the estimator $\widehat{\theta}\in K_b$ uniformly in $\theta\in\Theta$ by applying the operator $\mathsf{E}_\theta(\cdot/T)$ to the statistic $\widehat{\theta}$.

There is another interpretation of Theorem 4.2.1. Namely let S and T be two sufficient statistics for a family $\mathcal{P}$. If $\widehat{\theta}=\phi(T)$ where ϕ is a measurable function and S is a measurable function of T, then

$$\mathsf{E}_\theta(\widehat{\theta}_S-\theta)^2\le\mathsf{E}_\theta(\widehat{\theta}-\theta)^2$$

where $\widehat{\theta}_S=\mathsf{E}_\theta(\widehat{\theta}/S)$. In other words, one should find the so-called *minimal* sufficient statistics, that is those statistics for which any other sufficient statistic is a function of it. The procedure of the construction of an optimal estimator is as follows. One starts with a "bad" estimator $\widehat{\theta}$ and improves it with the help of sufficient statistics until the estimator becomes optimal.

Theorem 4.2.1 holds in the multidimensional case as well. In this case θ and $\widehat{\theta}$ are vectors of the space $\mathbf{R}^k$, $k>1$. As in the one-dimensional case let K_b be the class of estimators $\widehat{\theta}$ of the parameter θ with the bias $b(\theta)$.

THEOREM 4.2.2. *Let T be a sufficient statistic for a family $\mathcal{P}=(\mathsf{P}_\theta,\theta\in\Theta)$ and let $\widehat{\theta}\in K_b$. Then the estimator $\widehat{\theta}_T=\mathsf{E}_\theta(\widehat{\theta}/T)$ is such that*

1) $\widehat{\theta}_T\in K_b$;
2) *$\widehat{\theta}_T$ depends on x through $T(x)$;*
3) *for any vector $a\in\mathbf{R}^k$*

$$\mathsf{E}_\theta\big(a'(\widehat{\theta}_T-\theta)\big)^2\le\mathsf{E}_\theta\big(a'(\widehat{\theta}-\theta)\big)^2. \tag{4.2.3}$$

This inequality becomes an equality if and only if $\widehat{\theta}=\widehat{\theta}_S$ almost surely with respect to the measure P_θ.

PROOF. The first two statements of the theorem are obvious. Inequality (4.2.3) follows from Theorem 4.2.1, since the proof is reduced to the one-dimensional estimators $a'\widehat{\theta}$ of the parameter $a'\theta$ and since $\mathsf{E}_\theta(a'\widehat{\theta}/S)=a'\widehat{\theta}_S$. If (4.2.3) becomes an equality for all $a\in\mathbf{R}^k$, then $a'\widehat{\theta}=a'\widehat{\theta}_S$ almost surely with respect to the measure P_θ for all a, whence $\widehat{\theta}=\widehat{\theta}_S$ almost surely with respect to P_θ. □

REMARK 4.2.1. All the vectors a, θ, $\widehat{\theta}$, and $\widehat{\theta}_T$ in Theorem 4.2.2 are column-vectors.

Sufficient statistics and efficient estimators. We proved in Theorems 4.2.1 and 4.2.2 that if an estimator is not a function of a sufficient statistic, then it can be improved by using this sufficient statistic. However we still have no tool to construct an optimal estimator by following this idea.

On the other hand, if a set of regularity conditions holds, say (CR) or (R), and the Cramér–Rao inequality becomes an equality, then the estimator is optimal (this kind of optimality is called efficiency in Sections 3.4 and 3.5).

Below we consider the case where conditions (R) are satisfied. All other cases can be studied in a similar way. Let $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_k)'$ be an estimator of the parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$ constructed from an observation ξ. Let K_b be the class of estimators $\widehat{\theta}$ of the parameter θ with a bias $b(\theta)$, that is,

$$a(\theta) = \mathsf{E}_\theta \widehat{\theta} = \theta + b(\theta).$$

The following result contains a relationship between efficient estimators and sufficient statistics.

THEOREM 4.2.3. *Let conditions (R) hold. Let $\widehat{\theta}$ be an estimator of the class K_b such that* $\det D(\widehat{\theta};\theta) > 0$ *where $D(\widehat{\theta};\theta) = \mathsf{E}_\theta(\widehat{\theta} - a(\theta))(\widehat{\theta} - a(\theta))'$ is the covariation matrix of the estimator $\widehat{\theta}$. Then $\widehat{\theta}$ is an efficient estimator in the class K_b if and only if $\widehat{\theta}$ is a sufficient statistic of the parameter θ,*

$$f(x;\theta) = g\big(\widehat{\theta}(x);\theta\big)r(x) \quad \text{for all } \theta \in \Theta, \tag{4.2.4}$$

and all $r(x) = h(x)$ and $g(\widehat{\theta};\theta) = \exp(A(\theta)'\widehat{\theta} + C(\theta))$ and $h(x)$, $A(\theta)$, and $C(\theta)$ are the functions occurring in representation (3.5.14).

PROOF. By definition, $\widehat{\theta}$ is an efficient estimator of the parameter θ if and only if the Cramér–Rao inequality (3.5.13) becomes an equality. According to Theorem 3.5.2, the Cramér–Rao inequality (3.5.13) becomes an equality if and only if relation (3.5.14) holds. Note that relation (3.5.14) coincides with (4.2.4). By the factorization criterion (Theorem 4.1.2), the estimator $\widehat{\theta}$ is a sufficient statistic for the parameter θ. □

Note that if $\widehat{\theta}$ is an efficient estimator of the parameter θ in the class K_b and regularity conditions hold, then $\widehat{\theta}$ is also an optimal estimator of the parameter θ in the class K_b in the sense of the definition of Section 3.1. Generally speaking, the converse is not true, namely an estimator can be optimal in a class K_b, but the lower bound in the Cramér–Rao inequality is not attained for it. Thus an important role is played by those sufficient statistics that, by the Rao–Blackwell–Kolmogorov theorem, allow one to improve estimators and, in the case where one can construct the minimal sufficient statistics, to construct the optimal estimator.

Minimal sufficient statistics. We have seen above that there exist many sufficient statistics in the general situation. In particular, there always exists the so-called trivial sufficient statistic, namely $T(x) = x$. Nevertheless we are interested in those statistics that provide the best reduction of the data. However it is not always possible to find a sufficient statistic for which the reduction of the data is essentially better than that for the trivial sufficient statistic. To make the notion of the reduction of the data precise we introduce a partial order on the set of all sufficient statistics.

We say that a statistic T_1 is *subordinated* to a statistic T_2 if T_1 is a measurable function of T_2, that is, $T_1 = \phi(T_2)$. If a statistic T_1 is subordinated to a statistic T_2 and T_2 is subordinated to T_1, then the statistics T_1 and T_2 are called *equivalent*. A sufficient statistic T_0 is called *minimal* if it is subordinated to any other sufficient statistic.

The reduction of the data is best for a minimal sufficient statistic. If T is a minimal sufficient statistic, then a further reduction of the information as compared to T gives no result if the statistic remains sufficient.

We have seen above that the definition of sufficient statistics can be given in a more general form in terms of σ-algebras. If $\mathcal{B}' \subset \mathcal{B}$ is a σ-algebra, then $\mathcal{B}'$ is called a *sufficient σ-algebra* for the family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ (or, for the parameter θ) if there is a version of the conditional probability measure $\mathsf{P}_\theta(A/\mathcal{B}')$, $A \in \mathcal{B}$, that does not depend on θ. Let T be a statistic mapping a measurable space $(X, \mathcal{B})$ into a measurable space $(Y, \mathcal{S})$, and let $\mathcal{B}_T = T^{-1}(\mathcal{S})$ be the preimage of the σ-algebra $\mathcal{S}$ under the mapping T. If the σ-algebra $\mathcal{B}_T$ generated by the statistic T is sufficient, then T is a sufficient statistic. All the results on sufficient statistics can be stated in terms of sufficient σ-algebras. In particular, the factorization criterion remains true if the function $g(T(x); \theta)$ is substituted for a $\mathcal{B}'$-measurable function $g(x; \theta)$; in this case, $\mathcal{B}'$ is a sufficient σ-algebra.

Let T_1 and T_2 be two statistics. It is clear that T_1 is subordinated to T_2 if

$$\mathcal{B}_{T_1} \subset \mathcal{B}_{T_2}.$$

Thus the statistic T_1 reduces the data in a better way than the statistic T_2. Two statistics T_1 and T_2 are equivalent if and only if $\mathcal{B}_{T_1} = \mathcal{B}_{T_2}$.

A σ-algebra $\mathcal{B}^*$ is called the *minimal sufficient σ-algebra* if it belongs to any other sufficient σ-algebra, that is, $\mathcal{B}^* \subset \mathcal{B}'$ for any sufficient σ-algebra $\mathcal{B}'$. In other words, T_0 is a minimal sufficient statistic if $\mathcal{B}_{T_0} \subset \mathcal{B}_T$ for every sufficient statistic T.

A minimal sufficient statistic always exists for dominated families

$$\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$$

(see Theorem 4.2.4). To prove this result we use Theorem 4.1.1: for a family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ dominated by a σ-finite measure μ, there exists a discrete distribution $\mathbf{Q}$ on Θ such that the family $\mathcal{P}$ is dominated by the probability measure $\mathsf{P}_Q = \int \mathsf{P}_t \, \mathbf{Q}(dt)$. Then the density $p(x; \theta)$ of the measure P_θ with respect to the measure P_Q can be expressed as

$$\frac{d\mathsf{P}_\theta}{d\mathsf{P}_Q}(x) = p(x; \theta) = \frac{f(x; \theta)}{f(x; \mathbf{Q})} \qquad (\mu\text{-a.s.}) \tag{4.2.5}$$

where $f(x; \theta) = d\mathsf{P}_\theta / d\mu(x)$ and $f(x; \mathbf{Q}) = d\mathsf{P}_Q / d\mu(x)$. If T is a sufficient statistic, then $p(x; \theta)$ depends on x through $T(x)$ by the factorization criterion.

THEOREM 4.2.4. *Let a family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ be dominated by some σ-finite measure μ and let $\mathcal{B}^* = \sigma\{p(x; \theta);\ \theta \in \Theta\}$ be the σ-algebra in $(X, \mathcal{B})$ generated by the functions $p(x; \theta)$, $\theta \in \Theta$. Then $\mathcal{B}^*$ is the minimal sufficient σ-algebra.*

PROOF. According to (4.2.5) we have

$$f(x; \theta) = p(x; \theta) f(x; \mathbf{Q}) \qquad (\mu\text{-a.s.})$$

where the function $f(x;\mathbf{Q})$ does not depend on θ and the function $p(x;\theta)$ is $\mathcal{B}^*$-measurable for all $\theta \in \Theta$. From the factorization criterion for sufficient σ-algebras we obtain that $\mathcal{B}^*$ is a sufficient σ-algebra.

Now let $\mathcal{B}'$ be an arbitrary σ-algebra. Then $f(x;\theta) = g(x;\theta)r(x)$ (μ-a.s.) where $h(x)$ is a nonnegative $\mathcal{B}$-measurable function and $g(x;\theta)$ is a nonnegative $\mathcal{B}'$-measurable function. Consider the σ-algebra $\mathcal{B}_g = \sigma\{g(x;\theta);\theta \in \Theta\} \subset \mathcal{B}'$. It follows from (4.2.5) that

$$p(x;\theta) = \frac{g(x;\theta)}{\int g(x;t)\mathbf{Q}(dt)},$$

whence $\mathcal{B}^* \subset \mathcal{B}_g \subset \mathcal{B}'$. □

REMARK 4.2.2. Using Theorem 4.1.3 one can construct a minimal sufficient σ-algebra from the a posteriori distribution $\mathbf{Q}_x$. Let $\mathbf{Q}$ be an a priori measure such that its density $q(t)$ with respect to some other measure λ is positive for all $t \in \Theta$. Then a posteriori density is given by

$$q(t/x) = \frac{f(x;t)q(t)}{f(x;\mathbf{Q})} = p(x;t)q(t).$$

Thus the minimal sufficient σ-algebra $\mathcal{B}^*$ can be viewed as one generated by a posteriori distribution, that is, $\mathcal{B}^* = \sigma\{q(t/x);t \in \Theta\}$.

EXAMPLE 4.2.1. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the Poisson distribution with parameter $\theta \in \Theta = (0,\infty)$. We learned in Example 4.1.1 that

$$T_n = T_n(x) = \sum_{i=1}^{n} x_i$$

is a sufficient statistic. Here $x = (x_1, \dots, x_n)$ and $x_i \in \{0,1,2,\dots\}$ for all $i = 1,2,\dots,n$. Then T_n is the minimal sufficient statistic by Theorem 4.2.4, since the σ-algebra $\mathcal{B}_{T_n}$ coincides with the σ-algebra generated by the functions

$$p(x;\theta) = \frac{f(x;\theta)}{f(x;\theta_0)} = \left(\frac{\theta}{\theta_0}\right)^{T_n(x)} e^{n(\theta_0-\theta)}$$

where $\theta_0 \in (0,\infty)$ is some fixed value of the parameter. As the distribution $\mathbf{Q}$ on Θ we consider the distribution concentrated at the point θ_0.

EXAMPLE 4.2.2. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the uniform distribution on the interval $[0,\theta]$ where $\theta > 0$ is an unknown parameter. Let the sampling space be $\mathbf{R}_+^n = \{x = (x_1, x_2, \dots, x_n): x_i > 0 \text{ for all } i = 1,2,\dots,n\}$. As in Example 4.1.5 we prove that $T_n(x) = x_{n,n} = \max_{1\le i\le n} x_i$ is a sufficient statistic. Moreover it is the minimal sufficient statistic. Indeed, let $\mathbf{Q}$ be some distribution on $(0,\infty)$ whose density $q(t)$ is positive for all $t \in (0,\infty)$. Then

$$f(x;\theta) = \theta^{-n} I_{(-\infty,\theta]}(x_{n,n}), \qquad x \in \mathbf{R}_+^n,$$

$$f(x;\mathbf{Q}) = \int_0^\infty f(x;t)q(t)\,dt = \int_{T_n(x)}^\infty t^{-n}q(t)\,dt > 0 \quad \text{for all } x \in \mathbf{R}_+^n.$$

It is also clear that $T_n(x) = \sup\{\theta: f(x;\theta)/f(x;\mathbf{Q}) = 0\}$, $x \in \mathbf{R}_+^n$. This means that the statistic T_n is measurable with respect to the minimal sufficient σ-algebra

$\mathcal{B}^* = \sigma\{f(x;\theta)/f(x;\mathbf{Q});\ \theta \in \Theta\}$, $\mathcal{B}_{T_n} \subset \mathcal{B}^*$. Therefore T_n is the minimal sufficient statistic.

REMARK 4.2.3. There is another method to construct minimal sufficient statistics based on partitions of the sampling space generated by sufficient statistics. This method can be found in [**36**], also see [**5**] and [**19**].

Complete statistics and optimal estimators. In this section we consider complete sufficient statistics and use them to construct optimal estimators of a parameter. Let $T = T(x)$ be a statistic mapping a measurable space $(X, \mathcal{B})$ into a measurable space $(Y, \mathcal{S})$. Assume that the dimension of the space Y is l, that is, $Y \subset \mathbf{R}^l$. A usual assumption is that $l > k$ where k is the dimension of the parameter θ.

Let $\Gamma = \{\mathbf{G}_\theta; \theta \in \Theta\}$ be some family of probability measures on $(\mathbf{R}^l, \mathcal{B}^l)$. A family Γ is called *complete* if the relation

$$\int \phi(x)\, \mathbf{G}_\theta(dx) = 0 \quad \text{for all } \theta \in \Theta \tag{4.2.6}$$

implies $\phi(x) = 0$ ($\mathbf{G}_\theta$-a.s.) for all $\theta \in \Theta$. Equation (4.2.6) is considered in the class of functions $\phi\colon \mathbf{R}^l \to \mathbf{R}^k$ for which the integral (4.2.6) exists.

Let $\mathcal{P}^T = \{\mathsf{P}_\theta^T; \theta \in \Theta\}$ be a family of probability measures on $(Y, \mathcal{S})$ generated by the mapping $T\colon X \to Y$ where $\mathsf{P}_\theta^T(B) = \mathsf{P}_\theta(T^{-1}(B))$ and $B \in \mathcal{S}$, $\theta \in \Theta$. A statistic T is called *complete* if the family of distributions $\mathcal{P}^T$ is complete. Equation (4.2.6) for the statistic T can be written in the following form:

$$\mathsf{E}_\theta \phi(T(\xi)) = 0 \quad \text{for all } \theta \in \Theta. \tag{4.2.7}$$

THEOREM 4.2.5. *A statistic T is complete if and only if for some $b_0(\theta)$ a $\mathcal{B}_T$-measurable estimator $\widehat{\theta}$ is unique in the class of all $\mathcal{B}_T$-measurable estimators of the class K_{b_0} where $b_0(\theta)$ is the bias of the estimator.*

If a $\mathcal{B}_T$-measurable estimator is unique in the class K_{b_0}, then any $\mathcal{B}_T$-measurable estimator is unique in any other class K_b of estimators with the bias $b(\theta)$.

PROOF. Let $\widehat{\theta}_1 = \phi_1(T)$ and $\widehat{\theta}_2 = \phi_2(T)$ be two $\mathcal{B}_T$-measurable estimators of K_{b_0}. Then $\mathsf{E}_\theta \phi_i(T(\xi)) = b_0(\theta)$, $i = 1, 2$, and $\mathsf{E}_\theta(\phi_1(T(\xi)) - \phi_2(T(\xi))) = 0$ for all $\theta \in \Theta$. Since T is a complete statistic, $\phi_1(y) = \phi_2(y)$ (P_θ^T-a.s.) for all $\theta \in \Theta$. Conversely, let $\mathsf{E}_\theta \phi(T(\xi)) = 0$ for all $\theta \in \Theta$ and $\widehat{\theta}_1 = \phi_1(T) \in K_{b_0}$. Then $\widehat{\theta}_2 = \phi_1(T) + \phi(T) \in K_{b_0}$. Since a $\mathcal{B}_T$-measurable estimator of K_{b_0} is unique, $\phi(T(x)) = 0$ (P_θ-a.s.) for all $\theta \in \Theta$, that is, $\phi(y) = 0$ (P_θ^T-a.s.) for all $\theta \in \Theta$.

The second statement of the theorem is obvious. □

THEOREM 4.2.6. *If a sufficient statistic T is complete and $\widehat{\theta} \in K_b$, then*

$$\widehat{\theta}_T = \mathsf{E}_\theta(\widehat{\theta}/T)$$

is a unique optimal estimator in the class K_b.

PROOF. Since T is complete, Theorem 4.2.5 implies that a $\mathcal{B}_T$-measurable estimator is unique in K_b.

Let $\widetilde{\theta}$ be any other estimator of the class K_b. Then $\widetilde{\theta}_T = \mathsf{E}_\theta(\widetilde{\theta}/T) \in K_b$, thus we get by Theorem 4.2.5 that $\widetilde{\theta}_T = \widehat{\theta}_T$ (P_θ-a.s.) for all $\theta \in \Theta$. This together with

the Rao–Blackwell–Kolmogorov theorem implies that

$$\mathsf{E}_\theta(\widehat{\theta}_T - \theta)^2 = \mathsf{E}_\theta(\widetilde{\theta}_T - \theta)^2 \le \mathsf{E}_\theta(\widetilde{\theta} - \theta)^2 \quad \text{for all } \theta \in \Theta.$$

Moreover the inequality becomes an equality if and only if $\widetilde{\theta} = \widehat{\theta}_T$ (P_θ-a.s.) for all $\theta \in \Theta$. □

COROLLARY 4.2.1. *If T is a complete sufficient statistic and $\widehat{\theta}$ is an unbiased estimator of the parameter θ, then $\widehat{\theta}_T = \mathsf{E}_\theta(\widehat{\theta}/T)$ is a unique optimal unbiased estimator of the parameter θ.*

PROOF. It is necessary to apply Theorem 4.2.6 to the class K_{b_0} where $b_0(\theta) = 0$ for all $\theta \in \Theta$. □

EXAMPLE 4.2.3. Let an observation $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the Poisson distribution with parameter $\theta \in \Theta = (0, \infty)$. Consider the estimator $\widehat{\theta}_n(\xi^{(n)}) = \xi_1$. It is clear that this estimator is unbiased for all $\theta \in \Theta$, that is, $\mathsf{E}_\theta\widehat{\theta}_n = \mathsf{E}_\theta\xi_1 = \theta$. At the same time it is not consistent, since it does not depend on n. Consider the statistic

$$T_n(x) = \sum_{i=1}^{n} x_i, \qquad x = (x_1, x_2, \dots, x_n),$$

where $x_i \in \{0, 1, 2, \dots\}$ for all $i = 1, 2, \dots, n$. We learned in Examples 4.1.1 and 4.2.1 that T_n is a minimal sufficient statistic for the parameter θ. Moreover Example 4.1.1 implies that the conditional distribution of ξ_1 given T_n is of the form

$$\mathsf{P}_\theta\left\{\xi_1 = x/T(\xi^{(n)}) = y\right\} = \binom{y}{x}\left(\frac{1}{n}\right)^x\left(1 - \frac{1}{n}\right)^{y-x}$$

where $y \in \{0, 1, 2, \dots, n\}$ and $x \in \{0, 1, 2, \dots, y\}$. Thus

$$\begin{aligned}\widehat{\theta}_{T_n}(x) &= \mathsf{E}_\theta\left(\xi_1/T_n(x)\right) = \sum_{k=0}^{T_n(x)} k\binom{T_n(x)}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{T_n(x)-k} \\ &= \frac{1}{n}T_n(x) = \overline{x}.\end{aligned} \tag{4.2.8}$$

Now we show that T_n is a complete statistic. Since the distribution of T_n is Poisson with parameter $n\theta$, equation (4.2.7) for T_n becomes of the form

$$\sum_{k=0}^{\infty} \phi(k)e^{-n\theta}\frac{(n\theta)^k}{k!} = 0 \quad \text{for all } \theta \in \Theta$$

or, equivalently,

$$v(z) = \sum_{k=0}^{\infty} \phi(k)\frac{z^k}{k!} = 0 \quad \text{for all } z \in \Theta. \tag{4.2.9}$$

The convergence of series (4.2.9) at $z = 1$ implies that the function $v(z)$ is analytic for $|z| < 1$, whence $\phi(k) = 0$ for all k. Taking into account equality (4.2.8), we obtain from Corollary 4.2.1 that $\overline{x}$ is a unique optimal estimator of the parameter θ in the class of unbiased estimators.

EXAMPLE 4.2.4. Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from the uniform distribution on the interval $[0, \theta]$ where $\theta \in \Theta = (0, \infty)$ is an unknown parameter. It is natural to consider $\mathbf{R}_+^n$ as the sample space in this case. We learned in Examples 4.1.5 and 4.2.2 that $T_n(x) = x_{n,n}$ is a sufficient and minimal statistic. Its distribution function is

$$\mathsf{P}_\theta\{T_n < y\} = \left(\frac{y}{\theta}\right)^n, \qquad 0 \le y \le \theta.$$

Thus equality (4.2.7) for T_n becomes of the form

$$\int_0^\theta \phi(y) \frac{n y^{n-1}}{\theta^n}\, dy = 0 \quad \text{for all } \theta \in \Theta,$$

whence it follows that $\phi(y) y^{n-1} = 0$ for almost all $y > 0$ and therefore $\phi(y) = 0$ for almost all $y > 0$. This means that T_n is a complete statistic.

Since θ is a scale parameter, Theorems 3.3.1 and 3.3.2 imply that

$$\widehat{\theta}_n = T_n(n+1)/n$$

is an optimal estimator in the class of equivariant unbiased estimators of the scale parameter. Note that $\widehat{\theta}_n$ is a $\mathcal{B}_{T_n}$-measurable estimator and Corollary 4.2.1 implies that $\widehat{\theta}_n$ is an optimal estimator in the class of all unbiased estimators of the parameter θ.

The latter result can also be obtained directly from Corollary 4.2.1 (or from Theorem 4.2.6) by considering the estimator $\widetilde{\theta} = 2x_1$ which obviously is unbiased. According to Corollary 4.2.1 $\widetilde{\theta}_{T_n} = \mathsf{E}_\theta(\widetilde{\theta}/T_n)$ is a unique optimal unbiased estimator of the parameter θ. The conditional density of ξ_1 given T_n is given by

$$f(x/y) = \begin{cases} \left(1 - \dfrac{1}{n}\right)\dfrac{1}{y}, & 0 < x < y, \\ \dfrac{1}{n}, & x = y. \end{cases} \tag{4.2.10}$$

Indeed, since $\xi_1, \xi_2, \dots, \xi_n$ are independent identically distributed random variables and the model is symmetric, we get

$$\mathsf{P}_\theta\{\xi_1 = T_n/T_n\} = \frac{1}{n}, \qquad \mathsf{P}_\theta\{\xi_1 < T_n/T_n\} = 1 - \frac{1}{n}.$$

Moreover given $\{T_n = y, \xi_1 < T_n\}$, the conditional distribution of ξ_1 is uniform on $(0, y)$, since the distribution of ξ_1 is uniform. This leads to equality (4.2.10). Using equality (4.2.10) we obtain

$$\mathsf{E}_\theta\{\xi_1/T_n = y\} = \left(1 - \frac{1}{n}\right)\frac{y}{2} + \frac{y}{n}.$$

This implies that

$$\widetilde{\theta}_{T_n} = \mathsf{E}_\theta(\widetilde{\theta}/T_n) = \left(1 - \frac{1}{n}\right) T_n + \frac{2}{n} T_n = \frac{n+1}{n} T_n.$$

Another consequence of Theorem 4.2.6 is an assertion on the optimality of a function $g(\theta)$ of the parameter θ in the class K^g of all unbiased estimators of the function $g(\theta)$.

COROLLARY 4.2.2. *If T is a complete sufficient statistic and $\widehat{g}$ is an unbiased estimator of the function $g(\theta)$, then $\widehat{g}_T = \mathsf{E}_\theta(\widehat{g}/T)$ is a unique optimal unbiased estimator of the function $g(\theta)$ in the class K^g.*

PROOF. It is sufficient to apply Theorem 4.2.6 to estimators of the class K_{b_0} where $b_0(\theta) = g(\theta) - \theta$ for all $\theta \in \Theta$. □

Corollary 4.2.2 is known as the *Lehmann–Scheffeé theorem*; see [**36**], Theorem 3.1.2.

COROLLARY 4.2.3. *If T is a complete sufficient statistic, then any function $H(T)$ of it is a unique unbiased optimal estimator of its own expectation, that is, of the function $g(\theta) = \mathsf{E}_\theta H(T)$.*

PROOF. To prove this result it is sufficient to consider the class of estimators K^g with $g(\theta) = \mathsf{E}_\theta H(T)$ and to apply Corollary 4.2.2 with $\widehat{g} = H(T)$. Then

$$H(T) = \widehat{g}_T = \mathsf{E}_\theta(\widehat{g}/T)$$

is a unique optimal estimator of the function $g(\theta) = \mathsf{E}_\theta H(T)$. □

In fact we have a series of results allowing one to find optimal estimators of the function $g(\theta)$ if a complete sufficient statistic T exists, namely:

1) *if there exists an unbiased estimator of a function $g(\theta)$, then there exists an unbiased estimator that is a function of T; if there is no unbiased estimators of the form $H(T)$, that is, the equation $\mathsf{E}_\theta H(T) = g(\theta)$ has no solution, then the class K^g of unbiased estimators of the function $g(\theta)$ is empty;*
2) *an optimal unbiased estimator of the function $g(\theta)$ (if such an estimator exists at all) is a function of T and it is determined uniquely by the equality $\mathsf{E}_\theta H(T) = g(\theta)$;*
3) *an optimal unbiased estimator g^* of the function $g(\theta)$ can be found as follows:*

$$g^* = \widehat{g}_T = \mathsf{E}_\theta(\widehat{g}/T) \tag{4.2.11}$$

where $\widehat{g}$ is an arbitrary unbiased estimator of the function $g(\theta)$.

The latter method is rarely used when finding optimal estimators, since it requires the evaluation of the conditional expectation (4.2.11) which usually meets technical problems. Instead the equation

$$\mathsf{E}_\theta H(T) = g(\theta), \qquad \theta \in \Theta, \tag{4.2.12}$$

is used to determine the function H. There are several methods for solving equation (4.2.12). For example, one can expand both sides of (4.2.12) into power series of θ and equate corresponding coefficients.

The following result provides a relationship between complete and minimal statistics.

THEOREM 4.2.7. *Any complete sufficient statistic T is a minimal sufficient statistic.*

PROOF. Let $\mathcal{B}^*$ be the minimal sufficient σ-algebra (this σ-algebra exists by Theorem 4.2.4).

First we assume that $\mathsf{E}_\theta T$ exists and consider the function $\phi = T - \mathsf{E}_\theta(T/\mathcal{B}^*)$. Since $\mathcal{B}^* \subset \mathcal{B}_T$ where $\mathcal{B}_T$ is the σ-algebra generated by the statistic T, the function ϕ is $\mathcal{B}_T$-measurable, whence $\phi = \phi(T)$. Denote by P_θ^T the distribution of the statistic T. Then $\mathsf{E}_\theta\phi(T) = 0$ for all $\theta \in \Theta$ or, equivalently,

$$\int \phi(y)\mathsf{P}_\theta^T(dy) = 0 \qquad \text{for all } \theta \in \Theta.$$

This implies that $\phi(y) = 0$ (P_θ^T-a.s.) for all $\theta \in \Theta$, since T is a complete statistic. This means that $T = \mathsf{E}_\theta(T/\mathcal{B}^*)$ (P_θ^T-a.s.). Therefore T is a $\mathcal{B}^*$-measurable statistic and hence $\mathcal{B}_T = \mathcal{B}^*$.

If $\mathsf{E}_\theta T$ does not exist, then one should consider the statistic $\arctan T$ instead of T. If T is either sufficient, or complete, or minimal, then so is $\arctan T$. □

It is easy to construct an example of a minimal sufficient statistic that is not a complete statistic (see, for example, [**36**]).

Exponential families of distributions. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$ whose distribution belongs to a family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ where $\theta = (\theta_1, \theta_2, \dots, \theta_k)'$ is a k-dimensional parameter of the set $\Theta \subset \mathbf{R}^k$, $k \geq 1$. Let a family $\mathcal{P}$ be dominated by some σ-finite measure μ and let $f(x;\theta) = d\mathsf{P}_\theta/d\mu(x)$, $x \in X$, be the density of the measure P_θ with respect to the measure μ.

A family $\mathcal{P}$ is called *exponential* if the density $f(x;\theta)$ is of the form

$$f(x;\theta) = h(x)\exp\left\{\sum_{j=1}^{k} c_j(\theta)U_j(x) + V(\theta)\right\} \tag{4.2.13}$$

where all the functions on the right-hand side are finite and measurable.

Various distributions known in the literature are exponential. For example, normal, Poisson, binomial, Gamma-distributions, and others form exponential families of distributions.

If an observation is a sample $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ from an exponential family of the form (4.2.13), then the distribution of the sample $\xi^{(n)}$ also belongs to some exponential family $\mathcal{P}_n$. Indeed, if $f_n(x;\theta)$ is the density of the distribution of the sample with respect to the measure μ^n, then

$$f_n(x;\theta) = h_n(x)\exp\{c(\theta)'T(x) + nV(\theta)\} \tag{4.2.14}$$

where

$$c(\theta) = (c_1(\theta), \dots, c_k(\theta))', \qquad T(x) = (T_1(x), \dots, T_k(x))',$$
$$h_n(x) = \prod_{i=1}^{n} h(x_i), \quad T_j(x) = \sum_{i=1}^{n} U_j(x_i), \qquad x = (x_1, \dots, x_n) \in X^n.$$

It follows from the factorization criterion that if the family of distributions $\mathcal{P}$ is exponential, then the statistic $U(x) = (U_1(x), \dots, U_n(x))$ is sufficient. Similarly, the statistic $T(x)$ is sufficient for a family of distributions $\mathcal{P}_n$ whose densities are of the form (4.2.14). It turns out that these statistics are minimal and sufficient. We prove this result for families $\mathcal{P}_n$ with densities (4.2.14).

Since the functions $c_j(\theta)$, $U_j(x)$, and $V(\theta)$ are finite, the exponent in (4.2.14) is positive everywhere. Thus as a measure $\mathbf{Q}$ in Theorem 4.2.4 one can take a distribution concentrated at any fixed point $\theta^{(0)} \in \Theta$. All the measures P_θ are absolutely continuous with respect to the measure $\mathsf{P}_{\mathbf{Q}} = \int \mathsf{P}_t \, \mathbf{Q}(dt) = \mathsf{P}_{\theta(0)}$ in this case. Theorem 4.2.4 implies that the σ-algebra $\mathcal{B}^*$ generated by the functions

$$\begin{aligned} p(x;\theta) &= f_n(x;\theta)/f_n(x;\theta^{(0)}) \\ &= \exp\bigl(\bigl(c(\theta) - c(\theta^{(0)})\bigr)'T(x) + n\bigl(V(\theta) - V(\theta^{(0)})\bigr)\bigr) \end{aligned}$$

is the minimal sufficient σ-algebra for all $\theta \in \Theta$.

THEOREM 4.2.8. *Let functions $c_0(\theta) \equiv 1$, $c_1(\theta), \dots, c_k(\theta)$ be linearly independent on Θ. Then $T(x)$ in representation* (4.2.14) *is a minimal sufficient statistic.*

PROOF. Since the functions $1, c_1(\theta), \dots, c_k(\theta)$ are linearly independent on Θ, it follows that the functions $c_1(\theta) - c(\theta^{(0)}), \dots, c_k(\theta) - c(\theta^{(0)})$ are linearly independent. This means that there are k points $\theta^{(1)}, \dots, \theta^{(k)}$ in Θ such that the numbers

$$c_{ij} = c_i(\theta^{(j)}) - c_i(\theta^{(0)})$$

form the matrix C whose determinant is nonzero. This implies that the system of equations

$$\bigl(c(\theta^{(j)}) - c(\theta^{(0)})\bigr)'T(x) = \ln p(x;\theta^{(j)}) - n\bigl(V(\theta^{(j)}) - V(\theta^{(0)})\bigr)$$

for $j = 1, 2, \dots, k$ has a unique solution $T(x)$. Thus

$$\mathcal{B}_T \subset \sigma\bigl\{p(x;\theta^{(j)}), j = 1, 2, \dots, k\bigr\} \subset \mathcal{B}^*. \qquad \square$$

Sometimes the assumptions of Theorem 4.2.8 are too restrictive if one proves only that T is a complete sufficient statistic. First we note that representation (4.2.14) yields

$$f_n(x;\theta) = g(T(x);\theta)r(x) \qquad (\mu^n\text{-a.s.}) \tag{4.2.15}$$

(see Theorem 4.1.2) where

$$g(y;\theta) = \exp\{c(\theta)'y + nV(\theta)\},$$

$$r(x) = h_n(x) = \prod_{i=1}^{n} h(x_i), \qquad x = (x_1, \dots, x_n).$$

Consider the following measure on $(\mathbf{R}^k, \mathcal{B}^k)$:

$$\nu(B) = \int_{T^{-1}(B)} h_n(x)\, \mu^n(dx), \qquad B \in \mathcal{B}^k,$$

where $T^{-1}(B) = \{x: T(x) \in B\}$. In what follows we need two auxiliary results.

LEMMA 4.2.1. *The distribution $\mathsf{P}_\theta^T(B) = \mathsf{P}_\theta\{x: T(x) \in B\}$, $B \in \mathcal{B}^k$, of the statistic T is absolutely continuous with respect to the measure ν and its density is $g(y;\theta)$.*

PROOF. It is sufficient to note that relation (4.2.15) implies

$$\mathsf{P}_\theta^T(B) = \int_{T^{-1}(B)} g(T(x);\theta)r(x)\,\mu^n(dx) = \int_B g(y;\theta)\,\nu(dy)$$

where the latter equality follows from the change of variables theorem for the Lebesgue integral. □

LEMMA 4.2.2. *Let G_1 and G_2 be two σ-finite measures on $(\mathbf{R}^k, \mathcal{B}^k)$ and for some parallelepiped $B \subset \mathbf{R}^k$ two integrals exist and are equal:*

$$\int e^{a'y}\,G_1(dy) = \int e^{a'y}\,G_2(dy)$$

for all $a \in B$. Then $G_1 = G_2$.

PROOF. We give the proof for the one-dimensional case, that is, for $k = 1$. Let $B = \{x\colon |x| \le \tilde{a}\}$. Then the functions

$$\psi_j(a) = \int e^{ay}\,G_j(dy), \qquad j = 1, 2,$$

are analytic for $|a| < \tilde{a}$. Moreover for all $b \in \mathbf{R}$ the functions

$$\psi_j(z) = \int e^{(a+ib)y}\,G_j(dy), \qquad j = 1, 2,$$

of a complex variable $z = a + ib$ are well defined. It is clear that $\psi_j(z)$ are analytic functions in the strip $|a| < \tilde{a}$, $-\infty < b < \infty$. By assumption, $\psi_1(z) = \psi_2(z)$ on the interval $|a| < \tilde{a}$ of the line $b = 0$; thus $\psi_1(z) = \psi_2(z)$ for all z of the above strip. Thus for all $b \in (-\infty, \infty)$

$$\int e^{iby}\,G_1(dy) = \int e^{iby}\,G_2(dy). \tag{4.2.16}$$

Since $\psi_j(0) = \int G_j(dy) < \infty$, without loss of generality one can assume that G_j are probability measures. Since the correspondence between characteristic functions and distributions is one-to-one, equality (4.2.16) implies that $G_1 = G_2$.

If the parallelepiped B is of the form $\{x\colon |x - a_0| \le \tilde{a}\}$, then we consider the measures $G_j^*(dy) = e^{a_0 y}\,G_j(dy)$ and follow the line of the above proof.

The proof for the multidimensional case $k > 1$ is the same. □

THEOREM 4.2.9. *Let representation* (4.2.14) *hold for the density $f_n(x;\theta)$ of a sample $\xi^{(n)}$ and let the density belong to a family $\mathcal{P}_n$ of distributions where the function $c(\theta)$ and the set Θ are such that the image of the set Θ under the mapping*

$$c\colon \Theta \to \mathbf{R}^k$$

contains some k-dimensional parallelepiped. Then the statistic T occurring in representation (4.2.14) *is complete and sufficient.*

PROOF. It is sufficient to show that if ϕ is a measurable function on $(\mathbf{R}^k, \mathcal{B}^k)$ and that there exists

$$\int \phi(y)\, \mathsf{P}_\theta^T(dy) = 0, \qquad \theta \in \Theta, \tag{4.2.17}$$

then $\phi(y) = 0$ (P_θ^T-a.s.) for all $\theta \in \Theta$ where P_θ^T is the distribution of the statistic T. Let $\phi = \phi^+ - \phi^-$ where $\phi^+ = 0 \vee \phi$ and $\phi^- = -(0 \wedge \phi)$. Then (4.2.17) implies that

$$\int \phi^+(y)\, \mathsf{P}_\theta^T(dy) = \int \phi^-(y)\, \mathsf{P}_\theta^T(dy) \quad \text{for all } \theta \in \Theta.$$

This yields by Lemma 4.2.1 that

$$\int \phi^+(y) g(y;\theta)\, \nu(dy) = \int \phi^-(y) g(y;\theta)\, \nu(dy) \quad \text{for all } \theta \in \Theta,$$

$$\int \phi^+(y) e^{c(\theta)'y}\, \nu(dy) = \int \phi^-(y) e^{c(\theta)'y}\, \nu(dy) \quad \text{for all } \theta \in \Theta.$$

Consider σ-finite measures $\nu_\pm(dy) = \phi^\pm(y)\, \nu(dy)$. By the assumptions of the theorem we get

$$\int e^{c'y}\, \nu_+(dy) = \int e^{c'y}\, \nu_-(dy)$$

for all c of some parallelepiped in $\mathbf{R}^k$. Now it remains to apply Lemma 4.2.2. □

COROLLARY 4.2.4. *Let all the assumptions of Theorem* 4.2.9 *hold. Let $\widehat{\theta}$ be an estimator of the parameter θ of the class K_b constructed from a sample $\xi^{(n)}$. Then $\widehat{\theta}_T = \mathsf{E}_\theta(\widehat{\theta}/T)$ is an optimal estimator of the parameter θ in the class K_b where T is the statistic occurring in representation* (4.2.14).

COROLLARY 4.2.5. *Let all the assumptions of Theorem* 4.2.9 *hold. Let $\widehat{g}$ be some estimator of a function $g(\theta)$ of a parameter θ of the class K^g constructed from a sample $\xi^{(n)}$. Then $\widehat{g}_T = \mathsf{E}_\theta(\widehat{g}/T)$ is an optimal estimator of the function $g(\theta)$ in the class K^g.*

PROOF. It is sufficient to apply Corollary 4.2.4 to estimators of the class K_b where $b(\theta) = g(\theta) - \theta$ for all $\theta \in \Theta$. □

Some applications of sufficient statistics. In a series of examples above we learned how to construct sufficient, or minimal, or complete and sufficient statistics. We consider in this section some applications of sufficient statistics for several models and construct optimal estimators of a parameter θ and a function $g(\theta)$.

EXAMPLE 4.2.5. Let an observation be a vector $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ where

$$\xi_i = \theta + \eta_i, \qquad i = 1, 2, \dots, n,$$

and random variables $\eta_1, \dots, \eta_n$ (being dependent, generally speaking) do not depend on θ and are such that $\mathsf{E}\eta_i = 0$ and $\mathsf{E}\eta_i^2 < \infty$. Consider the class L of linear unbiased estimators of the parameter θ, that is, the class of functions $l = l(x) = \sum_{i=1}^n c_i x_i$ where $\sum_{i=1}^n c_i = 1$, $x = (x_1, \dots, x_n)$. Our aim is to construct an optimal estimator $l^* = \sum_{i=1}^n c_i^* x_i$ in the class L with respect to the quadratic loss function.

Assume that random variables $\eta_1, \eta_2, \dots$ form an autoregression sequence of first order, more precisely let

$$\eta_1 = \varepsilon_1, \qquad \eta_j = \lambda\eta_{j-1} + \varepsilon_j, \quad j = 2, 3, \dots, \tag{4.2.18}$$

where $\varepsilon_1, \varepsilon_2, \dots$ are independent Gaussian random variables with $\mathsf{E}\varepsilon_j = 0$ and $\mathsf{E}\varepsilon_j^2 = \sigma_j^2$, $0 < \sigma_j^2 < \infty$, for all j; here $\lambda \neq 1$ is a known parameter.

Let $p(u_1, \dots, u_n)$ be the probability density of the vector $(\eta_1, \dots, \eta_n)$ and let $\phi_j(x)$ be the probability density of the normal $\mathcal{N}(0, \sigma_j^2)$ law. Since

$$\varepsilon_1 = \eta_1, \quad \varepsilon_2 = \eta_2 - \lambda\eta_1, \quad \dots, \quad \varepsilon_n = \eta_n - \lambda\eta_{n-1}$$

by (4.2.18), the density $p(u_1, \dots, u_n)$ is of the form

$$p(u_1, \dots, u_n) = \phi_1(u_1)\phi_2(u_2 - \lambda u_1) \cdots \phi(u_n - \lambda u_{n-1}).$$

This implies that the probability density of the vector $(\xi_1, \dots, \xi_n)$ is equal to

$$\begin{aligned} f(x_1, \dots, x_n; \theta) &= p(x_1 - \theta, \dots, x_n - \theta) \\ &= \phi_1(x_1 - \theta)\phi_2(x_2 - \theta - \lambda(x_1 - \theta)) \cdots \phi_n(x_n - \theta - \lambda(x_{n-1} - \theta)) \\ &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(x_1 - \theta)^2}{2\sigma_1^2} - \sum_{j=2}^{n} \frac{(x_j - \theta - \lambda(x_{j-1} - \theta))^2}{2\sigma_j^2} \right\} \\ &= C(\theta) R(x_1, \dots, x_n) \exp\left\{ \theta \left[\frac{x_1}{\sigma_1^2} + (1 - \lambda) \sum_{j=2}^{n} \frac{x_j - \lambda x_{j-1}}{\sigma_j^2} \right] \right\}. \end{aligned} \tag{4.2.19}$$

In what follows we do not need the explicit expressions for $C(\theta)$ and $R(x_1, \dots, x_n)$.

It follows from (4.2.19) that the linear statistic

$$\begin{aligned} l(x) &= \frac{x_1}{\sigma_1^2} + (1 - \lambda) \sum_{j=2}^{n} \frac{x_j - \lambda x_{j-1}}{\sigma_j^2} \\ &= \left(\frac{1}{\sigma_1^2} - \frac{\lambda(1 - \lambda)}{\sigma_2^2} \right) x_1 + (1 - \lambda) \sum_{j=2}^{n-1} \left(\frac{1}{\sigma_j^2} - \frac{1}{\sigma_{j+1}^2} \right) x_j + \frac{1 - \lambda}{\sigma_n^2} x_n \end{aligned} \tag{4.2.20}$$

is sufficient for the parameter θ. It is also seen from (4.2.19) that the linear statistic given by (4.2.20) is complete and sufficient (according to Theorem 4.2.9). It is clear that the optimal linear estimator $l^*(x)$ is a function of the statistic $l(x)$. Thus the coefficients c_j^* are proportional to the corresponding coefficients of the statistic $l(x)$. Thus

$$\begin{aligned} c_1^* &= c \left(\frac{1}{\sigma_1^2} - \frac{\lambda(1 - \lambda)}{\sigma_2^2} \right), \\ c_j^* &= c(1 - \lambda) \left(\frac{1}{\sigma_j^2} - \frac{\lambda}{\sigma_{j+1}^2} \right), \qquad 2 \le j \le n - 1, \\ c_n^* &= c \frac{1 - \lambda}{\sigma_n^2} \end{aligned}$$

where the constant c is defined from the condition $\sum_{j=1}^{n} c_j^* = 1$.

The assumption that the distribution of the random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is Gaussian is not crucial. The same result can be obtained for a general distribution of random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ [15].

EXAMPLE 4.2.6. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the Bernoulli distribution, that is, $\mathsf{P}_\theta(\xi_i = 1) = \theta$ and $\mathsf{P}_\theta(\xi_i = 0) = 1 - \theta$, $i = 1, 2, \ldots, n$, where $\theta \in (0, 1)$ is an unknown parameter. The total number of "successes" $T(x) = \sum_{i=1}^n x_i$ in n Bernoulli trials is a sufficient statistic (see Example 4.1.7). We show that T is a complete statistic.

Note that T has the binomial distribution:

$$\mathsf{P}_\theta\{T = y\} = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \qquad y = 0, 1, 2, \ldots, n.$$

Let $\phi(y)$ be an arbitrary function on $\{0, 1, 2, \ldots, n\}$. Then condition (4.2.7) can be rewritten as

$$\sum_{y=0}^n \phi(y)\binom{n}{y}\theta^y(1-\theta)^{n-y} = 0 \quad \text{for all } \theta \in (0,1)$$

or, putting $x = \theta/(1-\theta)$,

$$\sum_{y=0}^n \phi(y)\binom{n}{y}x^y = 0 \quad \text{for all } x > 0.$$

This implies that all the coefficients of the latter polynomial are zero, that is, $\phi(y) = 0$ for all $y = 0, 1, \ldots, n$. Thus T is a complete sufficient statistic and $\widehat{\theta} = T/n$ is an optimal estimator of the parameter θ. Moreover, according to Corollary 4.2.3 any function of T is an optimal estimator of its own mean.

Since the moment generating function of the random variable T is

$$\mathsf{E}_\theta z^T = \phi(z;\theta) = (1 + (z-1)\theta)^n,$$

we put $(a)_k = a(a-1)\cdots(a-k+1)$, $k \geq 1$, and obtain

$$\mathsf{E}_\theta(T)_k = \left.\frac{\partial^k \phi(z;\theta)}{\partial z^k}\right|_{z=1} = (n)_k\theta^k.$$

This implies that for any integer k, $1 \leq k \leq n$, the statistic $(T)_k/(n)_k$ is an optimal estimator of the function θ^k. At the same time, other functions θ^i with $i > n$ cannot be estimated from a sample of size n in the class of unbiased estimators. Finally, according to Corollary 4.2.3 we get that $\widehat{g} = \sum_{j=1}^n c_j(T)_j/(n)_j$ is an optimal estimator for the polynomial $g(\theta) = \sum_{j=1}^k c_j\theta^j$ if $k \leq n$. Therefore

$$\widehat{\tau} = \frac{T}{n} - \frac{T(T-1)}{n(n-1)} = \frac{T(n-T)}{n(n-1)}$$

is an optimal estimator of the variance $\tau(\theta) = \theta(1-\theta)$ in the case of the binomial distribution.

EXAMPLE 4.2.7. Let a discrete random variable assume values $l, l+1, \dots$ with probabilities

$$f(x;\theta) = \mathsf{P}_\theta\{\xi = x\} = \frac{a(x)\theta^x}{f(\theta)}, \qquad x = l, l+1, \dots, \tag{4.2.21}$$

where $f(\theta) = \sum_{x=l}^{\infty} a(x)\theta^x$ is a series whose radius of convergence R is nonzero. We treat $\theta \in \Theta = (0, R)$ as an unknown parameter. Discrete distributions (4.2.21) are sometimes called *power series distributions* (see [**21**]).

Distributions (4.2.21) include many well-known discrete distributions with an infinite number of values, say Poisson ($f(\theta) = e^\theta$, $R = \infty$), logarithmic ($f(x;\theta) = \theta^x / \ln(1-\theta)^{-x}$, $x = 0, 1, 2, \dots, R = 1$), negative binomial ($f(\theta) = (1-\theta)^{-r}$, $R = 1$), and others as well as their truncated versions.

Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a sample from a distribution (4.2.21). Then the distribution of the sample is of the form

$$f_n(x;\theta) = \mathsf{P}_\theta\left\{\xi^{(n)} = x\right\} = \theta^{T(x)} \prod_{i=1}^{n} a(x_i) f^{-n}(\theta), \qquad \theta \in \Theta, \tag{4.2.22}$$

where $T(x) = \sum_{i=1}^{n} x_i$, $x = (x_1, \dots, x_n)$, $x_i = l, l+1, \dots$, for all $i = 1, 2, \dots, n$. This together with the factorization criterion implies that $T(x)$ is a sufficient statistic.

Since (4.2.21) is a distribution of the exponential type, Theorem 4.2.9 implies that $T(x)$ is a complete and sufficient statistic.

Note also that (4.2.21) yields

$$\mathsf{E}_\theta \xi = \sum_{x=l}^{\infty} x a(x) \theta^x / f(\theta) = \frac{\theta f'(\theta)}{f(\theta)} = \tau(\theta).$$

Then $\widehat{\tau} = n^{-1} T(x) = n^{-1} \sum_{i=1}^{n} x_i$, $x = (x_1, \dots, x_n)$, is an optimal unbiased estimator of the function $\tau(\theta) = \mathsf{E}_\theta \widehat{\tau}$ by Corollary 4.2.3.

Let a function $g(\theta)$ be represented as a power series $g(\theta) = \sum_{j=r}^{\infty} a_j \theta^j$ convergent on Θ. To estimate the function $g(\theta)$ we find the distribution of the statistic $T(x)$. We have

$$g(t;\theta) = \mathsf{P}_\theta\left\{T(\xi^{(n)}) = t\right\} = \sum_{\{x: T(x) = t\}} f_n(x;\theta) = \theta^t b_n(t) f^{-n}(\theta) \tag{4.2.23}$$

where $t = nl, nl+1, \dots$ and

$$b_n(t) = \sum_{x_1 + \dots + x_n = t} a(x_1) \cdots a(x_n).$$

Thus $b_n(t)$ is equal to the coefficient of z^t in the expansion of the function $f^n(z)$. It follows from (4.2.23) that the condition (4.2.12) can be rewritten as

$$\sum_{t=nl}^{\infty} H(t) b_n(t) \theta^t = f^n(\theta) g(\theta) = \sum_{i=nl}^{\infty} b_n(i) \theta^i \sum_{j=r}^{\infty} a_j \theta^j$$

$$= \sum_{k=nl+r}^{\infty} \theta^k \sum_{j=r}^{k-nl} a_j b_n(k-j).$$

This is an identity with respect to θ. Equating the coefficients of θ^t we get

$$H(t)b_n(t) = \begin{cases} \sum_{j=r}^{t-nl} a_j b_n(k-j), & \text{if } t \geq nl+r, \\ 0, & \text{if } t < nl+r. \end{cases}$$

This implies that the optimal estimator g^* of the function $g(\theta)$ is of the form

$$g^* = H(T) = \begin{cases} b_n^{-1}(T) \sum_{j=r}^{T-nl} a_j b_n(T-j), & \text{if } T \geq nl+r, \\ 0, & \text{if } T < nl+r. \end{cases}$$

In particular, if $g(\theta) = \theta^r$ for some $r \geq 1$, then

$$g^* = \begin{cases} b_n(T-r)/b_n(T), & \text{if } T \geq nl+r, \\ 0, & \text{if } T < nl+r. \end{cases}$$

Therefore one can construct optimal estimators for an arbitrary function of the parameter represented as a power series of θ in the case of discrete distributions (4.2.21).

General Methods for Constructing Estimators

In the previous two chapters we dealt with the optimal and efficient statistical estimators of unknown parameters of distributions or of functions of parameters. We considered some methods for constructing optimal and efficient estimators based on some properties of families of distributions. In particular, we considered the method based on sufficient statistics.

In this chapter, we consider general methods of forming estimators, namely the method of moments, the maximum likelihood method, the Bayes method, and the integral estimation method.

5.1. Method of moments

The oldest general method proposed to construct estimators of unknown parameters is the *method of moments* introduced by K. Pearson (1894). This method can be described as follows.

Let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ be a sample from a distribution belonging to a family of distributions $\{\mathsf{P}_\theta, \theta \in \Theta\}$ where $\theta = (\theta_1, \dots, \theta_k)$ is an unknown parameter of a set $\Theta \subset \mathbf{R}^k$, $k \geq 1$. Assume that $\mathsf{E}_\theta |\xi_1|^k < \infty$. Then the following k functions $\alpha_j(\theta) = \alpha_j(\theta_1, \dots, \theta_k) = \mathsf{E}_\theta \xi_1^j$, $j = 1, 2, \dots, k$, are well defined. Let a_j be the sampling moment of order j constructed from the sample $\xi^{(n)}$, that is,

$$a_j = \frac{1}{n} \sum_{i=1}^{n} \xi_i^j, \qquad j = 1, 2, \dots, k.$$

Consider the system of equations

$$\alpha_j(\theta_1, \dots, \theta_k) = a_j, \qquad j = 1, 2, \dots, k, \tag{5.1.1}$$

with unknowns $\theta_1, \dots, \theta_k$. Solutions $\widehat{\theta}_j$, $j = 1, 2, \dots, k$, of system (5.1.1), if they exist, are called *method of moments estimators of a parameter*.

Note that $\mathsf{E}_\theta |\xi_1|^j < \infty$ for all $j = 1, 2, \dots, k$ if $\mathsf{E}_\theta |\xi_1|^k < \infty$. Using results of Section 1.2 we obtain $a_j \to \alpha_j(\theta_1, \dots, \theta_k)$ as $n \to \infty$ in probability P_θ for all $j = 1, 2, \dots, k$. Assume that the functions $\alpha_j(\theta_1, \dots, \theta_k)$, $j = 1, 2, \dots, k$, determine a continuous one-to-one correspondence between vectors $(\theta_1, \theta_2, \dots, \theta_k)$ and $(a_1, \dots, a_k)$, that is, there exist continuous functions ϕ_j, $j = 1, 2, \dots, k$, such that $\alpha_j = \phi_j(a_1, \dots, a_k)$, $j = 1, 2, \dots, k$. Then solutions of system (5.1.1) can be represented as

$$\widehat{\theta}_j = \phi_j(a_1, \dots, a_k), \qquad j = 1, 2, \dots, k. \tag{5.1.2}$$

Thus estimators (5.1.2) are consistent by Theorem 1.2.2.

The sampling moments a_j are asymptotically normal if $\mathsf{E}_\theta |\xi_1|^{2j} < \infty$ (see Theorem 1.2.3). Moreover, we noticed in Section 1.2 that the asymptotic normality

can be proved for a continuous function of a finite number of sampling moments a_j provided suitable conditions on the function are posed (more details are given in [**5, 9**]). Therefore conditions can be posed on moments of random variables ξ_i to guarantee that estimators (5.1.2) are asymptotically normal.

On the other hand, Fisher (1921) pointed out that estimators (5.1.2) are not asymptotically efficient. Moreover, the method of moments cannot be applied in the cases where the corresponding moments do not exist (say, in the case of the Cauchy distribution). Nevertheless the method of moments has an advantage because of its practical expediency. Estimators (5.1.2) can be treated sometimes as a first approximation used for other methods to construct estimators of a higher efficiency.

REMARK 5.1.1. One can also use another form of the method of moments, namely one can use the moments $m_j(\theta) = \mathsf{E}_\theta g_j(\xi_1)$, $j = 1, 2, \ldots, k$, instead of moments $\alpha_j(\theta)$, $j = 1, 2, \ldots, k$, where $g_j(x)$, $j = 1, \ldots, k$, are some measurable functions. Then we obtain the system of equations

$$m_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_j(\xi_i), \qquad j = 1, 2, \ldots, k, \tag{5.1.3}$$

instead of (5.1.1). Solutions of system (5.1.3) are also called *method of moments estimators* (more detail is given in [**5**]). Note that system (5.1.3) reduces to (5.1.1) if $g_j(x) = x^j$, $j = 1, 2, \ldots, k$.

EXAMPLE 5.1.1. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the Gamma distribution, so that the density is

$$f(x;\theta) = \frac{1}{\Gamma(\theta)} x^{\theta-1} e^{-x} I_{(0,\infty)}(x)$$

where $\theta \in \Theta = (0, \infty)$ is an unknown parameter. In this case $\alpha_1(\theta) = \mathsf{E}_\theta \xi_1 = \theta$ and therefore a solution of the equation

$$\alpha_1(\theta) = a_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

is a method of moments estimator and it is given by $\widehat{\theta} = n^{-1}\sum_{i=1}^{n} \xi_i$. This estimator is unbiased and consistent. On the other hand, this estimator is not asymptotically efficient whatever the parameter θ is (see Example 3.4.3).

EXAMPLE 5.1.2. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from a distribution with the density $f(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$ where $\theta \in \Theta = (0, \infty)$ is an unknown parameter. We use two functions $g_1(x) = x$ and $g_2(x) = x^2$ to construct estimators according to the method of moments (see Remark 5.1.1). Since $m_j(\theta) = \mathsf{E}_\theta g_j(\xi_1) = \mathsf{E}_\theta \xi_1^j = j\theta^{-j}$ for $j = 1, 2$, equations (5.1.3) are of the form

$$m_1(\theta) = a_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i,$$

$$m_2(\theta) = a_2 = \frac{1}{n}\sum_{i=1}^{n} \xi_i^2.$$

There are two solutions of these equations with respect to θ:

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)^{-1}, \qquad \widetilde{\theta} = \left(\frac{1}{2n}\sum_{i=1}^{n}\xi_i^2\right)^{-1/2}$$

Every solution is a method of moments estimator of the parameter θ. One can show (see [5]) that both estimators $\widehat{\theta}$ and $\widetilde{\theta}$ are asymptotically normal with parameters $\mathcal{N}(\theta,\, n^{-1}\theta^2)$ and $\mathcal{N}(\theta, (5/4)n^{-1}\theta^2)$, respectively. Thus the estimator $\widehat{\theta}$ is better than $\widetilde{\theta}$, since $n^{-1}\theta^2 < (5/4)n^{-1}\theta^2$. The Fisher information of $\widehat{\theta}$ is $I(\theta) = \theta^{-2}$, whence it follows that $\widehat{\theta}$ is asymptotically efficient in the weak sense.

EXAMPLE 5.1.3. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the Gamma distribution with the density

$$f(x;\theta) = \frac{1}{\theta_1^{\theta_2}\Gamma(\theta_2)} x^{\theta_2 - 1} e^{-x/\theta_1} I_{(0,\infty)}(x)$$

where $\theta = (\theta_1, \theta_2) \in \Theta = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0\}$. It is clear that

$$\alpha_j(\theta) = \mathsf{E}_\theta \xi_1^j = \theta_1^j \frac{\Gamma(\theta_2 + j)}{\Gamma(\theta_2)} = \theta_1^j \theta_2(\theta_2 + 1)\cdots(\theta_2 + j - 1).$$

In particular, $\alpha_1(\theta) = \theta_1\theta_2$ and $\alpha_2(\theta) = \theta_1^2\theta_2(\theta_2 + 1)$. Thus solutions of the system of equations (5.1.1) are of the form

$$\widehat{\theta}_1 = \frac{a_2 - a_1^2}{a_1}, \qquad \widehat{\theta}_2 = \frac{a_1^2}{a_2 - a_1^2}.$$

REMARK 5.1.2. When considering maximum likelihood estimators for samples from the normal distribution, we will show that estimators obtained by the method of moments and by the maximum likelihood method coincide and both are efficient. This is an exceptional case where the method of moments estimators are efficient.

5.2. The maximum likelihood method

From a theoretical point of view, the most important general method of estimation of parameters is the *method of moments*. In particular cases, this method is already used by F. Gauss. As a general method of estimation it was first introduced by Fisher (1912) and afterwards it was further developed by the same author. In 1925 Fisher studied asymptotic properties of maximum likelihood estimators.

Maximum likelihood estimators. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$ and whose distribution is determined by a measure of a family of measures $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ where

$$\theta = (\theta_1, \ldots, \theta_k)$$

is an unknown parameter belonging to a set $\Theta \subset \mathbf{R}^k$, $k \geq 1$. Let the family $\mathcal{P}$ be dominated by a σ-finite measure μ and let $f(x;\theta)$ be the density of the measure P_θ with respect to the measure μ. The function $f(x;\theta)$ of an argument θ is called the *likelihood function*, while $L(x;\theta) = \ln f(x;\theta)$ is called the *logarithmic likelihood function*.

A statistic $\widehat{\theta} = \widehat{\theta}(x)$ such that

$$L(x; \widehat{\theta}(x)) = \sup_{\theta \in \Theta} L(x; \theta), \qquad x \in X, \tag{5.2.1}$$

is called the *maximum likelihood estimator of the parameter* θ if such a point

$$\widehat{\theta}(x) \in \Theta$$

exists. Otherwise, if there is no $\widehat{\theta}(x)$ satisfying (5.2.1), then we take an arbitrary point of Θ as $\widehat{\theta}(x)$. If the function $L(x; \theta)$ is continuous with respect to θ and the set Θ is close, then the supremum on the right-hand side of (5.2.1) is attained, thus the maximum likelihood estimator is well defined.

The maximum likelihood estimator $\widehat{\theta}_n$ can be defined in an equivalent way as a statistic maximizing the likelihood function $f(x; \theta)$. Below we consider the case where the likelihood functions are differentiable with respect to θ. One can substitute the closure $\overline{\Theta}$ instead of Θ in (5.2.1) in this case, the supremum on the right-hand side of (5.2.1) is attained, and the maximum likelihood estimator exists. If the supremum on the right-hand side of (5.2.1) is attained at an interior point of Θ and the function $L(x; \theta)$ is differentiable with respect to θ, then one can seek the maximum likelihood estimator $\widehat{\theta}$ among solutions of the system of equations

$$\frac{\partial}{\partial \theta_j} L(x; \theta) = 0, \qquad j = 1, 2, \ldots, k. \tag{5.2.2}$$

The equations of system (5.2.2) are called *likelihood equations*.

The following are two properties of the maximum likelihood estimators:

1) *if there exists an efficient unbiased estimator* $T = T(x)$ *of a scalar parameter* θ, *then the maximum likelihood estimator* $\widehat{\theta}$ *exists and coincides with the estimator* T;
2) *if there exists a sufficient statistic* $T = T(x)$ *and the maximum likelihood estimator exists and is unique, then* $\widehat{\theta}$ *is a function of* T.

It is sufficient to apply Theorem 3.4.1 (or Theorem 3.4.3) to prove property 1), and the factorization criterion to prove property 2).

Consider some examples of maximum likelihood estimators.

EXAMPLE 5.2.1. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the normal $\mathcal{N}(\theta_1, \theta_2)$ distribution where $\theta = (\theta_1, \theta_2)$ is an unknown parameter such that $\theta_1 \in (-\infty, \infty)$ and $\theta_2 > 0$. Then the logarithmic likelihood function for the distribution of the sample is given by

$$L(x; \theta) = -\frac{n}{2} \ln(2\pi\theta_2) - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2, \qquad x = (x_1, \ldots, x_n).$$

The system of likelihood equalities (5.2.2) is of the form in this case:

$$\frac{\partial L(x; \theta)}{\partial \theta_1} = \frac{1}{\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2 = 0,$$

$$\frac{\partial L(x; \theta)}{\partial \theta_2} = \frac{1}{2\theta_2^2} \sum_{i=1}^{n} (x_i - \theta_1)^2 - \frac{n}{2\theta_2} = 0.$$

Solving this system of equations with respect to θ_1 and θ_2 we obtain the following maximum likelihood estimators:

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\theta}_1)^2. \tag{5.2.3}$$

It is easy to see that these estimators coincide with the estimators of the parameters θ_1 and θ_2 obtained by the method of moments. Further, it is easy to check that the maximum of the function $L(x;\theta)$ is attained at the point $(\widehat{\theta}_1, \widehat{\theta}_2)$. Thus the maximum likelihood estimator exists, is unique, and is defined by (5.2.3). Note that this maximum likelihood estimator is a function of the sufficient statistic $T = (T_1, T_2)$ considered in Example 4.1.3. Moreover the maximum likelihood estimator $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ itself is a sufficient statistic (see Example 4.1.3).

EXAMPLE 5.2.2. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the uniform distribution on the interval $[0, \theta]$ where $\theta > 0$ is an unknown parameter. The likelihood function in this case is of the form

$$f(x;\theta) = \theta^{-n} I_{[0,\infty)}(x_{n,1}) I_{(-\infty,\theta]}(x_{n,n})$$

(see Example 4.1.5). The function $f(x;\theta)$ is discontinuous with respect to θ and moreover $f(x;\theta) = 0$ for $\theta < x_{n,n}$ and $f(x;\theta) = \theta^{-n}$ for $\theta \geq x_{n,n}$ if $x_{n,1} \geq 0$. Then the maximum likelihood estimator is $\widehat{\theta} = x_{n,n}$. If $x_{n,1} < 0$, then $f(x;\theta) = 0$ for all θ and any number can be taken as the maximum likelihood estimator, in particular one can put $\widehat{\theta} = x_{n,n}$. If the sampling space is $\mathbf{R}_+^n$, then the maximum likelihood estimator $\widehat{\theta} = x_{n,n}$ is unique and is a complete sufficient statistic (see Example 4.2.4). We learned from Example 4.2.4 that the estimator $\widetilde{\theta} = \frac{n+1}{n} x_{n,n}$ is optimal in the class of unbiased estimators of the parameter θ.

EXAMPLE 5.2.3. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the uniform distribution on the interval $[\theta, \theta + 1]$ where $\theta \in (-\infty, \infty)$ is an unknown parameter. The likelihood function in this case is given by

$$f(x;\theta) = I_{[\theta,\infty)}(x_{n,1}) I_{(-\infty,\theta+1]}(x_{n,n}), \quad x = (x_1, \ldots, x_n).$$

The maximum likelihood estimator is not unique in this case. In particular, $\widehat{\theta} = x_{n,1}$ is one of the maximum likelihood estimators, another one is $\widetilde{\theta} = x_{n,n} - 1$. Note that $T(x) = (x_{n,1}, x_{n,n})$ is a sufficient statistic in this case.

The invariance principle for maximum likelihood estimators. The following result is known as the invariance principle for maximum likelihood estimators with respect to the change of a parameter.

THEOREM 5.2.1. *Let $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ be a family of probability measures on $(X, \mathcal{B})$ defining the distribution of an observation ξ, and let $g = g(\theta)$ be a one-to-one mapping of Θ into some set G. If $\widehat{\theta}$ is a maximum likelihood estimator of the parameter θ constructed from an observation ξ, then $\widehat{g} = g(\widehat{\theta})$ is a maximum likelihood estimator of the function $g(\theta)$ constructed from the observation ξ.*

PROOF. Let $\theta(\gamma)$ be the inverse function to $g(\theta)$ and let $\mathbf{Q}_\gamma = \mathsf{P}_{\theta(\gamma)}$ for all $\gamma \in G$. Then the logarithmic likelihood function for the family $\{\mathbf{Q}_\gamma, \gamma \in G\}$ is of the form

$$M(\gamma; x) = \ln(d\mathbf{Q}_\gamma/d\mu(x)) = L(\theta(\gamma); x). \tag{5.2.4}$$

Let $\widehat{\theta}$ be the maximum likelihood estimator of the parameter θ and let $\widehat{g}$ be the maximum likelihood estimator of the parameter γ. Then equality (5.2.4) implies that $\widehat{\theta} = \theta(\widehat{g})$ or, equivalently, $\widehat{g} = g(\widehat{\theta})$. □

EXAMPLE 5.2.4. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the logarithmic distribution, that is, $\mathcal{L}(\ln \xi_1) = \mathcal{N}(\theta_1, \theta_2)$ where $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. Let $\theta = (\theta_1, \theta_2)$. It is not hard to show that

$$\gamma_1 = \mathsf{E}_\theta \xi_1 = \exp\left\{\theta_1 + \frac{1}{2}\theta_2\right\}, \qquad \gamma_2 = \mathsf{D}_\theta \xi_1 = \theta_1^2 \left(e^{\theta_2} - 1\right).$$

Consider the function $g(\theta_1, \theta_2) = (\gamma_1, \gamma_2)$ and find the maximum likelihood estimator $(\widehat{\gamma}_1, \widehat{\gamma}_2)$ of the function $g(\theta_1, \theta_2)$. We obtain from Example 5.2.1 that the maximum likelihood estimator $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ is of the form

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n \eta_i, \qquad \widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (\eta_i - \widehat{\theta}_1)^2$$

where $\eta_i = \ln \xi_i$. We obtain from the invariance principle for maximum likelihood estimators that

$$\widehat{\gamma}_1 = \exp\left\{\widehat{\theta}_1 + \frac{1}{2}\widehat{\theta}_2\right\}, \qquad \widehat{\gamma}_2 = \widehat{\theta}_1^2 \left(e^{\widehat{\theta}_2} - 1\right).$$

Asymptotic properties of maximum likelihood estimators. In this section we consider asymptotic properties of maximum likelihood estimators, namely we prove that a maximum likelihood estimator is consistent, asymptotically normal, and asymptotically efficient.

Let an observation $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from a distribution belonging to a family $\{\mathsf{P}_\theta, \theta \in \Theta\}$ dominated by some σ-finite measure μ. Denote by $f(x; \theta)$ the density of the measure P_θ with respect to μ. Then $f_n(x; \theta) = \prod_{i=1}^n f(x_i; \theta)$, $x = (x_1, \ldots, x_n)$, is the likelihood function, while $L_n(x; \theta) = \sum_{i=1}^n \ln f(x_i; \theta)$ is the logarithmic likelihood function. We denote by $\widehat{\theta}_n = \widehat{\theta}_n(x)$ the maximum likelihood estimator.

First we consider the case of a one-dimensional parameter θ, that is, we consider the case $k = 1$. The following result is an assertion on asymptotic properties of the maximum likelihood estimator $\widehat{\theta}_n$ of a one-dimensional parameter θ.

THEOREM 5.2.2. *Let Θ be an open interval. Assume that*

1) *for all $\theta \in \Theta$ there exist the derivatives $\partial^j \ln f(x; \theta)/\partial\theta^j$, $j = 1, 2, 3$, for μ-almost all x;*
2) *for all $\theta \in \Theta$ there exist nonnegative functions $F_1(x)$, $F_2(x)$, and $H(x)$ depending on x and such that*

$$\left|\frac{\partial^j f(x;\theta)}{\partial\theta^j}\right| \le F_j(x), \quad j = 1, 2; \qquad \left|\frac{\partial^3 \ln f(x;\theta)}{\partial\theta^3}\right| \le M(x);$$

the functions $F_j(x)$, $j = 1, 2$, *are integrable with respect to the measure* μ *and* $\mathsf{E}_\theta H(\xi_1) \le M < \infty$ *where the constant* M *does not depend on* θ;

3) $0 < I(\theta) = \mathsf{E}_\theta(\partial \ln f(\xi_1;\theta)/\partial\theta)^2 < \infty$ *for all* $\theta \in \Theta$.

Then the likelihood equation (5.2.2) *has a solution* $\widehat{\theta}_n = \widehat{\theta}_n(\xi^{(n)})$ *that converges in probability* P_{θ_0} *to the true value* θ_0 *of the parameter*; *moreover the maximum likelihood estimator* $\widehat{\theta}_n$ *is an asymptotically normal and asymptotically efficient estimator of the parameter* θ.

PROOF. Let $\theta_0 \in \Theta$ be the true value of the parameter θ. First we show that there exists a solution of the likelihood equation that converges in probability to θ_0.

Expanding $\partial L_n(\xi^{(n)};\theta)/\partial\theta$ into the Taylor series in a neighborhood of the point $\theta = \theta_0$ we get

$$\begin{aligned}\frac{\partial L_n(\xi^{(n)};\theta)}{\partial\theta} &= \sum_{i=1}^n \frac{\partial \ln f(\xi_i;\theta)}{\partial\theta} \\ &= \sum_{i=1}^n \left[\left(\frac{\partial \ln f(\xi_i;\theta)}{\partial\theta}\right)_{\theta_0} + (\theta-\theta_0)\left(\frac{\partial^2 \ln f(\xi_i;\theta)}{\partial\theta^2}\right)_{\theta_0} \right. \\ &\qquad \left. + \frac{1}{2}\lambda_i(\theta-\theta_0)^2 H(\xi_i)\right]\end{aligned} \tag{5.2.5}$$

where $|\lambda_i| \le 1$ and the symbol $(\cdot)_{\theta_0}$ stands for $(\phi(\theta))_{\theta_0} = \phi(\theta_0)$. Substituting (5.2.5) into the likelihood equation (5.2.2) we get

$$B_0 + B_1(\theta-\theta_0) + \frac{1}{2}\lambda B_2(\theta-\theta_0)^2 = 0 \tag{5.2.6}$$

where $|\lambda| \le 1$ and

$$\begin{aligned} B_j &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial^{j+1} \ln f(\xi_i;\theta)}{\partial\theta^{j+1}}\right)_{\theta_0}, \quad j = 0, 1, \\ B_2 &= \frac{1}{n}\sum_{i=1}^n H(\xi_i). \end{aligned} \tag{5.2.7}$$

Now we show that P_{θ_0} approaches 1 as $n \to \infty$ for an arbitrary $\delta > 0$ where P_{θ_0} is the probability that equation (5.2.6) has a solution on the interval $(\theta_0 - \delta, \theta_0 + \delta)$. To prove this result we study the limit behavior of B_j as $n \to \infty$.

Assumptions 1) and 2) imply that

$$\int \frac{\partial f(x;\theta)}{\partial\theta}\,\mu(dx) = \int \frac{\partial^2 f(x;\theta)}{\partial\theta^2}\,\mu(dx) = 0.$$

Thus

$$\mathsf{E}_{\theta_0}\left(\frac{\partial \ln f(\xi_1;\theta)}{\partial\theta}\right)_{\theta_0} = \int \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)_{\theta_0} \mu(dx) = 0, \tag{5.2.8}$$

$$\begin{aligned}\mathsf{E}_{\theta_0}&\left(\frac{\partial^2 \ln f(\xi_1;\theta)}{\partial\theta^2}\right)_{\theta_0} \\ &= \int \left(\frac{\frac{\partial^2 f(x;\theta)}{\partial\theta^2}}{f(x;\theta)} - \left(\frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)}\right)^2\right)_{\theta_0} f(x;\theta_0)\,\mu(dx) \\ &= -\int \left(\frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)}\right)^2_{\theta_0} f(x;\theta_0)\,\mu(dx) \\ &= -\mathsf{E}_{\theta_0}\left(\frac{\partial \ln f(\xi_1;\theta)}{\partial\theta}\right)^2_{\theta_0} = -I(\theta_0).\end{aligned} \tag{5.2.9}$$

According to the law of large numbers we have $B_0 \to 0$ and $B_1 \to -I(\theta_0)$ as $n \to \infty$ in probability P_{θ_0}. Similarly $B_2 \to \mathsf{E}_{\theta_0} H(\xi_1) \le M < \infty$ as $n \to \infty$ in probability P_{θ_0}.

Let $\delta > 0$ and $\varepsilon > 0$ be two arbitrary numbers. Then there exists an integer number $n_0 = n_0(\delta, \varepsilon)$ such that

$$p_1 = \mathsf{P}_{\theta_0}\{|B_0| \ge \delta^2\} < \frac{\varepsilon}{3},$$

$$p_2 = \mathsf{P}_{\theta_0}\left\{B_1 \ge -\frac{1}{2}I(\theta_0)\right\} < \frac{\varepsilon}{3},$$

$$p_3 = \mathsf{P}_{\theta_0}\{B_1 \ge 2M\} \le \frac{\varepsilon}{3}$$

for $n > n_0$. Consider events $S = \left\{|B_0| < \delta^2, B_1 < \frac{1}{2}I(\theta_0), B_2 < 2M\right\}$. Then

$$\mathsf{P}_{\theta_0}(\Omega \setminus S) \le p_1 + p_2 + p_3 < \varepsilon,$$

whence $\mathsf{P}_{\theta_0}(S) > 1 - \varepsilon$ for all $n > n_0$.

The left-hand side of equality (5.2.6) is equal to $B_0 \pm B_1\delta + 2^{-1}\lambda B_2\delta^2$ at the point $\theta = \theta_0 \pm \delta$. If the event S occurs, then $|B_0 + 2^{-1}\lambda B_2\delta^2| < \delta^2 + |\lambda| M\delta^2 < (M+1)\delta^2$ and $B_1\delta < -2^{-1}I(\theta_0)\delta$. Thus the sign of $B_0 \pm B_1\delta + 2^{-1}\lambda B_2\delta^2$ is defined by the second term if $\delta < I(\theta_0)2^{-1}(M+1)^{-1}$, whence $\partial L_n(\xi^{(n)};\theta)/\partial\theta > 0$ for $\theta = \theta_0 - \delta$ and $\partial L_n(\xi^{(n)};\theta)/\partial\theta < 0$ for $\theta = \theta_0 + \delta$. Assumption 1) implies that $\partial L_n(x;\theta)/\partial\theta$ is a continuous function of $\theta \in \Theta$ for μ-almost all x. Therefore for arbitrary $\delta > 0$ and $\varepsilon > 0$ the likelihood equation (5.2.6) has, with probability greater than $1-\varepsilon$, a solution belonging to the interval $(\theta_0 - \delta, \theta_0 + \delta)$ if $n > n_0(\delta, \varepsilon)$. This proves the first part of the theorem.

Now we prove that the maximum likelihood estimator $\widehat{\theta}_n$ is asymptotically normal and asymptotically efficient.

Let $\widehat{\theta}_n = \widehat{\theta}_n(\xi^{(n)})$ be a solution of the likelihood equation. It follows from (5.2.6) and (5.2.7) that

$$\widehat{\theta}_n - \theta_0 = \frac{B_0}{-B_1 - 2^{-1}\lambda B_2(\widehat{\theta}_n - \theta_0)}.$$

This implies that

$$\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{nI(\theta_0)}} \sum_{i=1}^{n} \left(\frac{\partial \ln f(\xi_i;\theta)}{\partial \theta} \right)_{\theta_0}}{-\frac{B_1}{I(\theta_0)} - \frac{1}{2}\lambda B_2 \frac{\widehat{\theta}_n - \theta_0}{I(\theta_0)}}. \tag{5.2.10}$$

Thus the denominator on the right-hand side of (5.2.10) converges in probability P_{θ_0} to 1 as $n \to \infty$. Every term of the sum in the numerator on the right-hand side of (5.2.10) has expectation 0 and variance $I(\theta_0)$ according to equalities (5.2.8) and (5.2.9). Thus the central limit theorem implies that the numerator on the right-hand side of (5.2.10) is asymptotically $\mathcal{N}(0,1)$ normal. Hence (5.2.10) implies that the random variable $\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0)$ is asymptotically $\mathcal{N}(0,1)$ normal. Therefore the estimator $\widehat{\theta}_n$ is asymptotically efficient. □

Asymptotic properties of the maximum likelihood estimator $\widehat{\theta}_n$ of a multidimensional parameter $\theta = (\theta_1, \ldots, \theta_k)$ are listed in the following result.

THEOREM 5.2.3. *Let Θ be an open nondegenerate k-dimensional parallelepiped. Assume that*

1) *for all $\theta \in \Theta$ and for μ-almost all x there exist partial derivatives up to third order inclusive of the function $\ln f(x;\theta)$ with respect to θ;*
2) *for all $\theta \in \Theta$*

$$\left| \frac{\partial f(x;\theta)}{\partial \theta_p} \right| \le F_1(x), \quad \left| \frac{\partial^2 f(x;\theta)}{\partial \theta_p \partial \theta_q} \right| \le F_2(x),$$

$$\left| \frac{\partial^3 \ln f(x;\theta)}{\partial \theta_p \partial \theta_q \partial \theta_s} \right| \le H(x),$$

the functions $F_j(x)$ are integrable with respect to the measure μ, and there is a constant M such that $\mathsf{E}_\theta H(\xi_1) \le M < \infty$ for all θ;

3) *for all $\theta \in \Theta$ the matrix*

$$B(\theta, \theta_0) = \left\| \mathsf{E}_{\theta_0} \frac{\partial \ln f(\xi_1;\theta)}{\partial \theta_p} \frac{\partial \ln f(\xi_1;\theta)}{\partial \theta_q} \right\|$$

is nonsingular and $\det B(\theta, \theta_0) < \infty$.

Then the system of likelihood equations (5.2.2) *has a solution that is a consistent, asymptotically $\mathcal{N}(\theta_0, n^{-1}B^{-1}(\theta_0, \theta_0))$ normal, and asymptotically efficient estimator of the parameter θ where θ_0 is the true value of the parameter.*

The proof of Theorem 5.2.3 is similar to that of Theorem 5.2.2, thus we omit it.

EXAMPLE 5.2.5. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the normal $\mathcal{N}(\theta_1, \theta_2)$ distribution where $\theta = (\theta_1, \theta_2)$ is an unknown parameter, $-\infty < \theta_1 < \infty$, $\theta_2 > 0$. We learned in Example 5.2.1 that the maximum likelihood estimators of the parameter θ are of the form (5.2.3), and moreover they coincide with the estimators obtained by the method of moments. By Theorem 5.2.3 estimators (5.2.3) are consistent, asymptotically $\mathcal{N}(\theta_0, n^{-1}I^{-1}(\theta_0))$ normal, and asymptotically efficient. Here θ_0 is the true value of the parameter and the Fisher information matrix $I(\theta)$ is of the form

$$I(\theta) = \begin{pmatrix} \theta_2^{-1} & 0 \\ 0 & 2^{-1}\theta_2^{-2} \end{pmatrix}$$

(see Example 3.5.1). Note that the estimator $\widehat{\theta}_1$ defined in (5.2.3) is unbiased, while the estimator $\widehat{\theta}_2$ is biased. On the other hand, we learned in Example 3.5.1 that the estimator $\widehat{\theta}_n = (\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n})$ is unbiased and asymptotically efficient. Moreover it is easy to see that the estimator $\widehat{\theta}_n$ is also asymptotically $\mathcal{N}(\theta_0, n^{-1}I^{-1}(\theta_0))$ normal.

EXAMPLE 5.2.6. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a sample from the Gamma distribution, so that the density is

$$f(x;\theta) = \frac{1}{\Gamma(\theta)} x^{\theta-1} e^{-x} I_{(0,\infty)}(x)$$

where $\theta \in \Theta = (0, \infty)$ is an unknown parameter. We showed in Examples 3.4.3 and 5.1.1 that the estimator $\widehat{\theta}_n = n^{-1} \sum_{i=1}^n \xi_i$ of the parameter θ obtained by the method of moments is unbiased and consistent but it is not asymptotically efficient whatever the parameter θ is. On the other hand, the maximum likelihood method leads to the equation

$$\frac{1}{n} \sum_{i=1}^n \ln \xi_i - \frac{d \ln \Gamma(\theta)}{d\theta} = 0 \tag{5.2.11}$$

and the maximum likelihood estimator $\widetilde{\theta}_n$ is a unique positive solution of this equation. According to Theorem 5.2.2 the estimator $\widetilde{\theta}_n$ is asymptotically

$$\mathcal{N}\left(\theta_0, \left(n \frac{d^2 \ln \Gamma(\theta)}{d\theta^2}\right)_{\theta_0}^{-1}\right)$$

normal and asymptotically efficient. This can easily be obtained explicitly from equation (5.2.11), since

$$\mathsf{E}_\theta \ln \xi_1 = \frac{d \ln \Gamma(\theta)}{d\theta}, \qquad \mathsf{D}_\theta \ln \xi_1 = \frac{d^2 \ln \Gamma(\theta)}{d\theta^2}$$

and the random variable $n^{-1} \sum_{i=1}^n \ln \xi_i$ is asymptotically

$$\mathcal{N}\left(\left(\frac{d \ln \Gamma(\theta)}{d\theta}\right)_{\theta_0}, \left(\frac{1}{n} \frac{d^2 \ln \Gamma(\theta)}{d\theta^2}\right)_{\theta_0}\right)$$

normal by the central limit theorem.

Applications of regularity conditions for families of distributions for studying asymptotic properties of maximum likelihood estimators. If the derivatives up to third order of the logarithmic likelihood function $\ln f(x;\theta)$ exist, then Theorems 5.2.2 and 5.2.3 show that the maximum likelihood estimator $\widehat{\theta}_n$ is consistent, asymptotically normal, and asymptotically efficient under some extra assumptions that are, in fact, not necessary for these properties.

First we consider sufficient conditions for the consistency of the maximum likelihood estimator $\widehat{\theta}_n$. For all sets $A \subset \mathbf{R}^k$ such that $A \cap \Theta \neq \varnothing$ put

$$f_n^*(x; A) = \sup\{f_n(x;\theta); \theta \in \Theta \cap A\}$$

where $x = (x_1, \dots, x_n)$. For $n = 1$ we have $f^*(x; A) = f_1^*(x; A)$.

The following result contains sufficient conditions for the consistency of the maximum likelihood estimator $\widehat{\theta}_n$ as $n \to \infty$.

THEOREM 5.2.4. *Let $\theta_0 \in \Theta$ be the true value of the parameter. Assume that*

1) *if $\theta \neq \theta_0$, then $\int |f(x;\theta) - f(x;\theta_0)|\, \mu(dx) > 0$;*
2) *for all x the density $f(x;\theta)$ is a semicontinuous function with respect to θ on the set Θ, that is, for all $\theta' \in \Theta$*

$$\lim_{h \to 0} \sup \{f(x;\theta); |\theta - \theta'| < h\} = f(x;\theta');$$

3) *for some r*

$$\mathsf{E}_{\theta_0} \ln\bigl(f_r(\xi^{(r)};\theta_0)/f_r^*(\xi^{(r)};H)\bigr) > -\infty.$$

If H is a compact subset of Θ containing the point θ_0, then there exists $\widehat{\theta}_n \in H$ such that

$$f_n(x;\widehat{\theta}_n) = f_n^*(x;H), \qquad x = (x_1, \ldots, x_n),$$

and $\widehat{\theta}_n \to \theta_0$ with P_{θ_0}-probability 1 as $n \to \infty$. Moreover,

4) *if additionally*

$$\mathsf{E}_{\theta_0} \ln\bigl(f_r(\xi^{(r)};\theta_0)/f_r^*(\xi^{(r)};\Theta \setminus H)\bigr) > 0,$$

then with P_{θ_0}-probability 1 the likelihood function has the global maximum at the point $\widehat{\theta}_n$ if n is sufficiently large, that is, $f_n(x;\widehat{\theta}_n) = f_n^*(x;\Theta)$.

The proof of Theorem 5.2.4 can be found in [**25**]. Note that Theorem 5.2.4 claims that the maximum likelihood estimator $\widehat{\theta}_n$ approaches θ_0 with P_{θ_0}-probability 1 as $n \to \infty$. We say in this case that $\widehat{\theta}_n$ is a *strongly consistent* estimator of the parameter.

Analyzing the assumptions of Theorem 5.2.4 one can see that the continuity of the function $f(x;\theta)$ with respect to θ is close to being a necessary condition for the consistency of the maximum likelihood estimator $\widehat{\theta}_n$.

Now we discuss sufficient conditions for the asymptotic normality and asymptotic efficiency of the maximum likelihood estimator $\widehat{\theta}_n$. We consider the case where θ is a one-dimensional parameter.

THEOREM 5.2.5. *Let $\theta_0 \in \Theta$ be the true value of the parameter. Assume that*

1) *$f(x;\theta)$ is a measurable function with respect to the pair of variables $(x;\theta)$ and $\int |f(x;\theta) - f(x;\theta')|\, \mu(dx) > 0$ for all $\theta \neq \theta'$;*
2) *for all x the function $f(x;\theta)$ is absolutely continuous with respect to θ and $\int |\partial f(x;\theta)/\partial\theta|\, d\theta < \infty$ for μ-almost all x;*
3) *$\mathsf{E}_{\theta_0} |\partial \ln f(\xi_1;\theta)/\partial\theta|^{2+\delta}$ for some $\delta > 0$ and all $\theta \in \Theta$; the Fisher information $I(\theta) = \mathsf{E}_\theta (\partial \ln f(\xi_1;\theta)/\partial\theta)^2$ is a continuous function such that $I(\theta) \leq C(1 + |\theta|^p)$ for some $C > 0$ and $p \geq 0$;*
4) *$\sup_\theta |\theta - \theta_0|^\gamma \int (f(x;\theta) f(x;\theta_0))^{1/2}\, \mu(dx) < \infty$ for some $\gamma > 0$.*

Then the maximum likelihood estimator $\widehat{\theta}_n$ is asymptotically $\mathcal{N}(\theta_0, (nI(\theta_0))^{-1})$ normal and asymptotically efficient as $n \to \infty$.

The proof of Theorem 5.2.5 can be found in [**13**].

Note that the assumptions of Theorem 5.2.5 can be weakened (see [**13**]). Moreover Theorem 5.2.5 can be proved for a multidimensional parameter. We also note that the asymptotic normality and asymptotic efficiency of the maximum likelihood estimator $\widehat{\theta}_n$ is proved in [**5**] under conditions weaker than those in Theorems 5.2.2 and 5.2.3 but stronger than those in Theorem 5.2.5. In particular, it is assumed

in [5] that the function $\ln f(x;\theta)$ is twice continuously differentiable with respect to θ for μ-almost all x. It is also proved in [5] that the maximum likelihood estimator is asymptotically Bayes and asymptotically minimax. More properties of the maximum likelihood estimator can be found in [13].

5.3. Bayes and minimax methods

The Bayes approach. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathcal{B})$ and whose distribution belongs to a family $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta)$ where $\theta = (\theta_1, \dots, \theta_k)$ is an unknown parameter and $\Theta \subset \mathbf{R}^k$ is a Borel set. For the sake of simplicity we assume that Θ is an interval for $k = 1$ and Θ is a k-dimensional interval (parallelepiped) in $\mathbf{R}^k$ for $k > 1$. Let $r(y, \theta)$ be a nonnegative loss function, $\theta \in \mathbf{R}^k$, $y \in \mathbf{R}^k$, and let $\mathbf{Q}$ be a probability (a priori) measure on $(\mathbf{R}^k, \mathcal{B}^k)$ such that $\mathbf{Q}(\mathbf{R}^k \setminus \Theta) = 0$. For any estimator $T = T(\xi)$ of the parameter θ we introduce the risk function

$$R(T;\theta) = \mathsf{E}_\theta r(T(\xi), \theta), \qquad \theta \in \Theta. \tag{5.3.1}$$

Following ideas from Section 3.1 one can study estimators of a general function $g(\theta)$ of the parameter θ. Since a general function $g(\theta)$ can be studied similarly to the particular function $g(\theta) = \theta$, we restrict ourselves to the latter case.

Consider the risk of the estimator $T = T(\xi)$ defined as

$$R(T) = \int R(T;\theta)\, \mathbf{Q}(d\theta). \tag{5.3.2}$$

An estimator $\widetilde{\theta} = \widetilde{\theta}(\xi)$ is called a *Bayes estimator* of the parameter θ (with respect to the loss function $r(y, \theta)$ and the a priori measure $\mathbf{Q}$) if

$$R(\widetilde{\theta}) = \inf_T R(T) \tag{5.3.3}$$

where the infimum is taken over all estimators T of the parameter θ.

A posteriori Bayes estimators are also considered in the literature. Moreover the same name "Bayes estimators" is used for them. Usually this does not cause any misunderstanding, since the classes of a priori Bayes estimators and a posteriori Bayes estimators coincide in most cases.

When following the Bayes approach it is natural to treat the parameter θ as a random vector with the distribution $\mathbf{Q}$ and the measure P_y as the conditional distribution of the observation ξ given $\theta = y$, that is, $\mathsf{P}_y(A) = \mathsf{P}\{\xi \in A/\theta = y\}$, $A \in \mathcal{B}$. In this case the a posteriori measure $\mathbf{Q}_x(B) = \mathsf{P}\{\theta \in B/\xi = x\}$, $B \in \mathcal{B}^k$, is well defined. We define *a posteriori risk* $R(T/x)$ of the estimator $T = T(\xi)$ by putting

$$R(T/x) = \mathsf{E}\{r(T(\xi), \theta)/\xi = x\} = \int r(T(x), y)\, \mathbf{Q}_x(dy). \tag{5.3.4}$$

An estimator $\widetilde{\theta} = \widetilde{\theta}(x)$ is called an *a posteriori Bayes estimator* of the parameter θ (with respect to the loss function $r(y, t)$ and the a priori measure $\mathbf{Q}$) if the a posteriori risk (5.3.4) attains its minimum at this estimator; more precisely, if

$$R(\widetilde{\theta}/x) = \inf_T R(T/x) \qquad (\mu\text{-a.s.}). \tag{5.3.5}$$

There are sufficient conditions posed on the a priori measure $\mathbf{Q}$ and loss function $r(y,t)$ under which the infimum in (5.3.5) is attained and the corresponding a posteriori Bayes estimator is unique (see [**36**]). Under these conditions, the a posteriori Bayes estimator coincides with the Bayes estimator defined by (5.3.3).

Equalities (5.3.3) and (5.3.5) can be used to construct estimators of parameters. It is natural to refer to the method of estimation based on (5.3.3) and (5.3.5) as the *Bayes method*. An advantage of this approach is that the corresponding estimators are optimal in the sense that they minimize the risk of the estimators. Moreover one can freely choose the a priori measure $\mathbf{Q}$ and the loss function $r(y,t)$ to reflect the features of the case under consideration. A disadvantage of this approach is that the corresponding Bayes estimators are not easy to evaluate.

Below we give some examples of the evaluation of Bayes and a posteriori Bayes estimators. These examples also show some problems when following the Bayes approach (more examples of Bayes estimators can be found in Examples 3.1.5 and 3.1.6).

EXAMPLE 5.3.1. Let an observation be a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ from the normal $\mathcal{N}(\theta, 1)$ distribution where θ is an unknown parameter such that

$$-\infty < \theta < \infty.$$

Then $T_n = \sum_{i=1}^n \xi_i$ is a complete sufficient statistic for the parameter θ in this case (see Section 4.2). Let the a priori distribution $\mathbf{Q}$ of the parameter θ be normal $\mathcal{N}(0, \tau^2)$. Our aim is to get the a posteriori Bayes estimator $\widetilde{\theta}_n$ of the parameter θ with respect to the loss function

$$r(y,t) = \begin{cases} 0, & \text{if } |y-t| < \delta, \\ 1, & \text{if } |y-t| \geq \delta \end{cases}$$

where $\delta > 0$ is a certain fixed number. Since there exists a complete sufficient estimator T_n, a posteriori distribution depends on $x = (x_1, \dots, x_n)$ through $T_n(x)$ (see Remark 4.1.1 and Corollary 4.1.3). Then an a posteriori Bayes estimator should also be a function of the sufficient statistic T_n, that is, $\widetilde{\theta}_n = d(T_n)$.

Given the statistic T_n, an a posteriori distribution of the parameter θ is normal

$$\mathcal{N}\left(\frac{T_n}{n+\tau^{-2}}, \left(n + \frac{1}{\tau^2}\right)^{-1}\right).$$

This result can easily be derived from properties of the normal distribution. Thus a posteriori risk of the estimator $\widetilde{\theta}_n = d(T_n)$ is

$$\begin{aligned} R(\widetilde{\theta}_n/x) &= 1 - \mathsf{P}\{|d(T_n) - \theta| < \delta / T_n(x)\} \\ &= 1 - \Phi\left(\frac{d(T_n(x)) + \delta - T_n(x)/(n+\tau^{-2})}{(n+\tau^{-2})^{-1/2}}\right) \\ &\quad + \Phi\left(\frac{d(T_n(x)) - \delta - T_n(x)/(n+\tau^{-2})}{(n+\tau^{-2})^{-1/2}}\right). \end{aligned}$$

To minimize $R(\widetilde{\theta}_n/x)$ one should choose $d(T_n)$ to maximize

$$\Phi\left(\frac{d(T_n(x)) + \delta - T_n(x)/(n+\tau^{-2})}{(n+\tau^{-2})^{-1/2}}\right) - \Phi\left(\frac{d(T_n(x)) - \delta - T_n(x)/(n+\tau^{-2})}{(n+\tau^{-2})^{-1/2}}\right).$$

Consider the function $f(x) = \Phi(x + \varepsilon - \eta) - \Phi(x - \varepsilon - \eta)$. Differentiating it with respect to x we obtain a sufficient condition for x_0 to be a point of maximum of $f(x)$, namely

$$\phi(x_0 + \varepsilon - \eta) = \phi(x_0 - \varepsilon - \eta) \tag{5.3.6}$$

where $\phi(x)$ is the density of the normal $\mathcal{N}(0,1)$ distribution. Since $\phi(\varepsilon) = \phi(-\varepsilon)$, we get $x_0 = \eta$. It is easy to check that $x_0 = \eta$ is a unique solution of equation (5.3.6). The second derivative of $f(x)$ at the point $x_0 = \eta$ is equal to $-2\varepsilon\phi(\varepsilon)$. Thus x_0 is the point of maximum of the function $f(x)$. Putting

$$x = \frac{d(T_n)}{(n+\tau^{-2})^{-1/2}} \quad \text{and} \quad \eta = \frac{T_n}{(n+\tau^{-2})^{1/2}}$$

we prove that a unique a posteriori Bayes estimator of the parameter θ is given by

$$\widetilde{\theta}_n = \widetilde{\theta}_n(x) = \frac{T_n(x)}{n+\tau^{-2}}, \qquad x = (x_1, \ldots, x_n).$$

Note that

$$\lim_{\tau\to\infty} \widetilde{\theta}_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The following example shows that there are a posteriori Bayes estimators for which the risk is infinite and thus the evaluation of a Bayes estimator does not make any sense from a practical point of view.

EXAMPLE 5.3.2. Let an observation ξ be a random variable with the distribution belonging to the family of uniform distributions on $(0, |\theta|^{-1})$ where θ is a real number such that $1 \le |\theta| < \infty$, that is, $f(x;\theta) = |\theta|$ for $0 \le x \le |\theta|^{-1}$.

Let an a priori distribution $\mathbf{Q}$ be absolutely continuous with respect to the Lebesgue measure with the density

$$q(t) = \begin{cases} \frac{1}{2}|t|^{-2}, & \text{if } 1 \le |t| < \infty, \\ 0, & \text{if } |t| < 1. \end{cases}$$

An a posteriori distribution of the parameter θ given $\xi = x$ possesses the density

$$q(t/x) = \begin{cases} \frac{|t|^{-1}}{2\ln x^{-1}}, & \text{if } x \le |t|^{-1} \le 1, \\ 0, & \text{if } |t|^{-1} \notin [x, 1]. \end{cases}$$

Let the loss function $r(y,t)$ be quadratic, that is, $r(y,t) = |y-t|^2$. Then the a posteriori Bayes estimator $\widetilde{\theta}(x)$ minimizes the a posteriori risk

$$R(T/x) = \int_{\{x\le|t|^{-1}\le 1\}} (T-t)^2 \frac{|t|^{-1}}{2\ln x^{-1}}\, dt.$$

The only function $T(x)$ minimizing $R(T/x)$ is the a posteriori mean of the parameter θ given $\xi = x$, that is, $T(x) = \mathsf{E}\{\theta/\xi = x\}$. Since the a posteriori density $q(t/x)$ is symmetric with respect to $x = 0$, we get that $\widehat{\theta}(x) = 0$ (μ-a.s.) is an a posteriori Bayes estimator. The a posteriori risk of this estimator is given by

$$R(\widehat{\theta}/x) = \int_{\{x\le|t|^{-1}\le 1\}} \frac{|t|}{2\ln x^{-1}}\, dt = \frac{1-x^2}{2x^2\ln x^{-1}} < \infty \qquad (\mu\text{-a.s.}).$$

However the risk of the a posteriori Bayes estimator $\widehat{\theta}(x) = 0$ is

$$R(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2 = 2\int_1^\infty t^2 \frac{1}{2}|t|^{-2}\,dt = \infty.$$

Note that the a posteriori Bayes estimator $\widehat{\theta}(x) = 0$ (μ-a.s.) is of no interest at all, since it does not depend on observations. Moreover the estimator $\widehat{\theta}(x)$ assumes values outside the set of parameters $\Theta = (-\infty, 1] \cup (1, \infty)$.

Below are some concluding remarks.

REMARK 5.3.1. The problem of how to construct Bayes estimators in the case of the loss function $r(y,t) = |y-t|^m$, $m = 1, 2, \ldots$, is quite well studied in the literature. It is known in the case $m = 1$ that the median of an a posteriori distribution is the Bayes estimator (and the a posteriori Bayes estimator, as well) of a parameter θ (see Remark 3.2.4 and [**9**], pp. 178–179). If $m = 2$, then the expectation of an a posteriori distribution is the Bayes estimator (and the a posteriori Bayes estimator, as well) of a parameter θ (see Theorem 2.2.1 concerning the general regression). In the case of a general loss function $r(y,t) = w(|y-t|)$ De Groot and Rao (1963) obtained necessary and sufficient conditions that an estimator is a posteriori Bayes (see [**36**], Theorem 6.2.2).

REMARK 5.3.2. If the a priori distribution $\mathbf{Q}$ is such that a posteriori risk with respect to the loss function $r(y,t) = \lambda(t)(y-t)^2$ is finite (μ-a.s.) for all estimators $\widehat{\theta}(x)$, then an a posteriori Bayes estimator is given by

$$\widetilde{\theta}(x) = \frac{\mathsf{E}(\theta\lambda(\theta)/\xi)}{\mathsf{E}(\lambda(\theta)/\xi)} \qquad (\mu\text{-a.s.})$$

(see [**36**]). Here $0 < \lambda(t) < \infty$ for all $t \in \Theta$. If $R(T/x) < \infty$ (μ-a.s.) only for $T = T_0$, then $T_0(x)$ is an a posteriori Bayes estimator. In general, an a posteriori Bayes estimator is unique (see Theorem 6.2.1 in [**36**]).

REMARK 5.3.3. In Section 3.1 we defined a generalized Bayes estimator as an estimator minimizing the risk if the a priori measure $\mathbf{Q}$ is not a probability measure. Generalized Bayes estimators are sometimes defined as limits of Bayes estimators constructed with respect to a priori probability measures $\mathbf{Q}_m$ as $m \to \infty$ (see, for example, Theorem 3.2.6). The estimator

$$\theta^*(x) = \frac{\int t f(x;t)\,\mathbf{Q}(dt)}{\int f(x;t)\,\mathbf{Q}(dt)}$$

is also called a generalized Bayes estimator where $\mathbf{Q}$ is some σ-finite measure (see [**36**]). More results on Bayes estimators can be found in [**36**].

Asymptotic properties of Bayes estimators. Let an observation be a sample $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ from a distribution belonging to a family $\{\mathsf{P}_\theta, \theta \in \Theta\}$ dominated by some σ-finite measure μ. Let $f(x;\theta)$ be the density of the measure P_θ with respect to the measure μ. Let $\widetilde{\theta}_n$ be the Bayes estimator of the parameter θ with respect to the quadratic loss function $r(y,t) = (y-t)^2$ and the a priori measure $\mathbf{Q}$ possessing the density $q(t)$ with respect to the Lebesgue measure.

The following result describes the asymptotic behavior of $\widetilde{\theta}_n$ as $n \to \infty$ in the case of a one-dimensional parameter θ.

THEOREM 5.3.1. *Let assumptions* 1)–4) *of Theorem* 5.2.5 *hold. Let additionally*

5) *the function* $q(t)$ *is continuous in a neighborhood of the point* $t = t_0$, $q(t_0) \neq 0$, *and* $\sup_t q(t)(1 + |t|^{p_0})^{-1} < \infty$ *for some* $p_0 \geq 0$.

Then the Bayes estimator $\widetilde{\theta}_n$ *given* $\theta = t_0$ *is asymptotically* $\mathcal{N}(t_0, (nI(t_0))^{-1})$ *normal as* $n \to \infty$.

The proof of Theorem 5.3.1 can be found in [**13**].

The assumptions of Theorem 5.3.1 can be weakened (see [**13**]). Moreover this result can be generalized to the case of a multidimensional parameter θ.

The minimax approach. The minimax method allows one to construct estimators, called *minimax*, that minimize the maximum of the risk function. Some necessary and sufficient conditions for estimators to be minimax are given in Section 3.1 (see Theorems 3.1.3 and 3.1.4). In many cases the minimax estimator is a Bayes estimator with respect to the less favorable a priori distribution. In those cases the construction of a minimax estimator is reduced to the construction of an appropriate Bayes estimator.

Some examples of minimax estimators are given in Examples 3.1.5 and 3.1.6. Below we give another example related to the loss function introduced in Example 5.3.1.

EXAMPLE 5.3.3. We learned in Example 5.3.1 that

$$\widetilde{\theta}_\tau(x) = \overline{x}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1}$$

where $\overline{x}_n = n^{-1} \sum_{i=1}^n x_i$, $x = (x_1, \ldots, x_n)$, is a Bayes estimator with respect to the a priori normal $\mathcal{N}(0, \tau^2)$ distribution and the loss function $r(y,t)$ that equals 0 for $|y-t| < \delta$ and 1 for $|y-t| \geq \delta$, $\delta > 0$. Let us show that $\overline{x}_n$ is a minimax estimator. The risk function of the estimator $\overline{x}_n$ is equal to

$$R(\overline{x}_n, t) = \mathsf{P}_t\{|\overline{\xi}_n - t| > \delta\} = 2(1 - \Phi(\delta\sqrt{n})) = \rho^*$$

for all $t \in (-\infty, \infty)$ where $\overline{\xi}_n = n^{-1} \sum_{i=1}^n \xi_i$. It is easy to check that the risk of the estimator $\widetilde{\theta}_\tau$ is given by

$$R(\widetilde{\theta}_\tau, t) = 2 - \left\{ \Phi\left(\frac{\delta + t(1+n\tau^2)^{-1}}{(1+(n\tau^2)^{-1})^{-1}} \sqrt{n} \right) + \Phi\left(\frac{\delta - t(1+n\tau^2)^{-1}}{(1+(n\tau^2)^{-1})^{-1}} \sqrt{n} \right) \right\}. \tag{5.3.7}$$

Putting $\tau = \tau_1, \tau_2, \ldots$ we get a sequence of a priori $\mathcal{N}(0, \tau_m^2)$ distributions, $m = 1, 2, \ldots$. We denote the risk with respect to the a priori $\mathcal{N}(0, \tau^2)$ distribution by

$$R_\tau(\widetilde{\theta}_\tau) = \mathsf{E}_\tau R(\widetilde{\theta}_\tau; \theta) \tag{5.3.8}$$

where E_τ is the expectation with respect to the $\mathcal{N}(0, \tau^2)$ distribution. For the right-hand side of (5.3.8) we apply the estimate $R(\widetilde{\theta}_\tau; t) \leq 2$. Then the Lebesgue dominated convergence theorem and (5.3.7) imply

$$\lim_{\tau \to \infty} R_\tau(\widetilde{\theta}_\tau) = 2(1 - \Phi(\delta\sqrt{n})) = \rho^*.$$

Now we apply Theorem 3.1.4 for the sequence of a priori $\mathcal{N}(0,\tau_m)$ distributions and Bayes estimators $\widetilde{\theta}_{\tau_m}$ and obtain that $\overline{x}_n$ is the minimax estimator of the parameter θ.

REMARK 5.3.4. Further results on minimax estimators can be found in [**36**].

5.4. Confidence intervals and regions

The notion of a confidence interval. In the preceding sections we considered the problem of constructing *point* estimators of an unknown parameter or a function of a parameter. Every point estimator is a statistic assuming values in the region of values of the parameter (or of the function of the parameter). It is a useful method in practice to construct an interval or a region from the observation or from the sample. The idea behind this method is that the interval or region mentioned above should contain the parameter with a probability close to 1.

Let ξ be an observation that is a random element assuming values in a measurable space $(X,\mathcal{B})$. Let its distribution belong to a family $(\mathsf{P}_\theta,\theta\in\Theta)$ where $\theta=(\theta_1,\dots,\theta_k)$ is an unknown parameter such that $\theta\in\Theta\subset\mathbf{R}^k$, $k\geq 1$. First we consider the case of a one-dimensional parameter θ, that is, we consider the case $k=1$.

Let $T_1=T_1(\xi)$ and $T_2=T_2(\xi)$ be two statistics such that $T_1<T_2$ and let

$$\mathsf{P}_\theta\{T_1(\xi)<\theta<T_2(\xi)\}\geq\gamma \quad \text{for all } \theta\in\Theta$$

for a given $\gamma\in(0,1)$. In this case the interval (T_1,T_2) is called a *γ-confidence interval* or a *confidence interval of level γ* for the parameter θ. The number γ is called a *confidence probability* or a *confidence level*, while T_1 and T_2 are called the *lower* and *upper confidence limits*, respectively.

Constructing a confidence interval by a given statistic. Let $\widehat{\theta}$ be an estimator of a parameter θ. It is natural to seek a confidence interval of a level γ in the form of $(\widehat{\theta}-\Delta(\gamma,\xi),\widehat{\theta}+\Delta(\gamma,\xi))$. However the random variables $\Delta(\gamma,\xi)$ depend, generally speaking, on the unknown parameter θ, since these random variables are found from the equation

$$\mathsf{P}_\theta\{\widehat{\theta}-\Delta(\gamma,\xi)<\theta<\widehat{\theta}+\Delta(\gamma,\xi)\}\geq\gamma \quad \text{for all } \theta\in\Theta.$$

Along with the estimator $\widehat{\theta}$ one can use any other statistic T when constructing a confidence interval. Let $G_\theta(y)=\mathsf{P}_\theta\{T(\xi)<y\}$ be the distribution function of the statistic T. Assume that $G_\theta(y)$ depends on the parameter θ monotonically. More precisely let

$$G_{\theta_1}(y)\geq G_{\theta_2}(y) \quad \text{for all } y \text{ and } \theta_1<\theta_2. \tag{5.4.1}$$

If additionally the function $G_\theta(y)$ is continuous with respect to θ, then the equation

$$G_\theta(y)=\gamma \tag{5.4.2}$$

has a solution with respect to θ for every $\gamma\in(0,1)$. We denote this solution by $b(y,\gamma)$.

THEOREM 5.4.1. *Let $\gamma = \gamma_1 + \gamma_2$. If the distribution function $G_\theta(y)$ of the statistic T is continuous with respect to θ and satisfies condition* (5.4.1), *then the statistics*

$$T_1 = b(T, 1-\gamma_2), \qquad T_2 = b(T, \gamma_1)$$

are lower and upper limits of a confidence interval of level $1-\gamma$.

PROOF. The random variable $G_\theta(T(\xi))$ has the uniform distribution with respect to P_θ on the interval [0,1]. Thus

$$\mathsf{P}_\theta\{\gamma_1 < G_\theta(T(\xi)) < 1-\gamma_2\} = 1-\gamma$$

or, equivalently,

$$\mathsf{P}_\theta\{b(T(\xi), 1-\gamma_2) < \theta < b(T(\xi), \gamma_1)\} = 1-\gamma. \qquad \square$$

The inversion procedure of the function $G_\theta(T)$ used in the proof of Theorem 5.4.1 can be done in two steps. First one inverts the function $G_\theta(y)$ with respect to y, that is, one finds the quantiles of $G_\theta^{-1}(\gamma)$ that are solutions of equation (5.4.2). Then one solves the following equations with respect to θ:

$$G_\theta^{-1}(\gamma_1) = T, \qquad G_\theta^{-1}(1-\gamma_2) = T.$$

Solutions of these equations exist, since the function $G_\theta^{-1}(\gamma)$ is monotone and continuous with respect to θ for all $\gamma \in (0,1)$.

If the function $G_\theta(y)$ is not continuous with respect to θ, then Theorem 5.4.1 still holds and the above procedure still works. The only difference is that an equality in the definition of quantiles is substituted by the inequality

$$\mathbf{G}_\theta(G_\theta^{-1}(\gamma_1), G_\theta^{-1}(1-\gamma_2)) \geq 1-\gamma$$

where $\mathbf{G}_\theta$ is the measure on $(-\infty, \infty)$ generated by the distribution of $G_\theta(y)$. Since we assumed continuity in Theorem 5.4.1, the quantiles were evaluated from the corresponding equalities in the proof above.

The problem of finding the most precise estimator also exists in the case of the interval setting. We will solve this problem when studying hypotheses testing.

Constructing confidence intervals for the Bayes approach. Let a parameter θ be random with a priori distribution $\mathbf{Q}$ possessing the density $q(y)$ with respect to some σ-finite measure λ. Assume that a family $(\mathsf{P}_\theta, \theta \in \Theta)$ of distributions of the observation ξ is dominated by some σ-finite measure μ. Thus $f(x;\theta)$ is the density of the measure P_θ with respect to the measure μ. In this case there exists an a posteriori distribution of the parameter θ given $\xi = x$. Its density with respect to λ is given by

$$q(y/x) = \frac{f(x;y)q(y)}{\int f(x;t)q(t)\,\lambda(dt)}.$$

As lower and upper limits for a confidence interval of level $1-\gamma$ one can take statistics $T_1(x)$ and $T_2(x)$ such that

$$\int_{T_1(x)}^{T_2(x)} q(t/x)\,\lambda(dt) = 1-\gamma$$

or

$$\int_{T_1(x)}^{T_2(x)} q(t/x)\,\lambda(dt) \geq 1-\gamma$$

depending on the continuity or discontinuity of the function $\int_{-\infty}^{t} q(u/x)\,\lambda(du)$ with respect to t. In other words, as statistics T_1 and T_2 one should take a γ_1-quantile and a $(1-\gamma_2)$-quantile, respectively, of an a posteriori distribution for all γ_1 and γ_2 such that $\gamma_1+\gamma_2=\gamma$.

In contrast to the non-Bayesian approach, in the relation $T_1<\theta<T_2$ both T_1 and T_2, as well as the parameter θ, are now random.

Asymptotically confidence intervals. Assume that an observation

$$\xi^{(n)}=(\xi_1,\dots,\xi_n)$$

is a sample from a distribution belonging to a family $(\mathsf{P}_\theta,\ \theta\in\Theta)$. Let

$$\underline{T}_n=T_1(\gamma,\xi^{(n)}) \quad \text{and} \quad \overline{T}_n=T_2(\gamma,\xi^{(n)})$$

be two statistics such that

$$\liminf_{n\to\infty} \mathsf{P}_\theta\{\underline{T}_n<\theta<\overline{T}_n\}\geq\gamma \quad \text{for all } \theta\in\Theta. \tag{5.4.3}$$

Then the interval $(\underline{T}_n,\overline{T}_n)$ is called a *confidence interval of level* γ. In fact, now one should speak of a sequence of intervals $(\underline{T}_n,\overline{T}_n)$, $n=1,2,\dots$.

In the preceding sections we considered point estimators that in the majority of cases are asymptotically normal. Below we construct asymptotic confidence intervals from the point estimators.

Let $\widehat{\theta}_n$ be an asymptotically $\mathcal{N}(\theta,\sigma^2(\theta)/n)$ normal estimator where $\sigma(\theta)$ is a continuous function. Since $\widehat{\theta}_n\to\theta$ in P_θ-probability as $n\to\infty$, we also have $\sigma(\widehat{\theta}_n)\to\sigma(\theta)$ in P_θ-probability as $n\to\infty$. This implies that the sequence

$$\frac{(\widehat{\theta}_n-\theta)\sqrt{n}}{\sigma(\widehat{\theta}_n)}, \qquad n=1,2,\dots,$$

is asymptotically $\mathcal{N}(0,1)$ normal. Denote by z_δ a solution of the equation

$$\Phi(z)=1-\delta$$

with respect to z, that is, z_δ is a $(1-\delta)$-quantile of the distribution $\mathcal{N}(0,1)$. Here the symbol $\Phi(z)$ stands for the distribution function of the law $\mathcal{N}(0,1)$. If η is a random variable distributed according to the law $\mathcal{N}(0,1)$, then $\mathsf{P}\{|\eta|<z_\delta\}=1-2\delta$. Let $\beta=z_{\gamma/2}$ for a fixed number $\gamma>0$. Hence

$$\lim_{n\to\infty}\mathsf{P}_\theta\left\{\left|\frac{(\widehat{\theta}_n-\theta)\sqrt{n}}{\sigma(\widehat{\theta}_n)}\right|<\beta\right\}=1-\gamma$$

or, in other words,

$$\lim_{n\to\infty}\mathsf{P}_\theta\left\{\widehat{\theta}_n-\frac{\beta\sigma(\widehat{\theta}_n)}{\sqrt{n}}<\theta<\widehat{\theta}_n+\frac{\beta\sigma(\widehat{\theta}_n)}{\sqrt{n}}\right\}=1-\gamma.$$

Therefore relation (5.4.3) holds for the random variables

$$\underline{T}_n = \widehat{\theta}_n - \frac{\beta\sigma(\widehat{\theta}_n)}{\sqrt{n}}, \qquad \overline{T}_n = \widehat{\theta}_n + \frac{\beta\sigma(\widehat{\theta}_n)}{\sqrt{n}}. \tag{5.4.4}$$

Equalities (5.4.4) define lower and upper limits of an asymptotic confidence interval of level $1-\gamma$.

EXAMPLE 5.4.1. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a sample from the Gamma distribution, so that the density is $f(x;\theta) = \theta e^{-\theta x}$, $x > 0$, where $\theta \in \Theta = (0, \infty)$ is an unknown parameter. The random variable $T_n = \sum_{i=1}^n \xi_i$ is a complete sufficient statistic and moreover $\mathsf{E}_\theta T_n^{-1} = \theta/(n-1)$. Thus $\widehat{\theta}_n = (n-1)T_n^{-1}$ is an unbiased optimal estimator of the parameter θ. Further $\mathsf{D}_\theta\widehat{\theta}_n = \theta^2/(n-2)$ and therefore $\sigma^2(\theta) = \theta^2$. Therefore the limits defined by (5.4.4) become of the form

$$\underline{T}_n = \widehat{\theta}_n\left(1 - \frac{\beta}{\sqrt{n}}\right), \qquad \overline{T}_n = \widehat{\theta}_n\left(1 + \frac{\beta}{\sqrt{n}}\right).$$

The asymptotic confidence level of the interval $(\underline{T}_n, \overline{T}_n)$ is $1-\gamma$. One can find a precise confidence level of the interval $(\underline{T}_n, \overline{T}_n)$ for a fixed n by evaluating the probability

$$\mathsf{P}_\theta\left\{1 - \frac{\beta}{\sqrt{n}} < \frac{\theta T_n}{n-1} < 1 + \frac{\beta}{\sqrt{n}}\right\}$$

which is possible by taking into account that θT_n has the Gamma distribution with the density $x^{n-1}\, e^{-x}/(n-1)!$, $x > 0$.

The multidimensional case. If a parameter $\theta \in \Theta \subset \mathbf{R}^k$ is multidimensional, that is, $k > 1$, then we consider a confidence region instead of a confidence interval.

A random subset $\Theta^* = \Theta^*(\gamma, \xi)$ of the region of parameters Θ is called a *confidence region of level* γ if

$$\mathsf{P}_\theta\{\theta \in \Theta^*\} \geq \gamma \quad \text{for all } \theta \in \Theta.$$

In other words, a confidence region Θ^* of level γ contains the unknown parameter θ with a probability greater than or equal to γ.

If an observation is a sample $\xi^{(n)}$, then a random set $\Theta_n^* = \Theta^*(\gamma, \xi^{(n)}) \subset \Theta$ such that

$$\liminf_{n\to\infty} \mathsf{P}_\theta\{\theta \in \Theta_n^*\} \geq \gamma$$

is called an *asymptotic confidence region of level* γ.

The procedure for constructing confidence regions is the same as in the one-dimensional case.

Confidence intervals for normal distributions. We use the exact distributions of linear and quadratic forms of Gaussian random variables (see Section 1.4) to construct exact confidence intervals for parameters of the normal distribution.

EXAMPLE 5.4.2. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ be a sample from the normal $\mathcal{N}(\theta, \sigma^2)$ distribution where $\theta \in \Theta = (-\infty, \infty)$ is an unknown parameter, while the variance σ^2 is known. Thus $\mathsf{E}_\theta\xi_1 = \theta$ and $\mathsf{D}_\theta\xi_1 = \sigma^2$ for all $\theta \in \Theta$. Our goal is to

construct a confidence interval of level γ for the parameter θ. We use the estimator

$$\widehat{\theta}_n = n^{-1} \sum_{i=1}^{n} \xi_i$$

of the parameter θ to find a confidence interval $(\widehat{\theta}_n - \Delta_{n,\gamma}, \widehat{\theta}_n + \Delta_{n,\gamma})$ where $\Delta_{n,\gamma}$ is a solution of the equation

$$\mathsf{P}_\theta \left\{ \widehat{\theta}_n - \Delta_{n,\gamma} < \theta < \widehat{\theta}_n + \Delta_{n,\gamma} \right\} = \gamma.$$

Since the distribution of the estimator $\widehat{\theta}_n$ is $\mathcal{N}(\theta, \sigma^2/n)$, the latter equation is equivalent to

$$\mathsf{P}_\theta\{|\widehat{\theta}_n - \theta| < \Delta_{n,\gamma}\} = \mathsf{P}_\theta \left\{ \left| \frac{(\widehat{\theta}_n - \theta)\sqrt{n}}{\sigma} \right| < \frac{\sqrt{n}\Delta_{n,\gamma}}{\sigma} \right\} = 2\Phi\left(\frac{\sqrt{n}\Delta_{n,\gamma}}{\sigma} \right) - 1 = \gamma.$$

This implies that $\Delta_{n,\gamma} = \sigma t_{(1-\gamma)/2}/\sqrt{n}$ where t_p is a p-quantile of the law $\mathcal{N}(0,1)$, that is, $\Phi(t_p) = p$. Thus, a confidence interval is of the form

$$\left(\widehat{\theta}_n - \frac{\sigma}{\sqrt{n}} t_{(1-\gamma)/2}, \widehat{\theta}_n + \frac{\sigma}{\sqrt{n}} t_{(1-\gamma)/2} \right).$$

The procedure described in Theorem 5.4.1 can also be used to construct a confidence interval.

EXAMPLE 5.4.3. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the normal $\mathcal{N}(\alpha, \theta)$ distribution where α is known and $\theta \in (0, \infty)$ is an unknown parameter. Now we use the statistic $T_n = \sum_{i=1}^{n} (\xi_i - \alpha)^2$ which, as we know, is a sufficient statistic for the parameter θ. It is obvious that the distribution of the random variable T_n/θ is $\chi^2(n)$. Thus there are two numbers $\underline{r}_{n,\gamma}$ and $\overline{r}_{n,\gamma}$ such that

$$\mathsf{P}_\theta \left\{ \underline{r}_{n,\gamma} < T_n/\theta < \overline{r}_{n,\gamma} \right\} = \gamma.$$

Note that a solution of the latter equation is not unique. A confidence interval of level γ for the variance θ can be taken as follows:

$$(T_n/\overline{r}_{n,\gamma}, T_n/\underline{r}_{n,\gamma}).$$

Note that numbers $\underline{r}_{n,\gamma}$ and $\overline{r}_{n,\gamma}$ possessing this property are not unique.

EXAMPLE 5.4.4. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample from the normal $\mathcal{N}(\theta_1, \theta_2)$ distribution where $\theta = (\theta_1, \theta_2)$ is an unknown parameter such that $\theta_1 \in (-\infty, \infty)$ and $\theta_2 \in (0, \infty)$, that is, we assume that both the expectation θ_1 and variance θ_2 are unknown. First we construct a confidence interval of level $\gamma \in (0,1)$ for the expectation θ_1. We use the random variable $T_n = (a_1 - \theta_1) m_2^{-1/2}$ where

$$a_1 = \frac{1}{n} \sum_{i=1}^{n} \xi_i \quad \text{and} \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - a_1)^2$$

are the sampling mean and sampling variance, respectively. According to Theorem 1.4.2 the random variable T_n has the Student distribution with $n-1$ degrees of freedom. Let $c_{n,\gamma}$ be a constant such that

$$\mathsf{P}_\theta\{|T_n| < c_{n,\gamma}\} = \gamma.$$

Such a constant $c_{n,\gamma}$ exists and is unique. Its approximate value can be found from tables of the Student distribution. Therefore a confidence interval of level γ is of the form $(a_1 - c_{n,\gamma}\sqrt{m_2}, a_1 + c_{n,\gamma}\sqrt{m_2})$.

Now we construct a confidence interval of level $\gamma \in (0,1)$ for the variance θ_2. We use the random variable $S_n = nm_2/\theta_2$. According to Theorem 1.4.2 the distribution of the random variable S_n is $\chi^2(n-1)$. Thus one can find two numbers $\underline{r}_{n-1,\gamma}$ and $\overline{r}_{n-1,\gamma}$ (see Example 5.4.3) such that

$$\mathsf{P}_\theta\left\{\underline{r}_{n-1,\gamma} < nm_2/\theta_2 < \overline{r}_{n-1,\gamma}\right\} = \gamma.$$

This implies that a confidence interval of level γ for the variance θ_2 can be taken as $(nm_2/\overline{r}_{n-1,\gamma}, nm_2/\underline{r}_{n-1,\gamma})$. It is obvious that the numbers $\underline{r}_{n-1,\gamma}$ and $\overline{r}_{n-1,\gamma}$ satisfying the above equality are not unique and thus confidence intervals also are not unique.

References to Part 1

1. T. W. Anderson, *An introduction to multivariate statistical analysis*, Wiley, New York–London–Sydney, 1958.
2. ———, *The statistical analysis of time series*, Wiley, New York–London–Sydney, 1971.
3. J.-R. Barra, *Notions fondamentales de statistique mathématique*, Dunod, Paris, 1971; English transl., *Mathematical basis of statistics*, Academic Press, New York–London, 1981.
4. P. J. Bickel and K. A. Doksum, *Mathematical statistics*, Holden Day, San Francisco, CA, 1977.
5. A. A. Borovkov, *Mathematical statistics. Estimation of parameters. Testing of hypotheses*, "Nauka", Moscow, 1984; English transl., *Mathematical statistics*, Gordon & Breach, Amsterdam, 1998.
6. ———, *Mathematical statistics. Supplementary chapters*, "Nauka", Moscow, 1984; English transl., Gordon & Breach, Amsterdam, 1998.
7. D. R. Cox and P. A. M. Lewis, *The statistical analysis of series of events*, Methuen, London, 1966.
8. D. R. Cox and D. Hinkley, *Theoretical statistics*, Chapman & Hall, London, 1974.
9. H. Cramér, *Mathematical methods of statistics*, Princeton Univ. Press, Princeton, NJ, 1999.
10. H. A. David, *Order statistics*, Wiley, New York–London–Sydney, 1970.
11. I. I. Gikhman, A. V. Skorokhod, and M. I. Yadrenko, *Theory of probability and mathematical statistics*, "Vyshcha shkola", Kiev, 1988. (Russian)
12. B. V. Gnedenko, *The theory of probability*, "Fizmatgiz", Moscow, 1961; English transl., Chelsea, New York, 1967.
13. I. A. Ibragimov and R. Z. Khas'minskiĭ, *Statistical estimation. Asymptotic theory*, "Nauka", Moscow, 1979; English transl., Springer-Verlag, New York–Berlin, 1981.
14. G. I. Ivchenko and Yu. I. Medvedev, *Mathematical statistics*, "Vysshaya shkola", Moscow, 1984. (Russian)
15. A. M. Kagan, Yu. V. Kagan, and C. R. Rao, *Characterization problems in mathematical statistics*, "Nauka", Moscow, 1972; English transl., Wiley, New York–London–Sydney, 1973.
16. M. G. Kendall and A. Stuart, *The advanced theory of statistics. Distribution theory*, 3rd edition, vol. 1, Griffin, London, 1961.
17. ———, *The advanced theory of statistics. Inference and relationship*, vol. 2, Griffin, London, 1961.
18. ———, *The advanced theory of statistics. Design and analysis and time-series*, 2nd edition, vol. 3, Griffin, London, 1968.
19. M. V. Kozlov and A. V. Prokhorov, *Introduction to mathematical statistics*, Moscow University Press, Moscow, 1987. (Russian)
20. S. Kullback, *Information theory and statistics*, Dover, New York, 1968.
21. E. L. Lehmann, *Theory of point estimation*, Wiley, New York, 1983.
22. Yu. N. Lin'kov, *Asymptotic statistical methods for stochastic processes*, "Naukova dumka", Kiev, 1993; English transl., Amer. Math. Soc., Providence, RI, 2001.
23. M. Loéve, *Probability theory*, 4th edition, Springer-Verlag, New York–Heidelberg–Berlin, 1977.
24. A. K. Mitropol'skiĭ, *Technique of statistical computations*, "Nauka", Moscow, 1961. (Russian)
25. E. L. G. Pitman, *Some basic theory for statistical inference*, Chapman & Hall, London, 1979.
26. C. R. Rao, *Statistical inference and its applications*, Wiley, New York–London–Sydney, 1965.
27. Yu. A. Rozanov, *Theory of probability, random processes, and mathematical statistics*, "Nauka", Moscow, 1985. (Russian)
28. H. Scheffeé, *The analysis of variance*, Wiley and Chapman & Hall, New York–London, 1959.

29. B. A. Sevast'yanov, *A course in probability theory and mathematical statistics*, "Nauka", Moscow, 1982. (Russian)
30. A. N. Shiryaev, *Probability*, "Nauka", Moscow, 1989; English transl., Springer-Verlag, New York, 1996.
31. L. Shmetterer, *Introduction to mathematical statistics*, Springer-Verlag, New York, 1973.
32. N. V. Smirnov, *Theory of probability and mathematical statistics. Selected works*, "Nauka", Moscow, 1970. (Russian)
33. N. V. Tutubalin, *Theory of probability*, Moscow University Press, Moscow, 1972. (Russian)
34. B. L. van der Waerden, *Mathematische Statistik*, Springer-Verlag, Berlin–Göttingen–Heidelberg, 1957. (German)
35. S. S. Wilks, *Mathematical statistics*, Wiley, New York–London–Sydney, 1963.
36. S. Zacks, *The theory of statistical inference*, Wiley, New York–London–Sydney, 1971.

Part 2

Preface to Part 2

Part 1 of this book dealt with the estimation of unknown parameters, while Part 2 is devoted to testing statistical hypotheses.

The theory of hypotheses testing appears, in more or less detail, in practically any textbook or monograph on mathematical statistics. We mention here the books by Lehmann [**34**], and Hájek and Šidák [**22**] that are entirely devoted to statistical tests, as well as the book by Borovkov and Mogul'skiĭ [**10**] that is devoted to asymptotic problems in testing statistical hypotheses.

Part 2 begins with an exposition of a general theory of testing (Chapter 1), that is, of problems related to testing statistical hypotheses in the scheme of general statistical experiments according to Ibragimov and Khas'minskiĭ [**25**], Barra [**2**], and Soler [**49**]. First, in Section 1.1, we deal with testing two hypotheses, we study the structure of the set formed by type I and type II error probabilities, and we introduce Neyman-Pearson tests, Bayes tests, and minimax tests. In Section 1.2, the theory of testing a finite number of simple hypotheses is presented and the most powerful tests, Bayes tests, and minimax tests are introduced. Section 1.3 deals with testing composite hypotheses and discusses different approaches to the definition of optimal tests. A relationship between tests and confidence intervals is investigated.

Chapter 2 deals with problems for asymptotically distinguishable families of simple statistical hypotheses in the scheme of general statistical experiments following the books [**47**] and [**37**]. A complete group of types of families of statistical hypotheses that can be asymptotically distinguished is introduced and characterization theorems are given, which enables one to determine the type to which a family of hypotheses belongs (Section 2.2). Complete asymptotic testing under the strong law of large numbers (Section 2.3) or under weak convergence (Section 2.4) of the logarithm of the likelihood ratio are presented. Section 2.5 deals with testing contiguous families of hypotheses.

Chapter 3 is devoted to goodness-of-fit tests for independent observations. The Kolmogorov test (Section 3.1), the Pearson test (Section 3.2), and the Smirnov test (Section 3.3) are considered in detail. Section 3.4 focuses on some other well-known goodness-of-fit tests.

Chapter 4 presents elements of sequential analysis applied to the problem of testing statistical hypotheses. Section 4.1 deals with the Bayes theory of sequential testing of, generally speaking, composite hypotheses. Sections 4.2 and 4.3 are devoted to the Wald sequential test for testing two simple hypotheses. Section 4.2 presents the basic properties of the Wald test and Section 4.3 establishes that the Wald test is optimal.

The list of references at the end of the book contains only those references that are directly related to the topics we treat in the book and is by no means a complete list of references on testing statistical hypotheses.

In Part 2 we follow the same system of notational conventions as in Part 1. We also enumerate theorems, lemmas, formulas, etc., in the same way as we did in Part 1.

CHAPTER 1

General Theory of Hypotheses Testing

1.1. Testing two simple hypotheses

Statistical hypotheses and tests. Type I and type II error probabilities of a test. Let ξ be a random element assuming values in a measurable space $(X, \mathscr{B})$ and let $\mathscr{P} = (\mathsf{P}, \widetilde{\mathsf{P}})$ be a pair of probability measures defined on $(X, \mathscr{B})$. Assume that the distribution of the random element ξ is generated by one of the measures of the family $\mathscr{P}$. The random element ξ is called an *observation*. The problem is to make a decision about the distribution of the random element ξ by the observation $\xi = x$.

Any conjecture about the distribution of an observation ξ is called a *statistical hypothesis* or, simply, a *hypothesis*. If a statistical hypothesis uniquely determines the distribution of an observation, then it is called a *simple hypothesis*. Otherwise it is called a *composite hypothesis*.

Let H and $\widetilde{H}$ be two statistical hypotheses that the distribution of an observation ξ corresponds to the measure P and $\widetilde{\mathsf{P}}$, respectively. It is clear that the hypotheses H and $\widetilde{H}$ are simple. Therefore the problem is to decide by using the observation $\xi = x$ which of the hypotheses H or $\widetilde{H}$ is true. In other words, this is a problem of *distinguishing two simple hypotheses* H and $\widetilde{H}$ by an observation $\xi = x$.

Any measurable mapping $\delta\colon (X, \mathscr{B}) \to ([0,1], \mathscr{B}([0,1]))$ where $\mathscr{B}(A)$ is the Borel σ-algebra of the set A is called a *statistical test* for distinguishing hypotheses H and $\widetilde{H}$. We treat $\delta(x)$ as the probability of accepting the hypothesis $\widetilde{H}$ given $\xi = x$, while $1 - \delta(x)$ is the probability of accepting the hypothesis H given $\xi = x$. The mapping δ is sometimes called a *decision rule* or a *decision function*. If the function $\delta(x)$ assumes only two values 0 and 1, then it is called a *nonrandomized test*. Otherwise δ is called a *randomized test*.

If a test δ is nonrandomized, then $X = X_0 \cup X_1$ where $X_i = \{x\colon \delta(x) = i\}$, $i = 0, 1$, and $X_0 \cap X_1 = \varnothing$. In this case the hypothesis H is accepted for $x \in X_0$, while the hypothesis $\widetilde{H}$ is accepted for $x \in X_1$. Thus every nonrandomized test is of the form $\delta(x) = I_{X_1}(x)$, $x \in X$, where $I_A(x)$ is the indicator of the set A, that is, $I_A(x) = 1$ for $x \in A$ and $I_A(x) = 0$ for $x \in A^c = X \setminus A$.

Throughout this chapter we write $\delta = \delta(\xi)$. To measure the quality of a test δ we consider the two numbers

$$\alpha(\delta) = \mathsf{E}\delta \quad \text{and} \quad \beta(\delta) = \widetilde{\mathsf{E}}(1 - \delta) \tag{1.1.1}$$

where E and $\widetilde{\mathsf{E}}$ are expectations with respect to the measures P and $\widetilde{\mathsf{P}}$, respectively. If $f(x)$ is a measurable function, then we write $\mathsf{P}\{f(\xi) \in A\}$ or $\mathsf{P}\{f \in A\}$ and $\widetilde{\mathsf{P}}\{f(\xi) \in A\}$ or $\widetilde{\mathsf{P}}\{f \in A\}$ instead of $\mathsf{P}\{x\colon f(x) \in A\}$ and $\widetilde{\mathsf{P}}\{x\colon f(x) \in A\}$, respectively. The number $\alpha(\delta)$ is called the *type* I *error probability* or δ*-level* of the test δ.

Similarly, the number $\beta(\delta)$ is called the *type* II *error probability* of the test δ. The number $1-\beta(\delta)$ is called the *power* of the test δ.

It is natural to say that a test δ_1 is better than a test δ_2 if $\alpha(\delta_1) \le \alpha(\delta_2)$, $\beta(\delta_1) \le \beta(\delta_2)$, and at least one of these two inequalities is strict. However it is not always possible to compare tests δ_1 and δ_2 in the specified way. In what follows we consider the set $\mathfrak{N}$ of points $(\alpha(\delta), \beta(\delta))$ corresponding to all possible tests δ. It is clear that $\mathfrak{N} \subset [0,1] \times [0,1]$. The definition of the set $\mathfrak{N}$ implies that $(\alpha, \beta) \in \mathfrak{N}$ if and only if there is a test δ such that $\alpha(\delta) = \alpha$ and $\beta(\delta) = \beta$.

Properties of the set $\mathfrak{N}$. First we consider some properties of the set $\mathfrak{N}$ that hold for each pair of measures $(\mathsf{P}, \widetilde{\mathsf{P}})$.

LEMMA 1.1.1. *The set $\mathfrak{N}$ is convex.*

PROOF. Let δ_1 and δ_2 be two arbitrary tests. Then $(\alpha(\delta_1), \beta(\delta_1)) \in \mathfrak{N}$ and $(\alpha(\delta_2), \beta(\delta_2)) \in \mathfrak{N}$. Let $0 \le \lambda \le 1$ and

$$\alpha = \lambda\alpha(\delta_1) + (1-\lambda)\alpha(\delta_2), \qquad \beta = \lambda\beta(\delta_1) + (1-\lambda)\beta(\delta_2). \tag{1.1.2}$$

We prove that $(\alpha, \beta) \in \mathfrak{N}$ for all $\lambda \in [0,1]$. We get from (1.1.1) and (1.1.2) that

$$\alpha = \mathsf{E}[\lambda\delta_1 + (1-\lambda)\delta_2], \qquad \beta = \widetilde{\mathsf{E}}[1 - (\lambda\delta_1 + (1-\lambda)\delta_2)].$$

This implies that $\alpha = \alpha(\bar{\delta})$ and $\beta = \beta(\bar{\delta})$ where $\bar{\delta} = \lambda\delta_1 + (1-\lambda)\delta_2$. It is obvious that $\bar{\delta}$ is a test for any $\lambda \in [0,1]$. Thus $(\alpha, \beta) \in \mathfrak{N}$ for any $\lambda \in [0,1]$ and therefore $\mathfrak{N}$ is a convex set. □

LEMMA 1.1.2. *The points $(0,1)$ and $(1,0)$ belong to the set $\mathfrak{N}$.*

PROOF. Let $\delta_0(x) = 0$ for all $x \in X$. Then $\alpha(\delta_0) = 0$ and $\beta(\delta_0) = 1$, whence $(0,1) \in \mathfrak{N}$. Further let $\delta_1(x) = 1$ for all $x \in X$. Thus $\alpha(\delta_1) = 1$ and $\beta(\delta_1) = 0$, whence $(1,0) \in \mathfrak{N}$. □

LEMMA 1.1.3. *The set $\mathfrak{N}$ is symmetric about the point $(1/2, 1/2)$.*

PROOF. It is sufficient to prove that if $(\alpha, \beta) \in \mathfrak{N}$, then $(1-\alpha, 1-\beta) \in \mathfrak{N}$. Let δ be a test such that $\alpha(\delta) = \alpha$ and $\beta(\delta) = \beta$. It follows from (1.1.1) that

$$1 - \alpha(\delta) = \mathsf{E}(1-\delta), \qquad 1 - \beta(\delta) = \widetilde{\mathsf{E}}\delta.$$

Putting $\tilde{\delta} = 1 - \delta$, we get

$$\alpha(\tilde{\delta}) = 1 - \alpha(\delta) = 1 - \alpha, \qquad \beta(\tilde{\delta}) = 1 - \beta(\delta) = 1 - \beta,$$

that is, $(1-\alpha, 1-\beta) \in \mathfrak{N}$. □

REMARK 1.1.1. Lemmas 1.1.1 and 1.1.2 imply that the diagonal of the square $[0,1] \times [0,1]$ joining its corners $(0,1)$ and $(1,0)$ belongs to the set $\mathfrak{N}$. Lemma 1.1.3 implies that the subset of $\mathfrak{N}$ above this diagonal coincides with the image under the central symmetry about the point $(1/2, 1/2)$ of the subset of $\mathfrak{N}$ below the diagonal. Therefore one can derive all the properties for the set $\mathfrak{N}$ from their counterparts for one of the two parts of $\mathfrak{N}$ specified above.

Now we consider other properties of the set $\mathfrak{N}$ that depend on the measures P and $\widetilde{\mathsf{P}}$. We need some definitions and results from measure theory that can be found, for example, in [**19, 23, 31, 32**].

A measure $\widetilde{\mathsf{P}}$ is called *absolutely continuous with respect to a measure* P if

$$\widetilde{\mathsf{P}}(A) = 0$$

for all $A \in \mathscr{B}$ such that $\mathsf{P}(A) = 0$. We write in this case $\widetilde{\mathsf{P}} \ll \mathsf{P}$. If $\widetilde{\mathsf{P}} \ll \mathsf{P}$ and $\mathsf{P} \ll \widetilde{\mathsf{P}}$, then the measures P and $\widetilde{\mathsf{P}}$ are called *equivalent*. The equivalence of measures $\widetilde{\mathsf{P}}$ and P is denoted by $\widetilde{\mathsf{P}} \sim \mathsf{P}$.

LEMMA 1.1.4. *If* $\mathsf{P} \sim \widetilde{\mathsf{P}}$, *then for all tests* δ *we have*

$$\alpha(\delta) = 0 \iff \beta(\delta) = 1, \tag{1.1.3}$$

$$\alpha(\delta) = 1 \iff \beta(\delta) = 0. \tag{1.1.4}$$

PROOF. Since $0 \le \delta(x) \le 1$, we get

$$\begin{aligned}\alpha(\delta) = 0 &\iff \mathsf{E}\delta = 0 \iff \mathsf{P}\{x\colon \delta(x) \ne 0\} = 0 \\ &\iff \widetilde{\mathsf{P}}\{x\colon \delta(x) \ne 0\} = 0 \iff \widetilde{\mathsf{E}}(1-\delta) = 1\end{aligned}$$

and (1.1.3) is proved. Relation (1.1.4) is proved similarly. □

We say that a measure P *is not absolutely continuous with respect to a measure* $\widetilde{\mathsf{P}}$ (denoted by $\mathsf{P} \not\ll \widetilde{\mathsf{P}}$) if there is a set $C \in \mathscr{B}$ such that $\widetilde{\mathsf{P}}(C) = 0$ and $\mathsf{P}(C) > 0$.

LEMMA 1.1.5. *If* $\widetilde{\mathsf{P}} \ll \mathsf{P}$ *and* $\mathsf{P} \not\ll \widetilde{\mathsf{P}}$, *then for all tests* δ *we have*

$$\alpha(\delta) = 0 \Rightarrow \beta(\delta) = 1, \tag{1.1.5}$$

$$\alpha(\delta) = 1 \Rightarrow \beta(\delta) = 0. \tag{1.1.6}$$

Moreover there are tests δ' *and* δ'' *such that* $\beta(\delta') = 0$, $\alpha(\delta') < 1$, $\beta(\delta'') = 1$, *and* $\alpha(\delta'') > 0$.

PROOF. Since $0 \le \delta(x) \le 1$, we have

$$\alpha(\delta) = 0 \Rightarrow \mathsf{P}\{x\colon \delta(x) \ne 0\} = 0 \Rightarrow \widetilde{\mathsf{P}}\{x\colon \delta(x) \ne 0\} = 0 \Rightarrow \widetilde{\mathsf{E}}(1-\delta) = 1,$$

whence (1.1.5) follows. Relation (1.1.6) is proved similarly.

Further let $C \in \mathscr{B}$ be such that $\widetilde{\mathsf{P}}(C) = 0$ and $\mathsf{P}(C) > 0$. Putting

$$\delta'(x) = I_{X \setminus C}(x)$$

we get

$$\begin{aligned}\alpha(\delta') &= \mathsf{E}\delta' = \mathsf{P}(X \setminus C) = 1 - \mathsf{P}(C) < 1, \\ \beta(\delta') &= \widetilde{\mathsf{E}}(1-\delta') = \widetilde{\mathsf{P}}(C) = 0.\end{aligned}$$

If $\delta''(x) = I_C(x)$, then we get in a similar way that $\alpha(\delta'') = \mathsf{P}(C) > 0$ and

$$\beta(\delta'') = \widetilde{\mathsf{P}}(X \setminus C) = 1.$$

□

LEMMA 1.1.6. *If* $\mathsf{P} \ll \widetilde{\mathsf{P}}$ *and* $\widetilde{\mathsf{P}} \not\ll \mathsf{P}$, *then for all tests* δ *we have*

$$\beta(\delta) = 0 \Rightarrow \alpha(\delta) = 1,$$
$$\beta(\delta) = 1 \Rightarrow \alpha(\delta) = 0.$$

Moreover there are tests δ' *and* δ'' *such that* $\alpha(\delta') = 0$, $\beta(\delta') < 1$, $\alpha(\delta'') = 1$, *and* $\beta(\delta'') > 0$.

The proof is similar to that of Lemma 1.1.5 and thus is omitted.

LEMMA 1.1.7. *If* $\widetilde{\mathsf{P}} \not\ll \mathsf{P}$ *and* $\mathsf{P} \not\ll \widetilde{\mathsf{P}}$, *then there are tests* δ_1, δ_2, δ_3, *and* δ_4 *such that*

$$\alpha(\delta_1) = 0, \quad \beta(\delta_1) < 1, \qquad \alpha(\delta_2) < 1, \quad \beta(\delta_2) = 0,$$
$$\alpha(\delta_3) = 1, \quad \beta(\delta_3) > 0, \qquad \alpha(\delta_4) > 0, \quad \beta(\delta_4) = 1.$$

The proof is similar to that of Lemma 1.1.5 and thus is omitted.

REMARK 1.1.2. To prove Lemmas 1.1.5 and 1.1.6 one can put $\delta'' = 1 - \delta'$ and apply Lemma 1.1.3. Similarly, to prove Lemma 1.1.7 one can put $\delta_3 = 1 - \delta_1$ and $\delta_4 = 1 - \delta_2$ and apply Lemma 1.1.3.

Measures P and $\widetilde{\mathsf{P}}$ are called *singular* (denoted by $\mathsf{P} \perp \widetilde{\mathsf{P}}$) if there exists $C \in \mathscr{B}$ such that $\mathsf{P}(C) = 0$ and $\widetilde{\mathsf{P}}(X \setminus C) = 0$.

LEMMA 1.1.8. *If* $\mathsf{P} \perp \widetilde{\mathsf{P}}$, *then* $(0,0) \in \mathfrak{N}$.

PROOF. Let $C \in \mathscr{B}$ be such that $\mathsf{P}(C) = 0$ and $\widetilde{\mathsf{P}}(C) = 1$. Putting $\delta^0(x) = I_C(x)$ we get

$$\alpha(\delta^0) = \mathsf{E}\delta^0 = \mathsf{P}(C) = 0, \qquad \beta(\delta^0) = \widetilde{\mathsf{E}}(1 - \delta^0) = \widetilde{\mathsf{P}}(X \setminus C) = 0,$$

that is, $(0,0) \in \mathfrak{N}$. □

COROLLARY 1.1.1. *If* $\mathsf{P} \perp \widetilde{\mathsf{P}}$, *then* $\mathfrak{N} = [0,1] \times [0,1]$.

PROOF. According to Lemmas 1.1.8 and 1.1.3 we get $(1,1) \in \mathfrak{N}$. Since $(0,1) \in \mathfrak{N}$ and $(1,0) \in \mathfrak{N}$ by Lemma 1.1.2, we apply Lemma 1.1.1 and obtain

$$\mathfrak{N} = [0,1] \times [0,1].$$ □

We write $\mathsf{P} = \widetilde{\mathsf{P}}$ if $\mathsf{P}(A) = \widetilde{\mathsf{P}}(A)$ for all $A \in \mathscr{B}$. Put

$$\underline{\mathfrak{N}} = \{(\alpha, \beta) : \beta = 1 - \alpha \text{ for all } \alpha \in [0,1]\}, \qquad \overline{\mathfrak{N}} = [0,1] \times [0,1]. \tag{1.1.7}$$

It is clear that $\underline{\mathfrak{N}}$ is the diagonal of the square $[0,1] \times [0,1]$ joining its corners $(0,1)$ and $(1,0)$.

COROLLARY 1.1.2. *The following hold*:

$$\mathsf{P} = \widetilde{\mathsf{P}} \iff \mathfrak{N} = \underline{\mathfrak{N}}, \tag{1.1.8}$$

$$\mathsf{P} \perp \widetilde{\mathsf{P}} \iff \mathfrak{N} = \overline{\mathfrak{N}}. \tag{1.1.9}$$

PROOF. Let $\mathsf{P} = \widetilde{\mathsf{P}}$. Then for all tests δ

$$\beta(\delta) = \widetilde{\mathsf{E}}(1-\delta) = \mathsf{E}(1-\delta) = 1 - \alpha(\delta),$$

that is, $\mathfrak{N} = \underline{\mathfrak{N}}$. Thus the implication $\Rightarrow$ in (1.1.8) is proved.

Let $\mathfrak{N} = \underline{\mathfrak{N}}$. Then $\beta(\delta) = 1 - \alpha(\delta)$ for all tests δ. Consider $\delta(x) = I_A(x)$ for $A \in \mathscr{B}$. Then

$$\beta(\delta) = \widetilde{\mathsf{E}}(1-\delta) = 1 - \widetilde{\mathsf{P}}(A), \qquad \alpha(\delta) = \mathsf{E}\delta = \mathsf{P}(A). \tag{1.1.10}$$

Since $\beta(\delta) = 1 - \alpha(\delta)$ for all tests δ, we obtain from (1.1.10) that $\widetilde{\mathsf{P}}(A) = \mathsf{P}(A)$ for all $A \in \mathscr{B}$. Thus $\mathsf{P} = \widetilde{\mathsf{P}}$. Therefore the implication $\Leftarrow$ in (1.1.8) is also proved.

According to Corollary 1.1.1, relation (1.1.9) follows from the implication $\Leftarrow$ in (1.1.9).

Let $\mathfrak{N} = \overline{\mathfrak{N}}$. Then $(0,0) \in \mathfrak{N}$ and therefore there is a test δ^0 such that $\alpha(\delta^0) = 0$ and $\beta(\delta^0) = 0$. Thus $\mathsf{P}\{x\colon \delta^0(x) \neq 0\} = 0$ if $\alpha(\delta^0) = 0$. Similarly, the equality $\beta(\delta^0) = 0$ implies that $\widetilde{\mathsf{P}}\{x\colon \delta^0(x) \neq 1\} = 0$, that is, $\widetilde{\mathsf{P}}\{x\colon \delta^0(x) = 1\} = 1$. Putting $C = \{x\colon \delta^0(x) \neq 0\}$, we obtain from $\{x\colon \delta^0(x) = 1\} \subset C$ that $\mathsf{P}(C) = 0$ and $\widetilde{\mathsf{P}}(C) = 1$, that is, $\mathsf{P} \perp \widetilde{\mathsf{P}}$. Thus the implication $\Leftarrow$ in (1.1.9) is also proved. □

Likelihood ratio and Lebesgue decomposition. Let $\mathbf{Q}$ be some σ-finite measure on $(X, \mathscr{B})$ dominating the family $\mathscr{P} = (\mathsf{P}, \widetilde{\mathsf{P}})$. This means that $\mathsf{P} \ll \mathbf{Q}$ and $\widetilde{\mathsf{P}} \ll \mathbf{Q}$. It is obvious that such a measure exists. In particular, as the measure $\mathbf{Q}$ one can take $\mathbf{Q} = (\mathsf{P} + \widetilde{\mathsf{P}})/2$. Let $\mathfrak{z}(x) = d\mathsf{P}/d\mathbf{Q}(x)$ and $\tilde{\mathfrak{z}}(x) = d\widetilde{\mathsf{P}}/d\mathbf{Q}(x)$ be the Radon–Nikodym derivatives (densities) of the measures P and $\widetilde{\mathsf{P}}$ with respect to the measure $\mathbf{Q}$, respectively. Note that $0 \leq \mathfrak{z}(x) < \infty$ and $0 \leq \tilde{\mathfrak{z}}(x) < \infty$ almost everywhere with respect to the measure $\mathbf{Q}$. Moreover, $\mathsf{P}\{x\colon \mathfrak{z}(x) = 0\} = 0$ and $\widetilde{\mathsf{P}}\{x\colon \tilde{\mathfrak{z}}(x) = 0\} = 0$. We define the *likelihood ratios* as follows:

$$z(x) = \tilde{\mathfrak{z}}(x)/\mathfrak{z}(x), \qquad \tilde{z}(x) = \mathfrak{z}(x)/\tilde{\mathfrak{z}}(x). \tag{1.1.11}$$

If we agree that $0 = 0/0$, then $z(x)$ and $\tilde{z}(x)$ in (1.1.11) are well defined. Note that

$$\mathsf{P}\{x\colon \tilde{\mathfrak{z}}(x) = 0,\ \mathfrak{z}(x) = 0\} = 0,$$
$$\widetilde{\mathsf{P}}\{x\colon \tilde{\mathfrak{z}}(x) = 0,\ \mathfrak{z}(x) = 0\} = 0.$$

The following result provides the Lebesgue decomposition of one of the measures P or $\widetilde{\mathsf{P}}$ with respect to the other one.

LEMMA 1.1.9. *For all sets* $A \in \mathscr{B}$

$$\widetilde{\mathsf{P}}(A) = \int_A z(x)\,\mathsf{P}(dx) + \widetilde{\mathsf{P}}(A \cap \{x\colon \mathfrak{z}(x) = 0\}), \tag{1.1.12}$$

$$\mathsf{P}(A) = \int_A \tilde{z}(x)\,\widetilde{\mathsf{P}}(dx) + \mathsf{P}(A \cap \{x\colon \tilde{\mathfrak{z}}(x) = 0\}) \tag{1.1.13}$$

where $z(x)$ *and* $\tilde{z}(x)$ *are defined in* (1.1.11).

PROOF. Note that for all $A \in \mathscr{B}$ we have

$$\widetilde{\mathsf{P}}(A) = \widetilde{\mathsf{P}}(A \cap \{\mathfrak{z} > 0\}) + \widetilde{\mathsf{P}}(A \cap \{\mathfrak{z} = 0\}). \tag{1.1.14}$$

Since $\mathsf{P}(\mathfrak{z} = 0) = 0$, we get

$$\begin{aligned} \widetilde{\mathsf{P}}(A \cap \{\mathfrak{z} > 0\}) &= \mathsf{E}_Q \tilde{\mathfrak{z}} I(A \cap \{\mathfrak{z} > 0\}) \\ &= \mathsf{E}_Q \tilde{\mathfrak{z}} \frac{\mathfrak{z}}{\mathfrak{z}} I(A \cap \{\mathfrak{z} > 0\}) = \mathsf{E} \frac{\tilde{\mathfrak{z}}}{\mathfrak{z}} I(A \cap \{\mathfrak{z} > 0\}) \\ &= \mathsf{E}\, zI(A) = \int_A z(x)\, \mathsf{P}(dx) \end{aligned} \tag{1.1.15}$$

where E_Q is the integral with respect to the measure $\mathbf{Q}$ and $I(A)$ is the indicator of the set A (in other words, of the event $\{\xi \in A\}$). Thus $I(A) = 1$ if $x \in A$ or $\xi \in A$, while $I(A) = 0$ if $x \notin A$ or $\xi \notin A$. Decomposition (1.1.12) follows from (1.1.14) and (1.1.15). Decomposition (1.1.13) is proved similarly. □

Put

$$\overline{\alpha} = \mathsf{P}(\tilde{\mathfrak{z}} > 0), \qquad \overline{\beta} = \widetilde{\mathsf{P}}(\mathfrak{z} > 0). \tag{1.1.16}$$

The Lebesgue decompositions (1.1.12) and (1.1.13) imply that

$$\widetilde{\mathsf{P}} \ll \mathsf{P} \Longleftrightarrow \overline{\beta} = 1, \tag{1.1.17}$$

$$\mathsf{P} \ll \widetilde{\mathsf{P}} \Longleftrightarrow \overline{\alpha} = 1. \tag{1.1.18}$$

It follows from (1.1.11) and (1.1.16) that

$$\overline{\alpha} = \mathsf{P}(z > 0) = \mathsf{P}(\tilde{z} < \infty), \tag{1.1.19}$$

$$\overline{\beta} = \widetilde{\mathsf{P}}(\tilde{z} > 0) = \widetilde{\mathsf{P}}(z < \infty). \tag{1.1.20}$$

The Lebesgue decompositions yield the following result.

LEMMA 1.1.10. *If η is an arbitrary nonnegative and measurable function defined on $(X, \mathscr{B})$, then*

$$\widetilde{\mathsf{E}}\eta = \mathsf{E}\eta z + \widetilde{\mathsf{E}}\eta I(\mathfrak{z} = 0), \tag{1.1.21}$$

$$\mathsf{E}\eta = \widetilde{\mathsf{E}}\eta\tilde{z} + \mathsf{E}\eta I(\tilde{\mathfrak{z}} = 0). \tag{1.1.22}$$

The following result contains more properties of the set $\mathfrak{N}$.

LEMMA 1.1.11. *For all tests δ*

$$\beta(\delta) = 0 \Rightarrow \alpha(\delta) \geq \overline{\alpha}. \tag{1.1.23}$$

Moreover there exists a test δ' such that $\beta(\delta') = 0$ and $\alpha(\delta') = \overline{\alpha}$. Further, for all tests δ

$$\alpha(\delta) = 0 \Rightarrow \beta(\delta) \geq \overline{\beta} \tag{1.1.24}$$

and there exists a test δ'' such that $\alpha(\delta'') = 0$ and $\beta(\delta'') = \overline{\beta}$.

PROOF. If $\beta(\delta) = 0$, then $\widetilde{\mathsf{P}}(\delta = 1) = 1$. Taking into account equalities (1.1.22) and $\widetilde{\mathsf{P}}(\tilde{\jmath} = 0) = 0$, we obtain

$$\alpha(\delta) = \mathsf{E}\delta = \widetilde{\mathsf{E}}\delta\tilde{z} + \mathsf{E}\delta I(\tilde{\jmath} = 0) = \widetilde{\mathsf{E}}\tilde{z} + \mathsf{E}\delta I(\tilde{\jmath} = 0)$$
$$\geq \widetilde{\mathsf{E}}\tilde{z} I(\tilde{\jmath} > 0) = \mathsf{E}_Q \tilde{z}\tilde{\jmath}\, I(\tilde{\jmath} > 0) = \mathsf{E} I(\tilde{\jmath} > 0) = \overline{\alpha}.$$

Thus the implication (1.1.23) is proved. Putting $\delta' = I(\tilde{\jmath} > 0)$, we get

$$\alpha(\delta') = \mathsf{P}(\tilde{\jmath} > 0) = \overline{\alpha}, \qquad \beta(\delta') = \widetilde{\mathsf{P}}(\tilde{\jmath} = 0) = 0.$$

The proof of the implication (1.1.24) is similar and follows from (1.1.21). Putting $\delta'' = I(\jmath = 0)$, we obtain $\alpha(\delta'') = 0$ and $\beta(\delta'') = \overline{\beta}$. □

We derive the following useful relations from Lemma 1.1.11 and Corollary 1.1.2:

(1.1.25) $$\widetilde{\mathsf{P}} \perp \mathsf{P} \Longleftrightarrow \overline{\beta} = 0,$$

(1.1.26) $$\widetilde{\mathsf{P}} \perp \mathsf{P} \Longleftrightarrow \overline{\alpha} = 0.$$

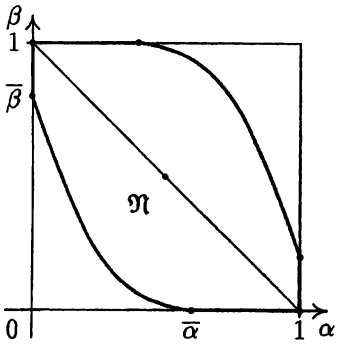

FIGURE 1.1.1

REMARK 1.1.3. The set $\mathfrak{N}$ is shown in Figure 1.1.1. The points $(\overline{\alpha}, 0)$ and $(0, \overline{\beta})$ depicted in Figure 1.1.1 are defined by equalities (1.1.16), (1.1.19), and (1.1.20). Using (1.1.17) and (1.1.25) we get

(1.1.27) $$0 < \overline{\beta} < 1 \Longleftrightarrow \widetilde{\mathsf{P}} \not\ll \mathsf{P},\ \widetilde{\mathsf{P}} \not\perp \mathsf{P}.$$

Similarly, it follows from (1.1.18) and (1.1.26) that

(1.1.28) $$0 < \overline{\alpha} < 1 \Longleftrightarrow \mathsf{P} \not\ll \widetilde{\mathsf{P}},\ \widetilde{\mathsf{P}} \not\perp \mathsf{P}.$$

The properties of the set $\mathfrak{N}$ proved above together with equivalences (1.1.27) and (1.1.28) completely describe the set $\mathfrak{N}$. Note that the tests corresponding to the points $(\overline{\alpha}, 0)$ and $(\overline{\beta}, 0)$ are defined in Lemma 1.1.11. Thus $(\overline{\alpha}, 0) \in \mathfrak{N}$ and $(0, \overline{\beta}) \in \mathfrak{N}$. Since $\mathfrak{N}$ is convex, the two segments of the straight lines joining the points $(\overline{\alpha}, 0)$ and $(1, 0)$ and $(0, \overline{\beta})$ and $(0, 1)$, respectively, belong to the set $\mathfrak{N}$. Since $\mathfrak{N}$ is symmetric about the point $(1/2, 1/2)$, two segments of the straight lines joining the points $(0, 1)$ and $(1 - \overline{\alpha}, 0)$ and $(1, 1 - \overline{\beta})$ and $(1, 0)$, respectively, also belong to the set $\mathfrak{N}$.

The most powerful, Bayes, and minimax tests. Consider the following two classes of tests:

$$\mathcal{K}_\alpha = \{\delta: \alpha(\delta) = \alpha\}, \qquad \mathcal{K}^\alpha = \{\delta: \alpha(\delta) \le \alpha\} \tag{1.1.29}$$

where α is some number of the interval $[0,1]$. It is clear that $\mathcal{K}_\alpha \subset \mathcal{K}^\alpha$ for all $\alpha \in [0,1]$. Put

$$\mathfrak{N}_\alpha = \{(\alpha(\delta), \beta(\delta)): \delta \in \mathcal{K}_\alpha\}, \qquad \mathfrak{N}^\alpha = \{(\alpha(\delta), \beta(\delta)): \delta \in \mathcal{K}^\alpha\}. \tag{1.1.30}$$

Then $\mathfrak{N}_\alpha \subset \mathfrak{N}^\alpha$.

A test $\delta^{*,\alpha}$ is called the *most powerful test of level* α if

$$\beta(\delta^{*,\alpha}) = \min\{\beta(\delta): \delta \in \mathcal{K}_\alpha\} \tag{1.1.31}$$

(in what follows we show that the minimum in (1.1.31) is attained for all $\alpha \in [0,1]$, indeed). It is clear that the test $\delta^{*,\alpha}$ has the maximal power $1-\beta(\delta)$ among tests δ of the class $\mathcal{K}_\alpha$. Also, it follows from (1.1.29), (1.1.30), and the definition of the set $\mathfrak{N}$ that the test $\delta^{*,\alpha}$ has the maximal power among the tests of the class $\mathcal{K}^\alpha$. This is an explanation of why we say that $\delta^{*,\alpha}$ is the *most powerful test in the class* $\mathcal{K}^\alpha$.

The intersection of the straight line $\alpha = \alpha_0$ and the lower bound of the set $\mathfrak{N}$ in Figure 1.1.2 determines the point A whose coordinates are $(\alpha_0, \beta(\delta^{*,\alpha_0}))$ and which corresponds to the most powerful test δ^{*,α_0} in the class $\mathcal{K}^{\alpha_0}$.

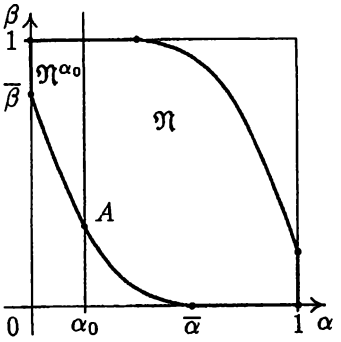

FIGURE 1.1.2

There is a different approach to compare tests. This is the *Bayes approach* based on the assumption that the tests H and $\widetilde{H}$ are random events and their probabilities $\pi = \mathsf{P}(H)$ and $\widetilde{\pi} = \mathsf{P}(\widetilde{H}) = 1 - \pi$ are known. The probabilities π and $\widetilde{\pi}$ are called *a priori error probabilities of tests* H *and* $\widetilde{H}$. The quality of a test δ is defined as the average of the error probabilities:

$$e_\pi(\delta) = \pi\alpha(\delta) + (1-\pi)\beta(\delta). \tag{1.1.32}$$

A test δ_π is called a *Bayes test with respect to the a priori distribution* $(\pi, 1-\pi)$ if

$$e_\pi(\delta_\pi) = \min_\delta e_\pi(\delta) \tag{1.1.33}$$

where the minimum is considered with respect to all tests δ. In Figure 1.1.3 the straight line $\pi\alpha + (1-\pi)\beta = c$ and set $\mathfrak{N}$ have only one common point B and

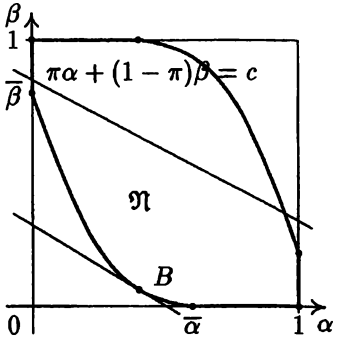

FIGURE 1.1.3

it corresponds to the Bayes test δ_π where c is some constant. It is clear that the Bayes test δ_π is the most powerful one in the class $\mathcal{K}^{\alpha_0}$ for $\alpha_0 = \alpha(\delta_\pi)$.

The following approach to compare tests is called *minimax* and is based on the maximal probability of errors of a test.

A test δ^* is called *minimax* if

$$\alpha(\delta^*) \vee \beta(\delta^*) = \min_{\delta}(\alpha(\delta) \vee \beta(\delta)) \tag{1.1.34}$$

where $\alpha \vee \beta$ is the maximum of two numbers α and β, while the minimum in (1.1.34) is considered with respect to all tests δ.

We will discuss Bayes and minimax tests in more detail when considering the problem of testing a finite number of simple tests.

The maximum likelihood test and the Neyman–Pearson fundamental lemma. Consider the test

$$\delta^{c,\varepsilon} = I(z > c) + \varepsilon I(z = c) \tag{1.1.35}$$

where z is the likelihood ratio defined by (1.1.11), and $c \in [0, \infty]$ and $\varepsilon \in [0, 1]$ are the parameters of the test. The test $\delta^{c,\varepsilon}$ defined by (1.1.35) is called the *maximum likelihood test*.

The following result is known as the *Neyman–Pearson fundamental lemma*. It shows that every maximum likelihood test is the most powerful one and, moreover, every most powerful test coincides (in a certain sense) with some maximum likelihood test.

THEOREM 1.1.1.

1) *For every $\alpha \in (0, \overline{\alpha})$ there exists a maximum likelihood test of level α.*
2) *The maximum likelihood test is the most powerful test of level α.*
3) *If $\delta^{*,\alpha}$ is the most powerful test of level $\alpha \in (0, \overline{\alpha})$, then there exists a constant c such that*

$$\mathsf{P}(S_c) = \widetilde{\mathsf{P}}(S_c) = 0$$

where

$$S_c = \{x: \delta^{*,\alpha}(x) \neq \delta^{c,\varepsilon}(x)\} \cap \{x: z(x) \neq c\} \tag{1.1.36}$$

and ε is an arbitrary constant of the interval $[0, 1]$.

PROOF. 1) Consider the function $F(c) = \mathsf{P}(z < c)$. Obviously $F(c) = 0$ for $c \leq 0$ and $F(0+) = 1 - \mathsf{P}(z > 0) = 1 - \mathsf{P}(\zeta > 0) = 1 - \overline{\alpha}$. Moreover,

$$F(\infty) = 1 - \mathsf{P}(z = \infty) = 1 - \mathsf{P}(\zeta = 0) = 1.$$

Let $c(\alpha)$ be the minimal solution of the system of inequalities

$$F(c) \leq 1 - \alpha \leq F(c+0) = F(c) + \mathsf{P}(z = c). \tag{1.1.37}$$

Further let $\varepsilon(\alpha) \in [0, 1]$ be such that

$$1 - \alpha = F(c(\alpha)) + (1 - \varepsilon(\alpha))\mathsf{P}(z = c(\alpha)). \tag{1.1.38}$$

If $\mathsf{P}(z = c(\alpha)) = 0$, then an arbitrary number of the interval $[0, 1]$ can be taken as $\varepsilon(\alpha)$. Otherwise, if $\mathsf{P}(z = c(\alpha)) \neq 0$, then we get from (1.1.38) that

$$\varepsilon(\alpha) = \frac{F(c(\alpha) + 0) - (1 - \alpha)}{\mathsf{P}(z = c(\alpha))}. \tag{1.1.39}$$

Equality (1.1.38) implies that the level of the maximum likelihood test $\delta^{c(\alpha),\varepsilon(\alpha)}$ is α.

2) Let $0 < \alpha < \overline{\alpha}$ and let $\delta^{c,\varepsilon}$ be the maximum likelihood test of level α. It is sufficient to show that $\beta(\delta) \geq \beta(\delta^{c,\varepsilon})$ for every test δ of level α. We have

$$\beta(\delta) - \beta(\delta^{c,\varepsilon}) = \widetilde{\mathsf{E}}(\delta^{c,\varepsilon} - \delta). \tag{1.1.40}$$

Since the levels of the tests $\delta^{c,\varepsilon}$ and δ are equal to α, it follows that

$$\alpha(\delta^{c,\varepsilon}) - \alpha(\delta) = \mathsf{E}(\delta^{c,\varepsilon} - \delta) = 0 \tag{1.1.41}$$

holds. When proving the first statement of the theorem we showed that

$$F(0+) = 1 - \overline{\alpha}, \qquad F(\infty) = 1.$$

This implies that the minimal solution $c = c(\alpha)$ of the system of inequalities (1.1.37) for $\alpha \in (0, \overline{\alpha})$ is such that $0 < c < \infty$. Multiplying (1.1.41) by c and subtracting the result from (1.1.40) we get

$$\beta(\delta) - \beta(\delta^{c,\varepsilon}) = \widetilde{\mathsf{E}}(\delta^{c,\varepsilon} - \delta) - c\mathsf{E}(\delta^{c,\varepsilon} - \delta). \tag{1.1.42}$$

Applying equality (1.1.21) we derive from (1.1.42) that

$$\beta(\delta) - \beta(\delta^{c,\varepsilon}) = \mathsf{E}(\delta^{c,\varepsilon} - \delta)(z - c) + \widetilde{\mathsf{E}}(\delta^{c,\varepsilon} - \delta)I(\zeta = 0). \tag{1.1.43}$$

Since $(\zeta = 0) \subset (z = 0) \cup (z = \infty)$, we have $(z = c, \zeta = 0) = \varnothing$ and

$$\widetilde{\mathsf{P}}(z < c, \zeta = 0) = \widetilde{\mathsf{P}}(\tilde{\zeta} = 0, \zeta = 0) = 0.$$

Thus

$$\begin{aligned} &\widetilde{\mathsf{E}}(\delta^{c,\varepsilon} - \delta)I(\zeta = 0) \\ &\quad = \widetilde{\mathsf{E}}(1 - \delta)I(z > c, \zeta = 0) \\ &\qquad + \widetilde{\mathsf{E}}(\varepsilon - \delta)I(z = c, \zeta = 0) - \widetilde{\mathsf{E}}I(z < c, \zeta = 0) \\ &\quad = \widetilde{\mathsf{E}}(1 - \delta)I(z > c, \zeta = 0) \geq 0. \end{aligned} \tag{1.1.44}$$

Taking into account (1.1.44) we obtain from (1.1.43) that

$$\begin{aligned}\beta(\delta) - \beta(\delta^{c,\varepsilon}) &\geq \mathsf{E}(\delta^{c,\varepsilon} - \delta)(z - c) \\ &= \mathsf{E}(1 - \delta)(z - c)I(z > c) + \mathsf{E}(-\delta)(z - c)I(z < c) \geq 0.\end{aligned} \tag{1.1.45}$$

Therefore $\beta(\delta) \geq \beta(\delta^{c,\varepsilon})$.

3) Let $\delta^{*,\alpha}$ be the most powerful test of level $\alpha \in (0, \overline{\alpha})$ and let $\delta^{c,\varepsilon}$ be the likelihood ratio test of level α. Since $\beta(\delta^{*,\alpha}) - \beta(\delta^{c,\varepsilon}) \leq 0$ and $\delta^{*,\alpha}$ is the most powerful test, we obtain that

$$\mathsf{E}(\delta^{c,\varepsilon} - \delta^{*,\alpha})(z - c)I(S_c) + \widetilde{\mathsf{E}}(\delta^{c,\varepsilon} - \delta^{*,\alpha})I(S_c \cap (\mathfrak{z} = 0)) = 0 \tag{1.1.46}$$

in view of relations (1.1.43) and (1.1.44) for $\delta = \delta^{*,\alpha}$ where S_c is the set defined by (1.1.36). Note that $(\delta^{c,\varepsilon} - \delta^{*,\alpha})(z - c) > 0$ on the set S_c and $\delta^{c,\varepsilon} - \delta^{*,\alpha} > 0$ on the set $S_c \cap (\mathfrak{z} = 0)$. Thus relation (1.1.46) implies that $\mathsf{P}(S_c) = 0$ and

$$\widetilde{\mathsf{P}}(S_c \cap (\mathfrak{z} = 0)) = 0.$$

Lebesgue decomposition yields

$$\widetilde{\mathsf{P}}(S_c) = \int_{S_c} z d\mathsf{P} + \widetilde{\mathsf{P}}(S_c \cap (\mathfrak{z} = 0)) = 0,$$

whence the third statement of Theorem 1.1.1 follows. □

Combining Lemmas 1.1.3 and 1.1.11 with Remark 1.1.3 and Theorem 1.1.1 we obtain the following result.

LEMMA 1.1.12. *The set $\mathfrak{N}$ is closed.*

REMARK 1.1.4. It is easy to see that the level of the likelihood ratio test $\delta^{\infty,\varepsilon}$ is $\alpha = 0$ for all $\varepsilon \in [0, 1]$. Indeed, according to relations (1.1.1) and (1.1.35) we have for all $\varepsilon \in [0, 1]$

$$\alpha(\delta^{\infty,\varepsilon}) = \mathsf{P}(z > \infty) + \varepsilon\mathsf{P}(z = \infty) = \varepsilon\mathsf{P}(\mathfrak{z} = 0) = 0.$$

Thus statement 1) of Theorem 1.1.1 holds for $\alpha = 0$, too. Further, as can be seen from the proof of Lemma 1.1.11, $\delta^{*,0} = I(\mathfrak{z} = 0)$ is the most powerful test of level $\alpha = 0$ and that $\beta(\delta^{*,0}) = \overline{\beta}$. Taking into account (1.1.20) we obtain from (1.1.1) and (1.1.35) that

$$\beta(\delta^{\infty,\varepsilon}) = \mathsf{P}(z < \infty) + (1 - \varepsilon)\mathsf{P}(z = \infty) = \overline{\beta} + (1 - \varepsilon)(1 - \overline{\beta}) \geq \overline{\beta}.$$

This implies that the test $\delta^{\infty,\varepsilon}$ is the most powerful only for $\varepsilon = 1$. This means that statement 2) of Theorem 1.1.1 does not hold in general. As we proved above it only holds for $\varepsilon = 1$. Note that $\delta^{\infty,1} = I(z = \infty)$. Finally, the set S_c defined by (1.1.36) is of the form

$$S_\infty = (\delta^{*,0} \neq \delta^{\infty,\varepsilon}) \cap (z \neq \infty) = (\mathfrak{z} = 0) \cap (z \neq \infty) = (\mathfrak{z} = 0, \tilde{\mathfrak{z}} = 0)$$

for all $\varepsilon \in [0, 1]$. Moreover $\mathsf{P}(S_\infty) = \widetilde{\mathsf{P}}(S_\infty) = 0$. Therefore statement 3) of Theorem 1.1.1 holds for $\alpha = 0$, too.

REMARK 1.1.5. According to equality (1.1.19) the level of the likelihood ratio test $\delta^{0,\varepsilon}$ is

$$\alpha(\delta^{0,\varepsilon}) = \mathsf{P}(z > 0) + \varepsilon \mathsf{P}(z = 0) = \overline{\alpha} + \varepsilon(1 - \overline{\alpha})$$

for $\varepsilon \in [0, 1]$. This implies that if $\overline{\alpha} < 1$, then $\alpha(\delta^{0,\varepsilon}) = \alpha$ for $\alpha \in [\overline{\alpha}, 1]$ and $\varepsilon = (\alpha - \overline{\alpha})/(1 - \overline{\alpha})$. On the other hand, if $\alpha = 1$, then $\alpha(\delta^{0,\varepsilon}) = 1$ for all $\varepsilon \in [0, 1]$. This shows that statement 1) of Theorem 1.1.1 holds for $\alpha \in [\overline{\alpha}, 1]$, too. Further, we have

$$\beta(\delta^{0,\varepsilon}) = \widetilde{\mathsf{P}}(z < 0) + (1 - \varepsilon)\widetilde{\mathsf{P}}(z = 0) = (1 - \varepsilon)\widetilde{\mathsf{P}}(\tilde{\jmath} = 0, \jmath \geq 0) = 0$$

for all $\varepsilon \in [0, 1]$, that is, the likelihood ratio test $\delta^{0,\varepsilon}$ of an arbitrary level $\alpha \in [\overline{\alpha}, 1]$ is the most powerful. Thus statement 2) of Theorem 1.1.1 holds for $\alpha \in [\overline{\alpha}, 1]$, too. The set S_c defined by (1.1.36) is of the form

$$S_0 = (\delta^{*,\alpha} \neq \delta^{0,\varepsilon}) \cap (z > 0) = (\delta^{*,\alpha} \neq 1) \cap (z > 0)$$

for $c = 0$ and arbitrary $\varepsilon \in [0, 1]$ where $\delta^{*,\alpha}$ is the most powerful test of level $\alpha \in [\overline{\alpha}, 1]$. It can be seen from the proof of Lemma 1.1.11 that the most powerful test of level $\alpha = \overline{\alpha}$ is given by $\delta^{*,\overline{\alpha}} = I(\tilde{\jmath} > 0)$. Moreover $\delta^{*,1} \equiv 1$ (see the proof of Lemma 1.1.2). Thus the most powerful test of level α is

$$\delta^{*,\alpha} = \frac{1 - \alpha}{1 - \overline{\alpha}} I(\jmath > 0) + \frac{\alpha - \overline{\alpha}}{1 - \overline{\alpha}} = I(\jmath > 0) + \frac{\alpha - \overline{\alpha}}{1 - \overline{\alpha}} I(\jmath = 0).$$

Note that $(\delta^{*,\alpha} \neq 1) = \varnothing$ for $\alpha = 1$ and $(\delta^{*,\alpha} \neq 1) = (\jmath = 0)$ for $\alpha < 1$. Since $(z > 0) = (\tilde{\jmath} > 0)$, it holds that $S_0 = \varnothing$, whence $\mathsf{P}(S_0) = \widetilde{\mathsf{P}}(S_0) = 0$. Therefore statement 3) of Theorem 1.1.1 holds for $\alpha \in [\overline{\alpha}, 1]$, too, that is, all the statements of Theorem 1.1.1 hold for $\alpha \in [\overline{\alpha}, 1]$.

Neyman–Pearson test. The fundamental Neyman–Pearson lemma (Theorem 1.1.1) and Remarks 1.1.4 and 1.1.5 imply that for any $\alpha \in [0, 1]$ there exists a likelihood ratio test $\delta^{c(\alpha),\varepsilon(\alpha)}$ of level α where $(c(\alpha), \varepsilon(\alpha))$ is some solution of the equation $\alpha(\delta^{c,\varepsilon}) = \alpha$ with respect to (c, ε). Moreover $\delta^{c(\alpha),\varepsilon(\alpha)}$ is the most powerful test in the class $\mathcal{K}_\alpha$ if $\varepsilon(0) = 1$ and $\alpha = 0$. The test $\delta^{c(\alpha),\varepsilon(\alpha)}$ for $\varepsilon(0) = 1$ is called the *Neyman–Pearson test of level* α for distinguishing the hypotheses H and $\widetilde{H}$. In what follows we denote this test by $\delta^{+,\alpha}$. One can see from the proof of Theorem 1.1.1 and Remarks 1.1.4 and 1.1.5 that the functions $c(\alpha)$ and $\varepsilon(\alpha)$ can be taken of the form

$$c(\alpha) = \begin{cases} \infty, & \alpha = 0, \\ \overline{c}(\alpha), & 0 < \alpha < \overline{\alpha}, \\ 0, & \overline{\alpha} \leq \alpha \leq 1, \end{cases} \qquad \varepsilon(\alpha) = \begin{cases} 1, & \alpha = 0, \\ \overline{\varepsilon}(\alpha), & 0 < \alpha < \overline{\alpha}, \\ \widetilde{\varepsilon}(\alpha), & \overline{\alpha} \leq \alpha \leq 1, \end{cases} \tag{1.1.47}$$

where $\overline{c}(\alpha)$ is the minimal number c such that

$$\mathsf{P}(z > c) \leq \alpha \leq \mathsf{P}(z \geq c), \tag{1.1.48}$$

$$\overline{\varepsilon}(\alpha) = \frac{\alpha - \mathsf{P}(z > \overline{c}(\alpha))}{\mathsf{P}(z = \overline{c}(\alpha))}, \qquad \widetilde{\varepsilon}(\alpha) = \frac{\alpha - \overline{\alpha}}{1 - \overline{\alpha}}. \tag{1.1.49}$$

If $\mathsf{P}(z = \overline{c}(\alpha)) = 0$, then $\mathsf{P}(z > \overline{c}(\alpha)) = \alpha$ which leads to an expression $\overline{\varepsilon}(\alpha) = 0/0$. In this case an arbitrary number of the interval $[0, 1]$ can be taken as $\overline{\varepsilon}(\alpha)$. If $\overline{\alpha} = 1$, then $[\overline{\alpha}, 1] = \{1\}$ and this also results in an expression $\widetilde{\varepsilon}(1) = 0/0$. In this case an

arbitrary number of the interval $[0,1]$ can be taken as $\widetilde{\varepsilon}(1)$. The definition of the test $\delta^{+,\alpha}$ and equalities (1.1.47) imply

$$\beta(\delta^{+,\alpha}) = \begin{cases} \overline{\beta}, & \alpha = 0, \\ \widetilde{\mathsf{P}}(z < \overline{c}(\alpha)) + (1 - \bar{\varepsilon}(\alpha))\widetilde{\mathsf{P}}(z = \overline{c}(\alpha)), & 0 < \alpha < \overline{\alpha}, \\ 0, & \overline{\alpha} \le \alpha \le 1. \end{cases} \tag{1.1.50}$$

It is clear that the function $\beta(\delta^{+,\alpha})$ determines the lower boundary of the set $\mathfrak{N}$.

EXAMPLE 1.1.1. Let an observation ξ be a Gaussian random variable with the normal $\mathcal{N}(a,1)$ distribution under the hypothesis H and let its distribution be $\mathcal{N}(\widetilde{a},1)$ under the hypothesis $\widetilde{H}$. Then the measures P and $\widetilde{\mathsf{P}}$ corresponding to the distribution of the observation ξ under the hypotheses H and $\widetilde{H}$, respectively, are absolutely continuous and

$$z(x) = \frac{d\widetilde{\mathsf{P}}}{d\mathsf{P}}(x) = \exp\left((\widetilde{a} - a)x + \frac{a^2 - \widetilde{a}^2}{2}\right).$$

It is obvious in this case that $\overline{\alpha} = 1$ and $\overline{\beta} = 1$. Moreover, the random variable $z = z(\xi)$ has a continuous distribution for both hypotheses H and $\widetilde{H}$. Thus equalities (1.1.37) and (1.1.48) defining the constant $c(\alpha)$ for $0 < \alpha < 1$ become $\mathsf{P}(z < c) = 1 - \alpha$. For the sake of definiteness let $\widetilde{a} > a$. Then

$$\mathsf{P}(z < c) = \mathsf{P}\left(\xi < \frac{\ln c}{\widetilde{a} - a} + \frac{\widetilde{a} + a}{2}\right) = \Phi\left(\frac{\ln c}{\widetilde{a} - a} + \frac{\widetilde{a} - a}{2}\right) = 1 - \alpha$$

where $\Phi(x)$ is the distribution function of the normal $\mathcal{N}(0,1)$ law. This implies that

$$\frac{\ln c(\alpha)}{\widetilde{a} - a} + \frac{\widetilde{a} - a}{2} = t_{1-\alpha}$$

where t_p is the p-quantile of the law $\mathcal{N}(0,1)$, that is, $\Phi(t_p) = p$. Thus we have for all $\alpha \in (0,1)$ that

$$c(\alpha) = \exp\left((\widetilde{a} - a)t_{1-\alpha} - \frac{(\widetilde{a} - a)^2}{2}\right).$$

The number $\varepsilon(\alpha)$ can be chosen arbitrarily from the interval $[0,1]$. Taking into account (1.1.50) we get

$$\beta(\delta^{+,\alpha}) = \widetilde{\mathsf{P}}(z < c(\alpha)) = \widetilde{\mathsf{P}}(\xi < t_{1-\alpha} + a) = \Phi(t_{1-\alpha} - \widetilde{a} + a).$$

Note also that the Neyman–Pearson test of level α can be represented in the form $\delta^{+,\alpha} = I(z(x) > c(\alpha)) = I(x > t_{1-\alpha} + a)$.

EXAMPLE 1.1.2. Let an observation ξ have the normal $\mathcal{N}(0,1)$ distribution under the hypothesis H and the exponential distribution with the density

$$\tilde{\mathfrak{z}}(x) = e^{-x} I_{(0,\infty)}(x)$$

with respect to the Lebesgue measure under the hypothesis $\widetilde{H}$. Then $\widetilde{\mathsf{P}} \ll \mathsf{P}$, $\mathsf{P} \not\ll \widetilde{\mathsf{P}}$, and moreover $z(x) = d\widetilde{\mathsf{P}}/d\mathsf{P}(x)$ where $z(x) = 0$ for $x \le 0$, while

$$z(x) = \sqrt{\frac{\pi}{2}} \exp\left(\frac{(x-1)^2}{2}\right)$$

for $x > 0$. In this case $\overline{\beta} = 1$ and $\overline{\alpha} = \mathsf{P}(\tilde{z}(x) > 0) = \mathsf{P}(x > 0) = 1/2$. The random variable $z = z(\xi)$ is continuous under the hypothesis H. Thus the constant $c(\alpha)$ for $\alpha \in (0, \overline{\alpha}) = (0, 1/2)$ can be determined from the equation $\mathsf{P}(z < c) = 1 - \alpha$ or, equivalently, from

$$\Phi\left(1 + \sqrt{\ln \frac{ec^2}{2\pi}}\right) - \Phi\left(1 - \sqrt{\ln \frac{ec^2}{2\pi}}\right) = 1 - \alpha.$$

According to Remark 1.1.5 the Neyman–Pearson test of level α can be represented as

$$\delta^{+,\alpha} = I(z(x) > c(\alpha)) = I\left((x-1)^2 > \ln \frac{ec^2(\alpha)}{2\pi}\right)$$

for $\alpha \in [0, \overline{\alpha})$ and as

$$\delta^{+,\alpha} = I(x > 0) + (2\alpha - 1) I(x \le 0)$$

for $\alpha \in [\overline{\alpha}, 1]$, since $\overline{\alpha} = 1/2$. Thus $\varepsilon(\alpha)$ determined by equalities (1.1.47) and (1.1.49) for $\alpha \in [\overline{\alpha}, 1]$ is equal to $\widetilde{\varepsilon}(\alpha) = 2\alpha - 1$. The type II error probability of the test $\delta^{+,\alpha}$ is given by relation (1.1.50). Moreover

$$\beta(\delta^{+,\alpha}) = \widetilde{\mathsf{P}}\{|\xi - 1| < C(\alpha)\} = e^{-(1-C(\alpha))\vee 0} - e^{-1-C(\alpha)}$$

for $\alpha \in (0, \overline{\alpha})$ where

$$C(\alpha) = \sqrt{\ln \frac{ec^2(\alpha)}{2\pi}}.$$

If one interchanges the hypotheses H and $\widetilde{H}$ in Example 1.1.2, then $\mathsf{P} \ll \widetilde{\mathsf{P}}$ and $\widetilde{\mathsf{P}} \not\ll \mathsf{P}$. In this case $\overline{\alpha} = 1$ and $\overline{\beta} < 1$. Details are left to the reader.

EXAMPLE 1.1.3. Let an observation ξ assume two values 1 and 0 with probabilities p and $q = 1-p$ under the hypothesis H and with probabilities $\widetilde{p}$ and $\widetilde{q} = 1-\widetilde{p}$ under the hypothesis $\widetilde{H}$. Then $\mathsf{P} \sim \widetilde{\mathsf{P}}$ and the likelihood ratio $z(x) = d\widetilde{\mathsf{P}}/d\mathsf{P}(x)$ is given by

$$z(x) = \left(\frac{\widetilde{p}}{p}\right)^x \left(\frac{\widetilde{q}}{q}\right)^{1-x}, \qquad x = 0, 1.$$

Thus the random variable $z = z(\xi)$ assumes two values $\widetilde{p}/p$ and $\widetilde{q}/q$ with probabilities p and q under the hypothesis H and with probabilities $\widetilde{p}$ and $\widetilde{q}$ under the hypothesis $\widetilde{H}$. For the sake of definiteness let $\widetilde{p} > p$. Then $\widetilde{q}/q < 1 < \widetilde{p}/p$. Since $\mathsf{P} \sim \widetilde{\mathsf{P}}$, we get $\overline{\alpha} = \overline{\beta} = 1$. Solving equation (1.1.48) and evaluating $c(\alpha)$ and $\varepsilon(\alpha)$ we obtain

$$c(\alpha) = \begin{cases} 0, & \alpha = 1, \\ \widetilde{q}/q, & p \le \alpha < 1, \\ \widetilde{p}/p, & 0 \le \alpha < p, \end{cases}$$

and

$$\varepsilon(\alpha) = \begin{cases} 1, & \alpha = 1, \\ (\alpha - p)/q, & p \le \alpha < 1, \\ \alpha/p, & 0 \le \alpha < p. \end{cases}$$

This together with (1.1.50) implies

$$\beta(\delta^{+,\alpha}) = \begin{cases} \widetilde{q} + \dfrac{\widetilde{p}}{p}(p-\alpha), & 0 \le \alpha < p, \\ \dfrac{\widetilde{q}}{q}(1-\alpha), & p \le \alpha \le 1. \end{cases}$$

The function $\beta(\delta^{+,\alpha})$ determines the lower boundary of the set $\mathfrak{N}$ for $0 \le \alpha \le 1$. The set $\mathfrak{N}$ is shown in Figure 1.1.4.

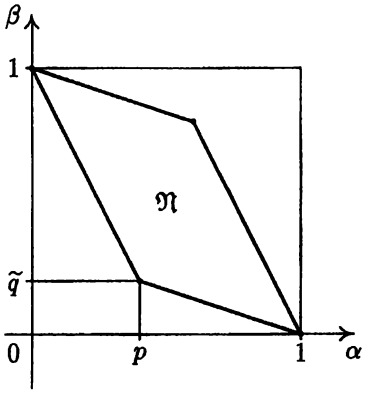

FIGURE 1.1.4

1.2. Distinguishing a finite number of simple hypotheses

Setting of the problem. The most powerful tests. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be the main probability space, let ξ be an observation that is a measurable mapping of the space $(\Omega, \mathscr{F})$ into some measurable space $(X, \mathscr{B})$, and let $\mathscr{P} = \{\mathsf{P}_1, \mathsf{P}_2, \dots, \mathsf{P}_N\}$, $1 < N < \infty$, be a family of probability measures defined on the space $(X, \mathscr{B})$. We assume that the distribution of the observation ξ is generated by some measure of the family $\mathscr{P}$.

Let $H_j = \{\theta = j\}$ be the hypothesis that the distribution of the observation ξ is generated by the measure P_j. We write in this case

$$\mathsf{P}_j(A) = \mathsf{P}_j\{\xi \in A\} = \mathsf{P}\{\xi \in A / H_j\}, \qquad A \in \mathscr{B},$$

and say that $\mathsf{P}_j(A)$ is the probability of the event $\{\xi \in A\}$ under the hypothesis $H_j = \{\theta = j\}$. The parameter θ assumes values in the set $\Theta = \{1, 2, \dots, N\}$ and is the index of the measure of the family $\mathscr{P}$ generating the distribution of the observation ξ. Thus we deal with N simple hypotheses $H_1, H_2, \dots, H_N$. Given an observation ξ, the problem is to decide which hypothesis of the set of hypotheses $H_1, H_2, \dots, H_N$ is true.

Any measurable mapping $\delta\colon (X, \mathscr{B}) \to \Theta$ is called a *statistical test for distinguishing N hypotheses $H_1, H_2, \dots, H_N$* by an observation ξ. The equality $\delta(x) = j$ means that the hypothesis H_j is accepted if $\xi = x$ (that is, $\theta = j$ in the parametric setting). A mapping δ is sometimes called a *decision rule* or a *decision function*. Every test δ uniquely determines a partition of the space (and vice versa) X into N disjoint measurable sets $X_j \in \mathscr{B}$, $j = 1, 2, \dots, N$, such that $X_j = \{x\colon \delta(x) = j\}$,

$j = 1, 2, \ldots, N$, $\bigcup_{j=1}^{N} X_j = X$. The test δ corresponding to a partition is closely related to the problem of estimation of an unknown parameter θ, namely $\delta(x)$ is an estimator of an unknown parameter θ if $\xi = x$. A test δ defined in this way is a *nonrandomized test* (see also Section 1.1). Any random variable $\delta = \delta(\xi)$ is called a test.

We consider below randomized tests defined as follows. Every measurable mapping $\delta\colon (X \times \Omega, \mathscr{B} \times \mathscr{F}) \to \Theta$ is called a *statistical test for distinguishing N hypotheses* $H_1, H_2, \ldots, H_N$ by an observation ξ. Given $\xi = x$, the hypothesis H_j is accepted if the random variable $\delta(x, \omega)$ is equal to j. If the function $\delta(x, \omega)$ does not depend on the variable ω, then the test δ is nonrandomized. Otherwise a test δ is called *randomized*. In general, every random variable $\delta = \delta(\xi(\omega), \omega)$, $\omega \in \Omega$, is called a *statistical test*.

Consider a family of functions $q^\delta(x) = (q_1^\delta(x), q_2^\delta(x), \ldots, q_N^\delta(x))$, $x \in X$, such that $q_j^\delta(x) = \mathsf{P}\{\delta = j/\xi = x\}$ is the conditional probability that the hypothesis H_j is accepted under the test δ given $\xi = x$ (that is, q_j^δ is the conditional probability of the event $\{\delta = j\}$ given $\xi = x$). It is clear that a test δ is uniquely determined by the family of conditional probabilities

$$q^\delta(x) = \big(q_1^\delta(x), q_2^\delta(x), \ldots, q_N^\delta(x)\big), \qquad x \in X.$$

Sometimes this family is called a *statistical test* (see Section 1.1). Note that

$$q_1^\delta(x) + \cdots + q_N^\delta(x) = 1$$

for all $x \in X$. If the functions $q_j^\delta(x)$ assume only two values 0 and 1, then the test δ is nonrandomized. Otherwise a test δ is called *randomized*. The decision domain of the hypotheses H_j, $j = 1, 2, \ldots, N$, is given by $X_j = \{x\colon q_j^\delta(x) = 1\}$ in the case of nonrandomized tests. Note that $X_j \cap X_i = \varnothing$, $i \neq j$, and $\bigcup_{j=1}^{N} X_j = X$.

The definition of a statistical test given in the preceding section differs to some extent from that given in this section, since the latter definition is inconvenient in the case $N = 2$. In what follows we use the simpler definition of the preceding section if we deal with the case of only two hypotheses.

To measure the quality of a test δ we introduce, as in the preceding section, the error probabilities:

$$\alpha_j(\delta) = \mathsf{P}\{\delta \neq j/H_j\} = \int_X (1 - q_j^\delta(x))\, \mathsf{P}_j(dx), \qquad j = 1, 2, \ldots, N. \tag{1.2.1}$$

The number $\alpha_j(\delta)$ is the probability to reject the hypothesis H_j by using the test δ if the hypothesis H_j is true. The number $\alpha_j(\delta)$ is called the *type j error probability of the test δ*.

It is natural to say that a test δ_1 is better than a test δ_2 if $\alpha_j(\delta_1) \leq \alpha_j(\delta_2)$ for all $j = 1, 2, \ldots, N$ and at least one of these inequalities is strict. However not all tests δ_1 and δ_2 can be compared in this way. In what follows we restrict the set of tests in order to have the possibility to compare them. Let

$$\mathcal{K}_{\alpha_1, \alpha_2, \ldots, \alpha_{N-1}} = \{\delta\colon \alpha_j(\delta) = \alpha_j, j = 1, 2, \ldots, N-1\} \tag{1.2.2}$$

where $\alpha_j \in [0, 1]$, $j = 1, 2, \ldots, N-1$, are some fixed numbers.

A test $\delta^* \in \mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$ is called the *most powerful test* (MP test, for short) *in the class* $K_{\alpha_1,\dots,\alpha_{N-1}}$ if

$$\alpha_N(\delta^*) \le \alpha_N(\delta)$$

for all tests $\delta \in \mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$.

Prior to constructing the most powerful tests in the class $\mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$ we consider two other approaches for comparing the tests, namely the Bayes and the minimax approaches.

The Bayes approach. Assume that the hypotheses $H_1, \dots, H_N$ are random events whose probabilities are known. Put

$$\mathsf{P}(H_j) = \pi_j, \qquad j = 1, 2, \dots, N.$$

The family of probabilities $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ is called the *a priori distribution of the hypotheses*. This family determines a distribution on the set $\Theta = \{1, 2, \dots, N\}$. The numbers $\mathsf{P}_j(A) = \mathsf{P}\{\xi \in A/H_j\}$ are conditional probabilities of the event $\{\xi \in A\}$ given the event H_j occurs. Moreover, we assume that the loss is A where $A = A_{ij}$ if the hypothesis H_j is accepted, while the hypothesis H_i is true. Therefore, the loss A is a random variable whose values are uniquely determined by the test δ and the index of the true hypothesis. To compare tests in the Bayes approach we use the *risk of a test* δ defined as the expectation of the loss:

$$R(\delta) = \mathsf{E}A = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} p^{\delta}_{j/i} \pi_i \tag{1.2.3}$$

where $p^{\delta}_{j/i} = \mathsf{P}\{\delta = j/H_i\}$. Since

$$p^{\delta}_{j/i} = \int_X q^{\delta}_j(x)\, \mathsf{P}_j(dx), \tag{1.2.4}$$

the risk (1.2.3) can be rewritten as

$$R(\delta) = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\pi_i \int_X q^{\delta}_j(x)\, \mathsf{P}_j(dx). \tag{1.2.5}$$

A test $\delta_{\pi,A}$ that minimizes the risk $R(\delta)$ is called the *Bayes test corresponding to a priori distribution* π *and loss* A.

Let μ be some σ-finite measure on $(X, \mathscr{B})$ dominating the family of probability measures $\mathscr{P} = \{\mathsf{P}_1, \mathsf{P}_2, \dots, \mathsf{P}_N\}$, $1 < N < \infty$, and let $p_i(x)$ be the density of the measure P_i with respect to the measure μ. Note that such a measure μ always exists. In particular, the measure $\mu = \sum_{i=1}^{N} c_i \mathsf{P}_i$ where $c_i > 0$ for all i possesses this property. The risk (1.2.5) can be expressed in terms of the measure μ and densities $p_i(x)$:

$$R(\delta) = \int_X \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} q^{\delta}_j(x) p_i(x) \pi_i\, \mu(dx). \tag{1.2.6}$$

Consider the measure $\overline{\mathsf{P}} = \sum_{i=1}^{N} \pi_i \mathsf{P}_i$ which defines the unconditional distribution of the observation ξ. Note that $\overline{\mathsf{P}}$ is absolutely continuous with respect to μ

and $f(x) = \sum_{i=1}^{N} \pi_i p_i(x)$ is the density of the measure $\overline{\mathsf{P}}$ with respect to the measure μ. Let $C = \{x: f(x) > 0\}$. Then $\overline{\mathsf{P}}(X \setminus C) = 0$ and $\mathsf{P}_i(X \setminus C) = 0$ for all i. For all $x \in C$ consider the functions

$$\pi_i(x) = \frac{p_i(x)\pi_i}{f(x)}, \qquad i = 1, 2, \ldots, N. \tag{1.2.7}$$

Equality (1.2.7) is the well-known Bayes formula for evaluating the conditional probability $\pi_i(x)$ of a hypothesis H_i given $\xi = x$. The numbers $\pi_i(x)$ are called *a priori probabilities of hypotheses* H_i.

THEOREM 1.2.1. *For all tests* δ *it holds that*

$$R(\delta) \geq \mathsf{E} \min_{1 \leq j \leq N} \sum_{i=1}^{N} A_{ij} \pi_i(\xi). \tag{1.2.8}$$

A test $\delta = \delta_{\pi,A}$ *is a Bayes test with respect to a priori distribution* π *and loss* A *if and only if*

$$q_k^{\delta}(x) = 1 \quad \text{if} \quad \sum_{i=1}^{N} A_{ik} \pi_i(x) = \min_{1 \leq j \leq N} \sum_{i=1}^{N} A_{ij} \pi_i(x) \tag{1.2.9}$$

almost surely with respect to $\overline{\mathsf{P}}$. *If* $\delta = \delta_{\pi,A}$, *then inequality* (1.2.8) *becomes an equality.*

PROOF. We obtain from equalities (1.2.6) and (1.2.7) that

$$\begin{aligned} R(\delta) &= \int_C \sum_{j=1}^{N} q_j^{\delta}(x) \sum_{i=1}^{N} A_{ij} \pi_i(x) f(x)\, \mu(dx) \\ &\geq \int_C \min_{1 \leq j \leq N} \sum_{i=1}^{N} A_{ij} \pi_i(x)\, \overline{\mathsf{P}}(dx) = \mathsf{E} \min_{1 \leq j \leq N} \sum_{i=1}^{N} A_{ij} \pi_i(\xi) \end{aligned} \tag{1.2.10}$$

where $C = \{x: f(x) > 0\}$, $\overline{\mathsf{P}} = \sum_{i=1}^{N} \pi_i \mathsf{P}_i$, and $f(x)$ is the density of the measure $\overline{\mathsf{P}}$ with respect to the measure μ. Thus inequality (1.2.8) is proved.

The sufficiency of condition (1.2.9) follows from (1.2.10). The case of an equality in (1.2.8) also follows from (1.2.10).

Now we prove the necessity of condition (1.2.9) by contradiction. Let $\delta = \delta_{\pi,A}$ be a Bayes test such that $q_k^{\delta}(x) = 1$ and

$$\sum_{i=1}^{N} A_{ik} \pi_i(x) > \sum_{i=1}^{N} A_{il} \pi_i(x) = \min_{1 \leq j \leq N} \sum_{i=1}^{N} A_{ij} \pi_i(x)$$

for $x \in A$, where A is some event of positive probability, $\overline{\mathsf{P}}(A) > 0$. Let δ_1 be a test that differs from δ only on the event A and such that $q_l^{\delta_1}(x) = 1$ for $x \in A$. Then

$$\begin{aligned} R(\delta_1) &= \int_A \sum_{i=1}^{N} A_{il} \pi_i(x)\, \overline{\mathsf{P}}(dx) + \int_{X \setminus A} \sum_{j=1}^{N} q_j^{\delta_1}(x) \sum_{i=1}^{N} A_{ij} \pi_i(x)\, \overline{\mathsf{P}}(dx) \\ &< \int_A \sum_{i=1}^{N} A_{ik} \pi_i(x)\, \overline{\mathsf{P}}(dx) + \int_{X \setminus A} \sum_{j=1}^{N} q_j^{\delta}(x) \sum_{i=1}^{N} A_{ij} \pi_i(x)\, \overline{\mathsf{P}}(dx) \\ &= R(\delta). \end{aligned}$$

This is a contradiction, since δ is a Bayes test. The necessity of condition (1.2.9) is proved. □

If $A_{ij} = 1 - \delta_{ij}$ where δ_{ij} is the Kronecker symbol (that is, $\delta_{ij} = 0$ for $i = j$ and $\delta_{ij} = 1$ for $i \neq j$), then the Bayes test $\delta_{\pi,A}$ for the loss A is called the *maximum a posteriori probability test*. In this case

$$\sum_{i=1}^{N} A_{ij}\pi_i(x) = \sum_{i \neq j} \pi_i(x) = 1 - \pi_j(x)$$

and condition (1.2.9) defining the Bayes test can be rewritten as

$$q_k^{\delta}(x) = 1 \quad \text{if } \pi_k(x) = \max_{1 \leq j \leq N} \pi_j(x). \tag{1.2.11}$$

The risk of an arbitrary test δ with loss $A_{ij} = 1 - \delta_{ij}$ is of the form

$$R(\delta) = \mathsf{E}A = \sum_{i=1}^{N} \sum_{j \neq i} p_{j/i}^{\delta}\pi_i = \mathsf{P}\{\delta \neq \theta\} = e_{\pi}(\delta) \tag{1.2.12}$$

where θ is the index of a hypothesis and $\{\theta = j\} = H_j$. We see that $R(\delta)$ in this case is the unconditional probability of a wrong decision $e_{\pi}(\delta)$ for the test δ. Thus the maximum a posteriori probability test minimizes the error probability $e_{\pi}(\delta)$ of the test δ. Taking into account (1.2.1) and (1.2.4) we obtain from (1.2.12) that

$$e_{\pi}(\delta) = \sum_{i=1}^{N} \mathsf{P}\{\delta \neq i/H_i\}\pi_i = \sum_{i=1}^{N} \alpha_i(\delta)\pi_i. \tag{1.2.13}$$

If $N = 2$ and $A_{ij} = 1 - \delta_{ij}$, then according to (1.2.11) the maximum a posteriori probability test is of the form

$$q_2^{\delta}(x) = \begin{cases} 1, & \pi_2(x) > \pi_1(x), \\ 0, & \pi_2(x) < \pi_1(x), \end{cases} \qquad q_1^{\delta}(x) = 1 - q_2^{\delta}(x). \tag{1.2.14}$$

Moreover, if $\pi_2(x) = \pi_1(x)$, then one can put either $q_2^{\delta}(x) = 1$ or $q_1^{\delta}(x) = 1$. Applying equality (1.2.7) one can rewrite equality (1.2.14) as

$$q_2^{\delta}(x) = \begin{cases} 1, & \pi_2 p_2(x) > \pi_1 p_1(x), \\ 0, & \pi_2 p_2(x) < \pi_1 p_1(x), \end{cases} \qquad q_1^{\delta}(x) = 1 - q_2^{\delta}(x). \tag{1.2.15}$$

Note that condition (1.2.9) does not uniquely determine the test $\delta_{\pi,A}$. In particular, it does not uniquely determine which hypothesis should be accepted if two or more numbers among $\sum_{i=1}^{N} A_{ij}\pi_i(x)$ are maximal. This is a matter of definition of the probabilities $q^{\delta}(x) = \left(q_1^{\delta}(x), q_2^{\delta}(x), \ldots, q_N^{\delta}(x)\right)$ on the boundaries

$$\Gamma_k = \left\{ x : \sum_{i=1}^{N} A_{ik}\pi_i(x) = \min_{j \neq k} \sum_{i=1}^{N} A_{ij}\pi_i(x) \right\}$$

of the sets

$$\widetilde{X}_k = \left\{ x : \sum_{i=1}^{N} A_{ik}\pi_i(x) < \min_{j \neq k} \sum_{i=1}^{N} A_{ij}\pi_i(x) \right\}.$$

Using condition (1.2.9) the hypothesis H_k is accepted on Γ_k according to the test $\delta_{\pi,A}$. Therefore the problem is to decide about the points of the boundary Γ_k to be

included into the set $\widetilde{X}_k$ where the hypothesis H_k is accepted. One of the possible approaches is to include points of Γ_k to any of the regions $\widetilde{X}_j$ adjacent to Γ_k; in this case condition (1.2.9) holds and the risk $R(\delta_{\pi,A})$ does not change. More precisely, if $A \subset \Gamma_{k_1} \cap \cdots \cap \Gamma_{k_l}$, then, according to the Bayes test, it makes no difference for $x \in A$ which hypothesis among $H_{k_1}, \ldots, H_{k_l}$ is accepted. Moreover, one can accept the hypotheses $H_{k_1}, \ldots, H_{k_l}$ randomly with probabilities $q^{\delta}_{k_1}(x), \ldots, q^{\delta}_{k_l}(x)$, $\sum_{i=1}^{l} q^{\delta}_{k_i}(x) = 1$. The risk $R(\delta_{\pi,A})$ does not change in this case.

The general definition of a Bayes test $\delta_{\pi,A}$ is based on the sets

$$\Gamma_{k_1,\ldots,k_l} = \bigcap_{i=1}^{l} \Gamma_{k_i} \bigcap_{j \neq k_1,\ldots,k_l} \overline{\Gamma}_j \tag{1.2.16}$$

where $\overline{\Gamma}_j = X \setminus \Gamma_j$. As $q^{\delta}(x)$ for $x \in \Gamma_{k_1,\ldots,k_l}$ one can take an arbitrary vector of the set $R_{k_1,\ldots,k_l}$ of vectors $(q_1, q_2, \ldots, q_N)$ with $q_1 \geq 0$, $q_2 \geq 0, \ldots, q_N \geq 0$ and $\sum_{i=1}^{N} q_i = 1$ and whose coordinates with indices different from $k_1, \ldots, k_l$ are zero. It is clear that the set R_k includes only one vector e_k whose k-th coordinate is 1, while all others are zero. Thus one should put $q^{\delta}(x) = e_k$ for $x \in \widetilde{X}_k$. This implies the following improvement of Theorem 1.2.1.

THEOREM 1.2.2. *A test δ is Bayes if and only if*

$$q^{\delta}(x) = \begin{cases} e_k, & \text{for } x \in \widetilde{X}_k, \\ R_{k_1,\ldots,k_l}, & \text{for } x \in \Gamma_{k_1,\ldots,k_l} \end{cases}$$

for $\overline{\mathsf{P}}$-almost all x where $\Gamma_{k_1,\ldots,k_l}$ are the sets defined by (1.2.16).

Theorems 1.2.1 and 1.2.2 show that randomized tests do not decrease the risk $R(\delta)$, however they enlarge the set of different Bayes tests $\delta_{\pi,A}$. Moreover Theorems 1.2.1 and 1.2.2 imply that among Bayes tests $\delta_{\pi,A}$ there is at least one nonrandomized test.

Let $N = 2$. Then

$$\widetilde{X}_1 = \left\{ x : \sum_{i=1}^{2} A_{i1}\pi_i(x) < \sum_{i=1}^{2} A_{i2}\pi_i(x) \right\},$$

$$\widetilde{X}_2 = \left\{ x : \sum_{i=1}^{2} A_{i2}\pi_i(x) < \sum_{i=1}^{2} A_{i1}\pi_i(x) \right\},$$

$$\Gamma_1 = \Gamma_2 = \left\{ x : \sum_{i=1}^{2} A_{i1}\pi_i(x) = \sum_{i=1}^{2} A_{i2}\pi_i(x) \right\}.$$

Hence we deal with a single set $\Gamma_{1,2} = \Gamma_1 = \Gamma_2$ instead of sets (1.2.16); moreover $R_{1,2} = \{(q_1, q_2) : q_1 \geq 0, q_2 \geq 0, q_1 + q_2 = 1\}$ in this case. According to Theorem 1.2.2 we obtain for the Bayes test $\delta_{\pi,A}$ that $q^{\delta}_1(x) = 1$ for $x \in \widetilde{X}_1$ and $q^{\delta}_2(x) = 1$ for $x \in \widetilde{X}_2$. As $q^{\delta}(x) = (q^{\delta}_1(x), q^{\delta}_2(x))$ for $x \in \Gamma_{1,2}$ one can take an arbitrary function with values in $R_{1,2}$.

In the case of $N = 2$, the Bayes test $\delta = \delta_{\pi,A}$ equals the maximum a posteriori probability for $A_{ij} = 1 - \delta_{ij}$. Applying (1.2.14) and (1.2.15) we represent the test

$\delta = \delta_{\pi,A}$ in the form

$$q_2^\delta(x) = \begin{cases} 1, & z_{2,1}(x) > c, \\ q(x), & z_{2,1}(x) = c, \\ 0, & z_{2,1}(x) < c, \end{cases} \qquad q_1^\delta(x) = 1 - q_2^\delta(x) \tag{1.2.17}$$

where $c = \pi_1/\pi_2$, $z_{2,1}(x) = p_2(x)/p_1(x)$ is the likelihood ratio (we assume that $0/0 = 0$), and $q(x)$ is an arbitrary measurable function with values in $[0,1]$. A test of the form (1.2.17) for an arbitrary function $q(x)$ is called the *likelihood ratio test*. Note that we considered in the preceding section a likelihood ratio test of the form (1.2.17) for a specific function $q(x)$ being constant on the set $\{z_{2,1}(x) = c\}$ (see (1.1.35)). Like the preceding section, we denote by $\delta^{c,q}$ the likelihood ratio test δ defined by relation (1.2.17).

The minimax approach. The quality of a test δ in the minimax approach is measured by

$$e(\delta) = \max_{1 \le j \le N} \alpha_j(\delta) = \max_\pi e_\pi(\delta) \tag{1.2.18}$$

where $e_\pi(\delta)$ is the unconditional error probability of the test δ defined in (1.2.13). Recall that $e_\pi(\delta)$ is the risk of the test δ if a priori distribution is determined by the vector $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ and loss is $A_{ij} = 1 - \delta_{ij}$.

A test δ^* such that

$$e(\delta^*) = \min_\delta e(\delta) \tag{1.2.19}$$

is called *minimax*, where $e(\delta)$ is the maximal error probability of the test δ (see relation (1.2.18)).

The following result contains a sufficient condition that a test is minimax.

THEOREM 1.2.3. *Let there exist a Bayes test $\bar\delta$ (with respect to some a priori distribution $\bar\pi = (\bar\pi_1, \dots, \bar\pi_N)$ and loss $A_{ij} = 1 - \delta_{ij}$) such that*

$$\alpha_1(\bar\delta) = \cdots = \alpha_N(\bar\delta). \tag{1.2.20}$$

Then the test $\bar\delta$ is minimax.

PROOF. For all tests δ it holds that

$$e(\delta) \ge \sum_{j=1}^N \bar\pi_j \alpha_i(\delta) \ge \sum_{j=1}^N \bar\pi_j \alpha_i(\bar\delta) = \max_{1 \le j \le N} \alpha_i(\bar\delta) = e(\bar\delta),$$

whence it follows that the test $\bar\delta$ is minimax. □

Let $\bar\pi$ be a test satisfying (1.2.20). The a priori distribution $\bar\pi$ corresponding to the test $\bar\pi$ is called the *worst* or the *least favorable*. This notion is explained by saying that the maximum

$$\max_\pi e_\pi(\delta_\pi) = \max_\pi \min_\delta e_\pi(\delta) \tag{1.2.21}$$

is attained at $\pi = \bar\pi$ where δ_π is the *Bayes test with respect to a priori distribution π and loss $A_{ij} = 1 - \delta_{ij}$*. Therefore the minimax test satisfying (1.2.20) is the Bayes test with the maximal error probability. The proof of equality (1.2.21) and of the

existence of the worse a priori distribution and the minimax test can be found in [8, 10]. More detail on the minimax approach can be found in [4, 52].

In the case $N = 2$, the minimax test δ^* can be found by applying the fundamental Neyman–Pearson lemma and the set $\mathfrak{N}$ described in Section 1.1. By Theorem 1.2.3 the minimax test δ^* is a Bayes test $\bar{\delta}$ with respect to some a priori distribution $\bar{\pi} = (\bar{\pi}_1, \bar{\pi}_2)$ and loss $A_{ij} = 1 - \delta_{ij}$ such that $\alpha_1(\bar{\delta}) = \alpha_2(\bar{\delta})$ provided such a test exists. Such a Bayes test $\bar{\delta}$ exists, indeed, and it corresponds to the point $A = (\alpha_1(\bar{\delta}), \alpha_2(\bar{\delta})) \in \mathfrak{N}$ of the intersection of the lower boundary of the set $\mathfrak{N}$ and the diagonal of the square $[0,1] \times [0,1]$ joining its corners $(0,0)$ and $(1,1)$ (see Figure 1.2.1).

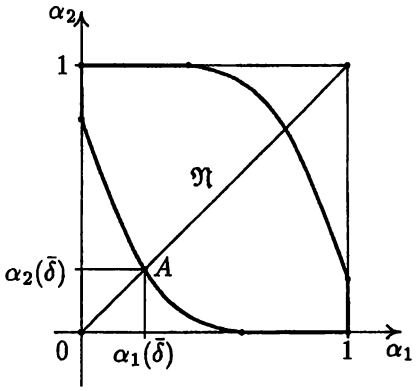

FIGURE 1.2.1

Moreover the following result gives an explicit form of the test $\bar{\delta}$.

THEOREM 1.2.4. *There is a likelihood test $\delta^{c,q}$ that is a minimax test. The parameters c and $q(x) \equiv q = \text{const}$ of the likelihood test $\delta^{c,q}$ are determined by the equation $\alpha_1(\delta^{c,q}) = \alpha_2(\delta^{c,q})$.*

PROOF. According to Theorem 1.2.3 it is sufficient to find a Bayes test $\bar{\delta}$ corresponding to some a priori distribution $\bar{\pi} = (\bar{\pi}_1, \bar{\pi}_2)$ and loss $A_{ij} = 1 - \delta_{ij}$ such that $\alpha_1(\bar{\delta}) = \alpha_2(\bar{\delta})$ and $\bar{\delta}$ coincides with a likelihood ratio test $\delta^{\bar{c},\bar{q}}$ for some parameters $\bar{c}$ and $\bar{q}(x) \equiv \bar{q} = \text{const}$. We have seen above that such a Bayes test exists and it corresponds to the point $A = (\alpha_1(\bar{\delta}), \alpha_2(\bar{\delta})) \in \mathfrak{N}$ of the intersection of the lower boundary of the set $\mathfrak{N}$ and the straight line joining the points $(0,0)$ and $(1,1)$ (see Figure 1.2.1). According to Theorems 1.2.1 and 1.2.2 this Bayes test $\bar{\delta}$ coincides with the likelihood ratio test $\delta^{\bar{c},\bar{q}}$ for some parameters $\bar{c}$ and $\bar{q}$ where $\bar{c} = \bar{\pi}_1/\bar{\pi}_2$ and $\bar{\pi} = (\bar{\pi}_1, \bar{\pi}_2)$ is the a priori distribution corresponding to the Bayes test $\bar{\delta}$. □

REMARK 1.2.1. Since an arbitrary Bayes test coincides with a likelihood ratio test $\delta^{c,q}$ for some constants c and q, the proof of Theorem 1.2.4 follows from the existence of a solution of the equation $\alpha_1(\delta^{c,q}) = \alpha_2(\delta^{c,q})$ with respect to (c,q). Note that this equation is of the form

$$\mathsf{P}_1(z_{2,1} > c) + \mathsf{P}_2(z_{2,1} > c) + q[\mathsf{P}_1(z_{2,1} = c) + \mathsf{P}_2(z_{2,1} = c)] = 1.$$

The proof that this equation has a solution with respect to (c,q) is the same as that of the existence of a solution of the equation $\alpha_1(\delta^{c,q}) = \alpha \in [0,1]$ used in Theorem 1.1.1 and in Remarks 1.1.4 and 1.1.5.

The most powerful tests. We turn to the construction of the most powerful test in the class $\mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$ defined by relation (1.2.2). Like the case of the minimax test, one can apply Bayes tests for this purpose, too.

THEOREM 1.2.5. *Let there exist an a priori distribution* $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ *such that*

$$\alpha_j(\delta_\pi) = \alpha_j, \qquad j = 1, 2, \dots, N-1, \tag{1.2.22}$$

where δ_π *is the Bayes test corresponding to a priori distribution* π *and loss* $A_{ij} = 1 - \delta_{ij}$. *Then* δ_π *is the most powerful test in the class* $\mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$.

PROOF. The definition of a Bayes test implies that

$$e_\pi(\delta_\pi) \le e_\pi(\delta)$$

for all tests δ, whence

$$\sum_{j=1}^{N} \pi_j \alpha_j(\delta_\pi) \le \sum_{j=1}^{N-1} \pi_j \alpha_j + \pi_N \alpha_N(\delta) \tag{1.2.23}$$

for all tests $\delta \in \mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$. Condition (1.2.22) implies that $\alpha_j(\delta_\pi) = \alpha_j$ for $j \le N-1$. Then it follows from (1.2.23) that $\alpha_N(\delta_N) \le \alpha_N(\delta)$, that is, δ_π is the most powerful test in the class $\mathcal{K}_{\alpha_1,\dots,\alpha_{N-1}}$. □

REMARK 1.2.2. One can treat equalities (1.2.22) (and equalities (1.2.20), too) as the system of $N-1$ equations and use it to evaluate the a priori distribution π and the corresponding Bayes test δ_π. Generally speaking, this Bayes test δ_π is randomized. See Section 1.1 for another method of finding the most powerful test for $N = 2$ that does not use the Bayes tests.

EXAMPLE 1.2.1 (Change point problem). Let an observation be a sample $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ where $\xi_1, \dots, \xi_n$ are independent random variables. The first $\theta - 1$ of them have a distribution G_1, while all other random variables have a distribution G_2 and $G_2 \ne G_1$. The number θ is called the *change point* (of the distribution). Possible values of θ are $1, 2, \dots, n$. Let $H_j = \{\theta = j\}$, $j = 1, 2, \dots, n$, be the statistical hypotheses about the parameter θ. Without loss of generality we assume that the measures G_1 and G_2 have densities $g_1(x)$ and $g_2(x)$, respectively, with respect to some σ-finite measure μ. Then the measure $P_j^{(n)}$ generating the distribution of $\xi^{(n)}$ under the hypothesis H_j is absolutely continuous with respect to the measure μ^n and its density is given by

$$p_j(x) = \prod_{i=1}^{j-1} g_1(x_i) \prod_{i=j}^{n} g_2(x_i), \qquad x = (x_1, x_2, \dots, x_n),$$

where we put $\prod_{i=1}^{0} = 1$.

Let δ^* be the maximum a posteriori probability test and let the a priori distribution be uniform, that is, δ^* is the Bayes test corresponding to the loss $A_{ij} = 1 - \delta_{ij}$ where δ_{ij} is the Kronecker symbol and the a priori distribution is $\pi = \pi_n^* = (1/n, 1/n, \dots, 1/n)$. According to (1.2.11) the test δ^* is such that

$$p_{\delta(x)^*}(x) = \max_{1 \le j \le n} p_j(x), \qquad x \in \mathbf{R}^n, \tag{1.2.24}$$

that is, the test δ^* maximizes the density $p_j(x)$. A test δ^* satisfying equality (1.2.24) is called the *maximum likelihood test* (see [9]). Condition (1.2.24) means that for all $j = 1, 2, \ldots, n$

$$\prod_{i=1}^{\delta^*(x)-1} g_1(x_i) \prod_{i=\delta^*(x)}^{n} g_2(x_i) \geq \prod_{i=1}^{j-1} g_1(x_i) \prod_{i=j}^{n} g_2(x_i). \tag{1.2.25}$$

Dividing inequality (1.2.25) by $p_1(x)$ one can prove that the test $\delta^*(x)$ is such that

$$\prod_{i=1}^{\delta^*(x)-1} \frac{g_1(x_i)}{g_2(x_i)} \geq \prod_{i=1}^{j-1} \frac{g_1(x_i)}{g_2(x_i)}$$

for all $j = 1, 2, \ldots, n$. Properties of the test $\delta^*(x)$ are studied in detail in [9], §72.

To complete a brief discussion of the change point problem we mention [9], a survey paper [30], and a monograph [46] where more detail is given about the change point analysis (however there is an extensive literature devoted to this topic).

1.3. Distinguishing composite hypotheses

The setting of the problem and main definitions. Let ξ be an observation that is a random element assuming values in a measurable space $(X, \mathscr{B})$ and let $\mathscr{P} = \{\mathsf{P}_\theta; \theta \in \Theta\}$ be a parametric family of probability measures defined on $(X, \mathscr{B})$ where Θ is some set containing more than two points. Let the distribution of the random element ξ be generated by a measure of the family $\mathscr{P}$.

Let $\mathscr{P}_i = \{\mathsf{P}_\theta; \theta \in \Theta_i\}$, $i = 1, 2$, where $\Theta_1 \cap \Theta_2 = \varnothing$ and $\Theta_1 \cup \Theta_2 = \Theta$, so that $\mathscr{P}_1 \cup \mathscr{P}_2 = \mathscr{P}$. Consider the hypotheses H_1 and H_2 that the distribution of the element ξ belongs to the sets $\mathscr{P}_1$ and $\mathscr{P}_2$, respectively. For the sake of brevity we write $H_i\colon \theta \in \Theta_i$, $i = 1, 2$. It is clear that at least one of the hypotheses H_1 and H_2 is *composite*, since at least one of the sets Θ_1 and Θ_2 contains at least two points. Consider the problem of distinguishing two hypotheses H_1 and H_2 by the observation $\xi = x$, $x \in X$.

As in Section 1.1 we consider a statistical test δ for distinguishing hypotheses H_1 and H_2 by the observation $\xi = x$. The test is a measurable mapping

$$\delta\colon (X, \mathscr{B}) \to \big([0, 1], \mathscr{B}([0, 1])\big).$$

We treat $\delta(x)$ as the probability that the hypothesis H_2 is accepted if $\xi = x$, while $1 - \delta(x)$ is the probability that the hypothesis H_1 is accepted if $\xi = x$. We also put $\delta = \delta(\xi)$.

To measure the quality of a test δ we consider the function

$$\beta(\delta; \theta) = \mathsf{E}_\theta \delta, \qquad \theta \in \Theta, \tag{1.3.1}$$

where E_θ is the expectation with respect to the distribution P_θ. The function $\beta(\delta; \theta)$ is called the *power function of the test* δ. It is clear that $\beta(\delta; \theta)$ for $\theta \in \Theta_1$ is the probability of a wrong decision, while for $\theta \in \Theta_2$ it is the probability of a correct decision.

DEFINITION 1.3.1. We say that a test δ_1 is *uniformly more powerful* than a test δ_2 if

$$\beta(\delta_1;\theta) \le \beta(\delta_2,\theta) \quad \text{for all } \theta \in \Theta_1, \tag{1.3.2}$$

$$\beta(\delta_1;\theta) \ge \beta(\delta_2,\theta) \quad \text{for all } \theta \in \Theta_2, \tag{1.3.3}$$

and at least one of the inequalities for at least one θ is strict.

DEFINITION 1.3.2. A test that is uniformly more powerful than any other test is called the *uniformly most powerful test* (UMP test).

There is a different approach for measuring the quality of tests, namely the Bayes approach. When following this approach we consider a random variable A treated as a loss: it assumes the value $A_i(\theta)$ if the hypothesis H_i is accepted and the true parameter is θ. The mean loss of the test δ is

$$\mathsf{E}_\theta^\delta A = A_1(\theta)\mathsf{E}_\theta(1-\delta) + A_2(\theta)\mathsf{E}_\theta\delta = (A_2(\theta) - A_1(\theta))\beta(\delta;\theta) + A_1(\theta) \tag{1.3.4}$$

if the parameter is θ. Here E_θ^δ is the expectation with respect to the distribution generated by the test δ if the true parameter is θ and $\beta(\delta;\theta)$ is the power function of the test δ defined by (1.3.1).

Let a σ-algebra $\mathscr{B}(\Theta)$ of measurable subsets of Θ be given and let a probability measure $\mathbf{Q}$ be defined on the measurable space $(\Theta, \mathscr{B}(\Theta))$, that is, θ is a random parameter and $\mathbf{Q}(B) = \mathsf{P}\{\theta \in B\}$, $B \in \mathscr{B}(\Theta)$. Applying (1.3.4) one can evaluate the mean loss:

$$\mathsf{E}^\delta A = \int_\Theta \mathsf{E}_t^\delta A \, \mathbf{Q}(dt) = \int_\Theta A_1(t)\, \mathbf{Q}(dt) + \int_\Theta (A_2(t) - A_1(t))\beta(\delta_i;t)\, \mathbf{Q}(dt) \tag{1.3.5}$$

where E^δ is the expectation with respect to the distribution generated by the test δ. The measure $\mathbf{Q}$ is sometimes called *a priori measure* or *a priori distribution.*

DEFINITION 1.3.3. A test $\delta_{A,Q}$ is called *Bayes* with respect to the loss A and a priori distribution $\mathbf{Q}$ if

$$\mathsf{E}^{\delta_{A,Q}} A \le \mathsf{E}^\delta A \tag{1.3.6}$$

for any test δ where $\mathsf{E}^\delta A$ is the mean loss defined by (1.3.5).

The following result allows one to compare the quality of tests in the Bayes approach and in an approach based on the power function.

THEOREM 1.3.1. *Let $\mathbf{Q}$ be an arbitrary a priori measure. Assume that the loss function $A_i(t)$ is such that*

$$A_1(t) \le A_2(t) \quad \textit{for all } t \in \Theta_1, \tag{1.3.7}$$

$$A_1(t) \ge A_2(t) \quad \textit{for all } t \in \Theta_2. \tag{1.3.8}$$

If a test δ_1 is uniformly more powerful than a test δ_2, then their mean losses are such that

$$\mathsf{E}^{\delta_1} A \le \mathsf{E}^{\delta_2} A. \tag{1.3.9}$$

PROOF. Let a test δ_1 be uniformly more powerful than a test δ_2. Then inequalities (1.3.2) and (1.3.3) hold. Taking into account inequalities (1.3.2) and (1.3.3) and conditions (1.3.7) and (1.3.8) we obtain from (1.3.5) that

$$\mathsf{E}^{\delta_1}A - \mathsf{E}^{\delta_2}A = \int_{\Theta_1} (A_2(t) - A_1(t))(\beta(\delta_1;t) - \beta(\delta_2;t))\,\mathbf{Q}(dt) + \int_{\Theta_2} (A_2(t) - A_1(t))(\beta(\delta_1;t) - \beta(\delta_2;t))\,\mathbf{Q}(dt) \le 0,$$

that is, inequality (1.3.9) is proved. □

COROLLARY 1.3.1. *If δ^* is a UMP test, then it also is a Bayes test with respect to an arbitrary a priori distribution $\mathbf{Q}$ and any loss function $A_i(t)$ satisfying conditions* (1.3.7) *and* (1.3.8).

REMARK 1.3.1. Conditions (1.3.7) and (1.3.8) posed on the loss functions $A_1(t)$ and $A_2(t)$ are natural in the sense that if $t \in \Theta_1$, then $A_1(t)$ is the loss due to the acceptance of the hypothesis H_1 if it is true, while $A_2(t)$ is the loss due to the acceptance of the alternative hypothesis H_2 if the hypothesis H_1 is true. Thus it is reasonable to assume that $A_1(t) \le A_2(t)$ for $t \in \Theta_1$. The same remark can be made regarding inequality (1.3.8).

REMARK 1.3.2. The method of comparing the quality of tests based on Definition 1.3.3 is sometimes called the *complete Bayes approach* (see [7]). Following this approach, one treats the numbers $\pi_i = \mathbf{Q}(\Theta_i) = \mathsf{P}\{\theta \in \Theta_i\}$, $i = 1, 2$, as a priori probabilities of the hypotheses H_1 and H_2, respectively. Considered in [7] the so-called *partial Bayes approach* does not require that the probabilities π_1 and π_2 are known. Instead, the distributions of the parameter θ are known on both sets Θ_1 and Θ_2.

Along with the Bayes approach we consider the minimax approach under which one seeks a test minimizing the maximum of the conditional mean loss $\mathsf{E}_t^{\delta}A$.

DEFINITION 1.3.4. A test $\bar{\delta}$ is called *minimax* for the loss A if

$$\sup_{t\in\Theta} \mathsf{E}_t^{\bar{\delta}}A \le \sup_{t\in\Theta} \mathsf{E}_t^{\delta}A \tag{1.3.10}$$

for all tests δ where $\mathsf{E}_t^{\delta}A$ is the conditional mean loss defined by (1.3.4).

Generally speaking, the UMP tests do not exist in the class of all possible tests. Thus we consider proper subsets of tests and look for the UMP tests there. The following example exhibits this idea. The method below is suitable for finding both Bayes and minimax tests.

EXAMPLE 1.3.1. Let an observation ξ be a vector $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ whose components are independent and have normal $\mathcal{N}(\theta, \sigma_i^2)$, $i = 1, 2, \ldots, n$, distributions. Put $\Theta_1 = (-\infty, 0)$ and $\Theta_2 = [0, \infty)$. To distinguish the hypotheses $H_1: \theta \in \Theta_1$ and $H_2: \theta \in \Theta_2$ we consider the class of linear nonrandomized tests $\delta(r)$ of the form $\delta(x; r) = I((x, r) \ge 0)$ where $x = (x_1, x_2, \ldots, x_n)$, $r = (r_1, r_2, \ldots, r_n)$, and $(x, r) = \sum_{i=1}^n r_i x_i$; the vector r is such that $\sum_{i=1}^n r_i = 1$. Put $G(r; \theta) = \beta(\delta(r); \theta)$. It is clear that

$$G(r;\theta) = \mathsf{P}_\theta\{(\xi, r) \ge 0\} = \Phi\left(\theta\left(\sum_{i=1}^n \sigma_i^2 r_i^2\right)^{-1/2}\right)$$

where $\Phi(x)$ is the distribution function of the law $\mathcal{N}(0,1)$. Taking into account inequalities (1.3.2) and (1.3.3) we prove that $\delta(\widetilde{r})$ is a UMP test if

$$\Phi\left(\theta\Big(\sum_{i=1}^n \sigma_i^2 \widetilde{r}_i^2\Big)^{-1/2}\right) \le \Phi\left(\theta\Big(\sum_{i=1}^n \sigma_i^2 r_i^2\Big)^{-1/2}\right) \quad \text{for all } \theta < 0,$$

$$\Phi\left(\theta\Big(\sum_{i=1}^n \sigma_i^2 \widetilde{r}_i^2\Big)^{-1/2}\right) \ge \Phi\left(\theta\Big(\sum_{i=1}^n \sigma_i^2 r_i^2\Big)^{-1/2}\right) \quad \text{for all } \theta \ge 0$$

for all vectors $r = (r_1, r_2, \dots, r_n)$ where $\widetilde{r} = (\widetilde{r}_1, \widetilde{r}_2, \dots, \widetilde{r}_n)$. The latter inequalities are equivalent to

$$\sum_{i=1}^n \sigma_i^2 \widetilde{r}_i^2 \le \sum_{i=1}^n \sigma_i^2 r_i^2 \quad \text{for all } r = (r_1, r_2, \dots, r_n).$$

Thus in order to find a UMP test it is necessary to find a vector $\widetilde{r}$ for which the function $\sum_{i=1}^n \sigma_i^2 r_i^2$ assumes its minimal value on the set of vectors r such that $\sum_{i=1}^n r_i = 1$. It is clear that the components of the vector $\widetilde{r}$ with this property are such that

$$\widetilde{r}_i = \sigma_i^{-2}\Big(\sum_{j=1}^n \sigma_j^{-2}\Big)^{-1}, \qquad i = 1, 2, \dots, n.$$

The following result provides necessary and sufficient conditions for the existence of a UMP test.

THEOREM 1.3.2. *In order that a test is UMP for distinguishing the hypotheses $H_1\colon \theta \in \Theta_1$ and $H_2\colon \theta \in \Theta_2$ it is necessary and sufficient that it is an MP test for distinguishing two arbitrary simple hypotheses $H_1'\colon \theta = \theta_1$ and $H_2'\colon \theta = \theta_2$ where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$.*

PROOF. *Necessity.* Let δ^* be a UMP test and let δ be an arbitrary test. Then $\beta(\delta^*;\theta) \le \beta(\delta;\theta)$ for all $\theta \in \Theta_1$ and $\beta(\delta^*;\theta) \ge \beta(\delta;\theta)$ for all $\theta \in \Theta_2$. Let $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$. Consider two simple hypotheses $H_1'\colon \theta = \theta_1$ and $H_2'\colon \theta = \theta_2$. Then type I and type II error probabilities of the test are such that

$$\begin{aligned} \alpha(\delta^*) &= \mathsf{E}_{\theta_1}\delta^* = \beta(\delta^*;\theta_1) \le \beta(\delta;\theta_1) = \alpha(\delta), \\ \beta(\delta^*) &= \mathsf{E}_{\theta_2}(1-\delta^*) = 1 - \beta(\delta^*;\theta_2) \le 1 - \beta(\delta;\theta_2) = \beta(\delta) \end{aligned} \tag{1.3.11}$$

(see Section 1.1), that is, δ^* is an MP test.

Sufficiency. Let δ^* be an MP test for distinguishing the hypotheses $H_1'\colon \theta = \theta_1$ and $H_2'\colon \theta = \theta_2$ for all $\theta_i \in \Theta_i$, $i = 1, 2$. The Neyman–Pearson fundamental lemma (see Theorem 1.1.1) implies that $\beta(\delta^*) \le \beta(\delta)$ for all tests δ of level $\mathsf{E}_{\theta_1}\delta^* = \beta(\delta^*;\theta_1)$. It follows from (1.3.11) that $\beta(\delta;\theta_2) \le \beta(\delta^*;\theta_2)$ for all $\theta_2 \in \Theta_2$ and all tests δ such that $\beta(\delta;\theta_1) = \beta(\delta^*;\theta_1)$. Interchanging the hypotheses H_1' and H_2' and applying the Neyman–Pearson fundamental lemma once more we prove that $\beta(\delta^*;\theta_1) \le \beta(\delta;\theta_1)$ for all tests δ such that $\beta(\delta;\theta_2) = \beta(\delta^*;\theta_2)$ and all $\theta_1 \in \Theta_1$. □

The following example is a continuation of Example 1.3.1. It exhibits an application of Theorem 1.3.2 for finding a UMP test for another restriction of the class of tests as compared to that studied in Example 1.3.1.

EXAMPLE 1.3.2. Let $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ and random variables ξ_i be independent and distributed according to the $\mathcal{N}(\theta, \sigma_i^2)$, $i = 1, 2, \ldots, n$, laws. Consider the hypotheses $H_1: \theta \leq 0$ and $H_2: \theta > 0$. In order to construct an UMP test we use Theorem 1.3.2 and find an MP test for distinguishing the hypotheses $H_1': \theta = \theta_1$ and $H_2': \theta = \theta_2$ for all $\theta_1 \leq 0$ and $\theta_2 > 0$. The Neyman–Pearson fundamental lemma implies that this can be done by constructing a likelihood ratio test for distinguishing the hypotheses H_1' and H_2'. Let $p(x; \theta_i)$ be the density of the distribution of the vector ξ in the case of $\theta = \theta_i$. Then

$$\Lambda(x) = \ln \frac{p(x; \theta_2)}{p(x; \theta_1)} = (\theta_2 - \theta_1) \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} + \frac{\theta_2^2 - \theta_1^2}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \tag{1.3.12}$$

where $x = (x_1, x_2, \ldots, x_n)$. This means that $\delta(x) = I(\Lambda(x) > c)$ is the desired likelihood ratio test where c is some constant. Since $\theta_2 > \theta_1$, we obtain from (1.3.12) that this test is of the form

$$\delta(x) = I\left(\sum_{i=1}^{n} \sigma_i^{-2} x_i > k\right) \tag{1.3.13}$$

where k is some constant. If the constant k is such that $\mathsf{E}_0 \delta(\xi) = \alpha$ where $\alpha \in (0, 1)$ is a certain number, then the test δ given by (1.3.13) for some constant k does not depend on θ_1 and θ_2. Now Theorem 1.3.2 implies that the test defined by (1.3.13) is UMP in the class of all tests such that $\beta(\delta; 0) = \alpha$ where the constant k is specified above.

The latter example suggests a general idea on how to restrict the class of tests under consideration.

DEFINITION 1.3.5. The number

$$\alpha_1(\delta) = \sup_{\theta \in \Theta_1} \beta(\delta; \theta)$$

is called a *level* or *type* I *error probability* of the test δ.

The number $\alpha_1(\delta)$ is sometimes called the *size of the test* δ. This is the maximal probability of rejecting the hypothesis H_1 if it is true.

Consider the class of tests

$$K_\alpha = \{\delta: \alpha_1(\delta) \leq \alpha\}$$

where α is some number of the interval $[0, 1]$.

DEFINITION 1.3.6. A test δ^* is called a *uniformly most powerful* (*UMP*) *test in the class* K_α if

$$\beta(\delta^*; \theta) \geq \beta(\delta; \theta), \qquad \theta \in \Theta_2,$$

for an arbitrary test $\delta \in K_\alpha$.

Similar definitions in the class K_α can be introduced for Bayes and minimax tests [7].

UMP test for distributions with a monotone likelihood ratio. Let

$$\Theta = (-\infty, \infty), \qquad \Theta_1 = (-\infty, \theta_0], \quad \text{and} \quad \Theta_2 = (\theta_0, \infty)$$

where θ_0 is a fixed point of Θ. The hypotheses $H_1: \theta \in \Theta_1$ and $H_2: \theta \in \Theta_2$ are called *one-sided hypotheses* in contrast to the case of $H_2: \theta \neq \theta_0$ and $H_1: \theta = \theta_0$ where they are called *two-sided* hypotheses, since the sign of $\theta - \theta_0$ can be arbitrary in the latter case.

In what follows we assume that the measure P_θ for all $\theta \in \Theta$ is absolutely continuous with respect to some σ-finite measure μ and the density is

$$p(x; \theta) = dP_\theta / d\mu(x), \qquad x \in X.$$

Moreover we assume that for all $\theta_1 < \theta_2$ the likelihood ratio

$$z(x; \theta_2, \theta_1) = \frac{p(x; \theta_2)}{p(x; \theta_1)}, \qquad x \in X, \tag{1.3.14}$$

is a monotone (either nondecreasing or nonincreasing) function of some statistic $T(x)$. We say in this case that the family $\mathscr{P} = \{\mathsf{P}_\theta; \theta \in \Theta\}$ has a *monotone likelihood ratio.* For the sake of definiteness we assume that the likelihood ratio (1.3.14) is a nondecreasing function of $T(x)$. The case where $z(x; \theta_2, \theta_1)$ is a nonincreasing function of $T(x)$ can be considered analogously.

THEOREM 1.3.3. *Let $\theta \in \Theta = (-\infty, \infty)$ be a one-dimensional parameter. Assume that a family $\mathscr{P} = \{\mathsf{P}_\theta; \theta \in \Theta\}$ has a monotone likelihood ratio $z(x; \theta_2, \theta_1)$. Then*

1. *In the class of tests K_α, $\alpha \in (0, 1)$, there exists a UMP test for distinguishing the hypotheses $H_1: \theta \in \Theta_1 = (-\infty, \theta_0]$ and $H_2: \theta \in \Theta_2 = (\theta_0, \infty)$. The test is given by*

$$\delta^*(x) = I(T(x) > c) + qI(T(x) = c) \tag{1.3.15}$$

where $c \in (-\infty; \infty)$ and $q \in [0, 1]$ are the parameters defined by

$$\mathsf{E}_{\theta_0} \delta^* = \mathsf{P}_\theta(T(\xi) > c) + q\mathsf{P}_{\theta_0}(T(\xi) = c) = \alpha. \tag{1.3.16}$$

2. *The power function $\beta(\delta^*; \theta)$ of the test δ^* defined by equalities* (1.3.15) *and* (1.3.16) *is a nondecreasing function of $\theta \in \Theta$.*
3. *For all $\theta' \in \Theta$ the test* (1.3.15) *is a UMP test in the class $K_{\beta(\delta^*; \theta')}$ for distinguishing the hypotheses $H_1': \theta \leq \theta'$ and $H_2': \theta > \theta'$.*
4. *For every $\theta < \theta_0$ the test δ^* defined by equalities* (1.3.15) *and* (1.3.16) *minimizes the function $\beta(\delta; \theta)$ in the class K_α.*

PROOF. Consider the two simple hypotheses $\widetilde{H}_1: \theta = \theta_0$ and $\widetilde{H}_2: \theta = \theta_2$ where $\theta_2 > \theta_0$. According to the Neyman–Pearson fundamental lemma (Theorem 1.1.1), a most powerful test for distinguishing the hypotheses $\widetilde{H}_1$ and $\widetilde{H}_2$ in the class of tests δ such that $\mathsf{E}_{\theta_0} \delta = \alpha$ is of the form (1.3.15), since the inequality $z(x; \theta_2, \theta_0) > c$ is equivalent to the inequality $T(x) > c$ in view of the monotonicity of the likelihood ratio where the constants c and q are defined by (1.3.16). Since the parameters c and q do not depend on θ_2, the test δ^* is the most powerful for distinguishing the hypotheses $\widetilde{H}_1: \theta = \theta_0$ and $\widetilde{H}_2: \theta = \theta_2$ for all $\theta_2 \in \Theta_2$. Thus Theorem 1.3.2 implies that the test δ^* maximizes $\beta(\delta; \theta)$ for all $\theta \in \Theta_2$ in the class of tests δ such that $\beta(\delta; \theta_0) = \alpha$.

Now let θ' and θ'' be two arbitrary points such that $\theta' < \theta''$. Again by the Neyman–Pearson fundamental lemma the test δ^* is the most powerful for distinguishing the simple hypotheses $\widetilde{H}_1'\colon \theta = \theta_1'$ and $\widetilde{H}_2'\colon \theta = \theta''$ in the class of tests of level $\alpha' = \beta(\delta^*; \theta')$. By the definition of the set $\mathfrak{N}$ (see Section 1.1) $\alpha(\delta^*) \le 1 - \beta(\delta^*)$ for the most powerful test δ^*. Thus

$$\alpha' = \beta(\delta^*; \theta') = \alpha(\delta^*) \le 1 - \beta(\delta^*) = \beta(\delta^*; \theta'').$$

Therefore $\beta(\delta^*; \theta') \le \beta(\delta^*; \theta'')$ for all $\theta' < \theta''$ and statement 2 of the theorem is proved.

Since the function $\beta(\delta^*; \theta)$ is nondecreasing, the test δ^* is such that $\beta(\delta^*; \theta) \le \alpha$ for all $\theta \le \theta_0$, that is, the test δ^* belongs to the class K_α. In its turn K_α belongs to a wider class $\{\delta\colon \mathsf{E}_{\theta_0}\delta = \alpha\}$. Since δ^* is a UMP test in the class $\{\delta\colon \mathsf{E}_{\theta_0}\delta = \alpha\}$, it also is a UMP test in the class K_α.

Statement 3 of the theorem can be proved in the same manner.

Statement 4 follows from statements 1–3 applied to the problem of distinguishing the hypotheses $H_1^0\colon \theta \ge \theta_0$ and $H_1^0\colon \theta < \theta_0$. A UMP test in the class $\{\delta\colon \sup_{\theta \ge \theta_0} \mathsf{E}_\theta\delta \le 1 - \alpha\}$ is $\delta^0(x) = 1 - \delta(x)$ for this problem and the power function $1 - \beta(\delta^*; \theta) = \mathsf{E}_\theta\delta^0$ is maximal for $\theta < \theta_0$. Therefore the test δ^* minimizes $\beta(\delta; \theta)$ for $\theta < \theta_0$ and for tests δ of the class K_α. □

REMARK 1.3.3. Equality (1.3.12) shows that the likelihood ratio in Example 1.3.2 is a monotone function of the statistic

$$T(x) = \sum_{i=1}^{n} \sigma_i^{-2} x_i, \quad x = (x_1, x_2, \ldots, x_n).$$

Thus Theorem 1.3.3 is applicable in this case and a UMP test exists for distinguishing the one-sided hypotheses.

EXAMPLE 1.3.3. Let $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables whose distribution depends on a parameter $\theta \in (0, 1)$ such that $\mathsf{P}_\theta\{\xi_i = 1\} = \theta$ and $\mathsf{P}_\theta\{\xi_i = 0\} = 1 - \theta$. The space X of possible values of the random vector ξ consists of the vectors $x = (x_1, x_2, \ldots, x_n)$ whose coordinates x_i are either 0 or 1. The distribution of the vector ξ is given by

$$\mathsf{P}_\theta(x) = \mathsf{P}_\theta\{\xi = x\} = \theta^{\sum_{i=1}^n x_i}(1 - \theta)^{n - \sum_{i=1}^n x_i}, \qquad x = (x_1, \ldots, x_n),$$

whence we obtain the likelihood ratio:

$$z(x; \theta_2, \theta_1) = \left(\frac{\theta_2}{\theta_1}\right)^{\sum_{i=1}^n x_i} \left(\frac{1 - \theta_2}{1 - \theta_1}\right)^{n - \sum_{i=1}^n x_i} = \left(\frac{\theta_2(1 - \theta_1)}{\theta_1(1 - \theta_2)}\right)^{\sum_{i=1}^n x_i} \left(\frac{1 - \theta_2}{1 - \theta_1}\right)^n.$$

We see that the likelihood ratio $z(x; \theta_2, \theta_1)$ for $\theta_1 < \theta_2$ is an increasing function of the statistic $T(x) = \sum_{i=1}^n x_i$. According to Theorem 1.3.3 there exists a UMP test for distinguishing one-sided hypotheses in the class K_α.

An important class of distributions for which the likelihood ratio is monotone is presented by the *one parameter exponential family*. The density $p(x; \theta)$ in this case is given by

$$p(x; \theta) = h(x) \exp\{a(\theta)T(x) + V(\theta)\}, \qquad x \in X, \tag{1.3.17}$$

where all the functions on the right-hand side are finite and measurable (see [38], Section 1.2). In view of the factorization criterion (Theorem 4.1.2 in [38]), the statistic $T(x)$ is sufficient. The likelihood ratio in this case is given by

$$z(x;\theta_2,\theta_1) = \exp\{(a(\theta_2) - a(\theta_1))T(x) + (V(\theta_2) - V(\theta_1))\}.$$

This implies that if $a(\theta_2) - a(\theta_1)$ does not change its sign for all $\theta_1 < \theta_2$, then the likelihood ratio $z(x;\theta_2,\theta_1)$ is a monotone function of the statistic $T(x)$.

Theorem 1.3.3 implies the following assertion.

COROLLARY 1.3.2. *Let the density $p(x,\theta)$ be of the form* (1.3.17) *where $a(\theta)$ is a monotone function. Then there exists a UMP test δ^* in the class K_α for distinguishing the hypotheses $H_1\colon \theta \le \theta_0$ and $H_2\colon \theta > \theta_0$. If the function $a(\theta)$ increases, then the test δ^* is defined by* (1.3.15) *and* (1.3.16). *If the function $a(\theta)$ decreases, then the test δ^* is defined by* (1.3.15) *and* (1.3.16) *where $T(x) \le c$ and $T(\xi) \le c$ substitute $T(x) > c$ and $T(\xi) > c$, respectively.*

REMARK 1.3.4. If we distinguish the hypothesis $H_1\colon \theta = \theta_0$ and its two-sided alternative $H_2\colon \theta \ne \theta_0$, then a UMP test does not exist in the case of exponential distributions (1.3.17). Indeed, for simplicity let the function $a(\theta)$ increase and let the P_θ-distribution of $T(\xi)$ for all θ possess the density. Then by the Neyman–Pearson fundamental lemma a most powerful test for distinguishing the hypotheses

$$H_1\colon \theta = \theta_0 \quad \text{and} \quad H_2'\colon \theta = \theta_2$$

with $\theta_2 > \theta_0$ is nonrandomized and moreover $\delta^*(x) = I(T(x) \ge c)$. On the other hand, if $\theta_2 < \theta_0$, then the most powerful test is $\delta'(x) = I(T(x) < c)$. Thus there is no unique UMP test for all $\theta_2 > \theta_0$ and $\theta_2 < \theta_0$. In a similar manner we get that there is no UMP test for distinguishing the hypotheses $H_1\colon \theta \in (\theta_1,\theta_2)$ and $H_2\colon \theta \notin (\theta_1,\theta_2)$ where $\theta_1 < \theta_2$. However if $H_1\colon \theta \notin (\theta_1,\theta_2)$ is the null hypothesis and $H_2\colon \theta \in (\theta_1,\theta_2)$ is its alternative, then a UMP test exists. This case is studied in the next section.

Two-sided null hypotheses. Exponential families of distributions. Let a distribution P_θ be absolutely continuous with respect to some σ-finite measure μ and let the density $p(x;\theta) = d\mathsf{P}_\theta/d\mu(x)$ be of the form (1.3.17).

THEOREM 1.3.4. *Let $p(x;\theta)$ be of the form* (1.3.17) *where the function $a(\theta)$ is monotone. Let $H_1\colon \theta \notin (\theta_1,\theta_2)$ be the null hypothesis and let $H_2\colon \theta \in (\theta_1,\theta_2)$ be its alternative where $\theta_1 < \theta_2$ are two fixed numbers. Then*

1) *in the class*

$$K_\alpha = \left\{\delta\colon \sup_{\theta \notin (\theta_1,\theta_2)} \mathsf{E}_\theta \delta(\xi) \le \alpha\right\}$$

there exists a UMP test δ^* such that

$$\delta^*(x) = I(c_1 < T(x) < c_2) + q_1 I(T(x) = c_1) + q_2 I(T(x) = c_2) \tag{1.3.18}$$

where c_1, c_2, q_1, and q_2 are the constants defined by

$$\mathsf{E}_{\theta_1}\delta^*(\xi) = \mathsf{E}_{\theta_2}\delta^*(\xi) = \alpha; \tag{1.3.19}$$

2) *the test δ^* defined by equalities* (1.3.18) *and* (1.3.19) *maximizes the power function $\beta(\delta;\theta)$ inside the interval (θ_1,θ_2) and minimizes it outside this interval*;

3) *for $0 < \alpha < 1$ the power function $\beta(\delta^*;\theta)$ attains its maximum at some point $\theta_0 \in (\theta_1, \theta_2)$; moreover it strictly decreases in both cases if the argument goes to the left of θ_0 or if it goes to the right of θ_0. Note that the case where the distribution of $T(\xi)$ is concentrated at two points is not excluded, that is, the case where there are t_1 and t_2 such that for all θ*

$$\mathsf{P}_\theta\{T(\xi) = t_1\} + \mathsf{P}_\theta\{T(\xi) = t_2\} = 1.$$

We omit the proof of Theorem 1.3.4 that can be found in [**7, 9**], or [**34**].

The generalized Neyman–Pearson fundamental lemma. The construction of MP and UMP tests requires, in fact, the solution of a variational problem and finding a maximum of a certain functional of the test satisfying some restrictions. In particular, we deal with the test δ^* in Theorem 1.3.4 for which we maximize the functional

$$\int_X \delta(x)p(x;\theta)\,\mu(dx)$$

in the class of tests δ such that

$$\int_X \delta(x)p(x;\theta_i)\,\mu(dx) = \alpha, \qquad i = 1,2.$$

The following result is sometimes called the *generalized Neyman–Pearson fundamental lemma.*

THEOREM 1.3.5. *Let $f_1, f_2, \dots, f_{m+1}$ be real Borel functions defined on $(X, \mathscr{B})$ that are integrable with respect to a measure μ. Consider the tests δ such that*

$$\int_X \delta(x)f_i(x)\,\mu(dx) = \alpha_i, \qquad i = 1,2,\dots,m, \tag{1.3.20}$$

where $\alpha_1, \alpha_2, \dots, \alpha_m$ are some numbers. Then the test $\delta^*(x)$ that maximizes the functional $\int_X \delta(x)f_{m+1}(x)\,\mu(dx)$ is of the form

$$\delta^*(x) = \begin{cases} 1, & \text{if } f_{m+1}(x) > \sum_{i=1}^n k_i f_i(x), \\ 0, & \text{if } f_{m+1}(x) < \sum_{i=1}^n k_i f_i(x), \end{cases}$$

where the constants $k_1, \dots, k_m$ are defined by conditions (1.3.20).

PROOF. Put

$$F_i(\delta) = \int_X \delta(x)f_i(x)\,\mu(dx), \qquad i = 1,2,\dots,m+1.$$

A test δ such that

$$F_i(\delta) = \alpha_i, \qquad i = 1,2,\dots,m,$$

maximizes $F_{m+1}(\delta)$ if and only if it maximizes $F_{m+1}(\delta) - \sum_{i=1}^m k_i F_i(\delta)$ for some constants $k_1, k_2, \dots, k_m$ ($\sum_{i=1}^m k_i F_i(\delta)$ is fixed in this expression). This is the case if the test $\delta(x)$ maximizes the functional

$$\int_X \left(f_{m+1}(x) - \sum_{i=1}^m k_i f_i(x)\right)\delta(x)\,\mu(dx).$$

The latter expression is maximal for the test δ such that $\delta(x) = 1$ if

$$f_{m+1}(x) - \sum_{i=1}^{m} k_i f_i(x) > 0$$

and $\delta(x) = 0$ otherwise. The constants $k_1, k_2, \ldots, k_m$ occurring in the definition of the test δ, as well as the values of $\delta(x)$ on the set

$$\left\{ f_{m+1}(x) = \sum_{i=1}^{m} k_i f_i(x) \right\},$$

should be chosen to satisfy conditions (1.3.20). □

Unbiased tests. Another restricted class we use to construct UMP tests consists of the so-called unbiased tests.

Consider the general problem of distinguishing the hypotheses $H_1: \theta \in \Theta_1$ and $H_2: \theta \in \Theta_2$ where $\Theta_1 \cap \Theta_2 = \varnothing$ and $\Theta_1 \cup \Theta_2 = \Theta$. Let δ be a test of the class

$$K_\alpha = \left\{ \delta: \alpha_1(\delta) = \sup_{\theta \in \Theta_1} \mathsf{E}_\theta \delta \leq \alpha \right\}.$$

If Θ_1 contains only a single point θ_1 and $\mathsf{E}_{\theta_1}\delta = \alpha$, then α is the probability of rejecting the hypothesis H_1 if it is true. It is natural to require that a test δ is such that the probability of rejecting the hypothesis H_1, if it is wrong, is bigger than α, that is, $\beta(\delta;\theta) \geq \alpha$ for all $\theta \in \Theta_2$. If this is not the case, then there are alternative hypotheses $\theta \in \Theta_2$ such that the probability of accepting the hypothesis H_1 is bigger than $1 - \alpha = 1 - \mathsf{E}_{\theta_1}\delta$ and the latter is the probability of accepting the hypothesis H_1 if it is true. It is reasonable to exclude such cases from our consideration.

DEFINITION 1.3.7. A test δ is called *unbiased* if

$$\inf_{\theta \in \Theta_2} \beta(\delta;\theta) \geq \sup_{\theta \in \Theta_1} \beta(\delta;\theta). \tag{1.3.21}$$

Condition (1.3.21) implies that a test $\delta \in K_\alpha$ of level $\alpha_1(\delta) = \alpha$ is unbiased if $\beta(\delta;\theta) \geq \alpha$ for all $\theta \in \Theta_2$. The class of unbiased tests of level α is denoted by $\widehat{K}_\alpha$. By $\partial\Theta_i$ we denote the boundary of the set Θ_i, that is, all the limit points of the set Θ_i.

LEMMA 1.3.1. *Let* $\Gamma = \partial\Theta_1 \cap \partial\Theta_2 \neq \varnothing$. *Assume that the density* $p(x;\theta)$ *is continuous in* θ *for* μ*-almost all* $x \in X$. *Then*

$$\beta(\delta;\theta) = \alpha \quad \textit{for all } \theta \in \Gamma \tag{1.3.22}$$

for any test $\delta \in \widehat{K}_\alpha$.

PROOF. Since

$$\beta(\delta;\theta) = \int_X \delta(x) p(x;\theta)\, \mu(dx), \qquad 0 \leq \delta(x) \leq 1,$$

and the function $p(x;\theta)$ is continuous in θ by Corollary 3.4.1 in [**38**], the power function $\beta(\delta;\theta)$ is also continuous for any test δ. This implies equality (1.3.22) for any test $\delta \in \widehat{K}_\alpha$. □

We denote by $\overline{K}_\alpha$ the class of tests δ satisfying condition (1.3.22).

LEMMA 1.3.2. *Let* $\Gamma = \partial\Theta_1 \cap \partial\Theta_2 \neq \varnothing$ *and* $\widehat{K}_\alpha \subset \overline{K}_\alpha$. *Then any UMP test in the class* $\overline{K}_\alpha \cap K_\alpha$ *is UMP in the class* $\widehat{K}_\alpha$, *too.*

PROOF. Let δ' be a UMP test in the class $\overline{K}_\alpha \cap K_\alpha$. It is sufficient to prove that $\delta' \in \widehat{K}_\alpha$ and $\widehat{K}_\alpha \subset \overline{K}_\alpha \cap K_\alpha$. The inclusion $\widehat{K}_\alpha \subset \overline{K}_\alpha \cap K_\alpha$ follows from $\widehat{K}_\alpha \subset \overline{K}_\alpha$ and $\widehat{K}_\alpha \subset K_\alpha$. Since δ' is a UMP test in the class $\overline{K}_\alpha \cap K_\alpha$ and the test $\delta \equiv \alpha$ belongs to $\overline{K}_\alpha \cap K_\alpha$, we have $\beta(\delta';\theta) \geq \beta(\delta;\theta) = \alpha$ for all $\theta \in \Theta_2$, whence

$$\inf_{\theta\in\Theta_2} \beta(\delta';\theta) \geq \inf_{\theta\in\Theta_2} \beta(\delta;\theta) = \alpha.$$

Therefore $\delta' \in \widehat{K}_\alpha$. □

Lemma 1.3.2 implies that the problem of finding an unbiased UMP test can be reduced to the problem of finding a usual UMP test satisfying condition (1.3.22). If the number of points of the boundary Γ is finite, then the conditions of Theorem 1.3.5 hold and one needs to prove that an optimal test δ^* does not depend on the point $\theta \in \Theta_2$ at which the test maximizes the functional $\beta(\delta;\theta) = \mathsf{E}_\theta\delta$. This means that δ^* is a UMP test.

Consider condition (1.3.22) for the following case. Let

$$\Theta = (-\infty,\infty), \qquad \Theta_1 = [\theta_1,\theta_2], \quad \text{and} \quad \Theta_2 = [\theta_1,\theta_2]^c = (-\infty,\infty) \setminus [\theta_1,\theta_2].$$

If $\theta_1 < \theta_2$, then the common boundary Γ of the sets Θ_1 and Θ_2 contains only two points θ_1 and θ_2. Therefore condition (1.3.22) becomes of the form $\beta(\delta;\theta_i) = \alpha$, $i = 1,2$. If $\Theta_1 = \{\theta_1\}$, then condition (1.3.22) is equivalent to $\beta(\delta;\theta_1) = \alpha$.

The following result gives a necessary condition for a test of level α to be unbiased.

LEMMA 1.3.3. *Let* δ *be a test of level* α *for distinguishing the hypotheses*

$$H_1\colon \theta = \theta_1 \quad \textit{and} \quad H_2\colon \theta \neq \theta_1.$$

Assume that the regularity conditions (R) *hold for the density* $p(x;\theta)$ *of the measure* P_θ *with respect to a* σ*-finite measure* μ. *If the test* δ *is unbiased, then*

$$\mathsf{E}_{\theta_1}\delta(\xi)S(\xi;\theta_1) = 0 \tag{1.3.23}$$

where $S(x;\theta) = \frac{\partial}{\partial\theta}\ln p(x;\theta)$.

PROOF. Since the test δ is unbiased, the power function $\beta(\delta;\theta)$ attains its minimum at the point θ_1. Since the function $p(x;\theta)$ satisfies the regularity conditions (R), the power function $\beta(\delta;\theta)$ is differentiable by Lemma 3.4.4 of [**38**]. Therefore the equality $\beta'(\delta;\theta_1) = 0$ is satisfied. Applying again Lemma 3.4.4 of [**38**], we obtain

$$\beta'(\delta;\theta_1) = \int_X \delta(x)p'_\theta(x;\theta_1)\,\mu(dx) = \int_X \delta(x)S(x;\theta_1)p(x;\theta_1)\,\mu(dx).$$

This together with the equality $\beta'(\delta;\theta_1) = 0$ implies (1.3.23). □

REMARK 1.3.5. The regularity conditions (R) can be found in [**38**] (also see [**7**] or [**9**]). Lemma 1.3.3 implies that an unbiased test of level α for distinguishing the hypotheses $H_1\colon\theta = \theta_1$ and $H_2\colon\theta \neq \theta_1$ is a solution of the following two equations:

$$\mathsf{E}_{\theta_1}\delta(\xi) = \alpha, \qquad \mathsf{E}_{\theta_1}\delta(\xi)S(\xi;\theta_1) = 0. \tag{1.3.24}$$

EXAMPLE 1.3.4. Let the distribution of ξ be exponential with a density of the form (1.3.17) where the functions $a(\theta)$ and $V(\theta)$ are differentiable. Then

$$S(x;\theta) = a'(\theta)T(x) + V'(\theta).$$

Since $\mathsf{E}_\theta S(\xi;\theta) = 0$, we get $V'(\theta) = -a'(\theta)\mathsf{E}_\theta T(\xi)$. Thus

$$\mathsf{E}_\theta \delta(\xi) S(\xi;\theta) = a'(\theta)\mathsf{E}_\theta \delta(\xi) T(\xi) - a'(\theta)\mathsf{E}_\theta \delta(\xi)\mathsf{E}_\theta T(\xi).$$

This implies that equations (1.3.24) become of the form

$$\mathsf{E}_{\theta_1}(\delta(\xi) - \alpha) = 0, \qquad \mathsf{E}_{\theta_1}(\delta(\xi) - \alpha)T(\xi) = 0.$$

The following result describes the form of an unbiased UMP test for distinguishing a null hypothesis and its two-sided alternative hypothesis for exponential families.

THEOREM 1.3.6. *Let $p(x;\theta)$ be of the form* (1.3.17) *where the function $a(\theta)$ is monotone. Assume that the problem is to distinguish the hypotheses $H_1\colon \theta \in [\theta_1, \theta_2]$ and $H_2\colon \theta \notin [\theta_1, \theta_2]$ where $\theta_1 \le \theta_2$. Then*

1. *In the class of tests $\widehat{K}_\alpha$ there exists a UMP test $\hat{\delta}$ such that*

$$\hat{\delta}(x) = I\big(T(x) \notin [c_1, c_2]\big) + q_1 I\big(T(x) = c_1\big) + q_2 I\big(T(x) = c_2\big) \tag{1.3.25}$$

where the constants c_i and q_i, $i = 1, 2$, are defined by

$$\mathsf{E}_{\theta_i}\hat{\delta}(\xi) = \alpha, \qquad i = 1, 2, \tag{1.3.26}$$

if $\theta_1 < \theta_2$, and by

$$\mathsf{E}_{\theta_1}\hat{\delta}(\xi) = \alpha, \qquad \mathsf{E}_{\theta_1}(\hat{\delta}(\xi) - \alpha)T(\xi) = 0 \tag{1.3.27}$$

if $\theta_1 = \theta_2$.

2. *The test $\hat{\delta}$ defined by* (1.3.25)–(1.3.27) *minimizes the power function $\beta(\delta;\theta)$ inside the interval $[\theta_1, \theta_2]$ if* (1.3.26) *holds and maximizes it outside the interval $[\theta_1, \theta_2]$ if* (1.3.26) *holds and $\theta_1 < \theta_2$ or* (1.3.27) *holds and $\theta_1 = \theta_2$.*
3. *If $0 < \alpha < 1$ and $\theta_1 < \theta_2$, then the function $\beta(\hat{\delta};\theta)$ attains its minimum at some point $\theta_0 \in (\theta_1, \theta_2)$; moreover it strictly decreases in both cases: either the argument goes to the left of θ_0 or it goes to the right of θ_0. Note that the case where the distribution of $T(\xi)$ is concentrated at two points is not excluded, that is, we do not exclude the case where there are t_1 and t_2 such that for all θ*

$$\mathsf{P}_\theta\{T(\xi) = t_1\} + \mathsf{P}_\theta\{T(\xi) = t_2\} = 1.$$

Theorem 1.3.6 is similar to Theorem 1.3.4; however the words "minimizes" and "maximizes" are interchanged and the case of $\theta_1 = \theta_2$ is not excluded in Theorem 1.3.6. The proof of Theorem 1.3.6 is omitted (it can be found in [**7**, **9**], or [**34**]).

EXAMPLE 1.3.5. Let $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ where $\xi_1, \xi_2, \dots, \xi_n$ are independent identically distributed random variables with the $\mathcal{N}(0, \sigma^2)$ distribution. Consider the hypotheses $H_1: \sigma = \sigma_0$ and $H_2: \sigma \neq \sigma_0$ where $\sigma_0 \in (0, \infty)$ is a fixed number. The density of the distribution belongs to the family of densities

$$(2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\right), \qquad \sigma \in (0, \infty),\ x = (x_1, \dots, x_n).$$

The statistic $T(x) = \sum_{i=1}^{n} x_i$ is essential for this family and the density of $T(\xi)$ is $\sigma^{-2} f_n(y/\sigma^2)$ where

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{(n/2)-1} e^{-y/2}, \qquad y > 0,$$

is the $\chi^2(n)$ density with n degrees of freedom.

An unbiased UMP test $\delta^*(x)$ of level α is of the form (1.3.25) where the constants c_1 and c_2 satisfy condition (1.3.27), while the constants q_1 and q_2 are arbitrary, since the distribution of the statistic $T(\xi)$ is continuous. Putting $q_1 = q_2 = 1$ one can represent the test $\delta^*(x)$ as

$$\delta^*(x) = I\left(\frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_i^2 \notin (C_1, C_2)\right)$$

where $C_i = c_i/\sigma_0^2, i = 1, 2$. Then condition (1.3.27) becomes of the form

$$\int_{C_1}^{C_2} f_n(y)\, dy = 1 - \alpha, \qquad \int_{C_1}^{C_2} y f_n(y)\, dy = n(1 - \alpha). \tag{1.3.28}$$

To determine the constants c_1 and c_2 one can use, for example, the tables of the $\chi^2(n)$ distribution. Using the equality $y f_n(y) = n f_{n+2}(y)$ the second equation in (1.3.28) can be rewritten as

$$\int_{C_1}^{C_2} f_{n+2}(y)\, dy = 1 - \alpha.$$

Another way to determine C_1 and C_2 is to integrate by parts the second equation in (1.3.28) and obtain

$$C_1^{n/2} e^{-C_1/2} = C_2^{n/2} e^{-C_2/2}.$$

A relationship between tests and confidence sets. Let ξ be an observation whose distribution belongs to a family $\{\mathsf{P}_\theta; \theta \in \Theta\}$.

DEFINITION 1.3.8. A random set $\Theta^*(\xi, \gamma)$ is called a *confidence set of level* γ if $\Theta^*(\xi, \gamma) \subseteq \Theta$ and

$$\mathsf{P}_\theta\{\theta \in \Theta^*(\xi, \gamma)\} \geq \gamma \tag{1.3.29}$$

for all $\theta \in \Theta$.

Put

$$X(\theta, \gamma) = \{x \in X: \theta \in \Theta^*(x, \gamma)\}. \tag{1.3.30}$$

Then the inclusions

$$\theta \in \Theta^*(x,\gamma) \quad \text{and} \quad x \in X(\theta,\gamma) \tag{1.3.31}$$

are equivalent.

In the definition of a confidence set we assume that the set $X(\theta,\gamma)$ in (1.3.30) is measurable, thus the probability in (1.3.29) is well defined. In view of the equivalence of inclusions (1.3.31), the latter probability is equal to

$$\mathsf{P}_\theta\{\theta \in \Theta^*(\xi,\gamma)\} = \mathsf{P}_\theta\{\xi \in X(\theta,\gamma)\}. \tag{1.3.32}$$

The following result describes a relationship between confidence sets and statistical tests for distinguishing the hypotheses $H(\theta_0)\colon \theta = \theta_0$ and

$$K(\theta_0)\colon \theta \in \Theta \setminus \{\theta_0\} = \Theta(\theta_0).$$

THEOREM 1.3.7. 1) *For every θ_0 let a nonrandomized test $\delta(\theta_0)$ of level $1-\gamma$ be given for distinguishing the hypotheses $H(\theta_0)$ and $K(\theta_0)$. Let $X(\theta_0,\gamma)$ be the acceptance set for the hypothesis $H(\theta_0)$ defined by* (1.3.30). *Then*

$$\Theta^*(\xi;\gamma) = \{\theta \in \Theta\colon \xi \in X(\theta_0,\gamma)\}$$

is a confidence set of level γ.

Conversely, if $\Theta^*(\xi;\gamma)$ is a confidence set of level γ and $\theta_0 \in \Theta^*(\xi;\gamma)$, then the acceptance set $X(\theta_0,\gamma)$ defined by (1.3.30) *for the hypothesis $H(\theta_0)$ determines a test for distinguishing the hypotheses $H(\theta_0)$ and $K(\theta_0)$.*

2) *Let $X(\theta_0,\gamma)$ be the set defined by* (1.3.30) *for the hypothesis $H(\theta_0)$. If $\delta(\theta_0)$ is a UMP test of level $1-\gamma$ for all θ_0, then the corresponding set $\Theta^*(\xi,\gamma)$ minimizes the probability*

$$\mathsf{P}_\theta\{\theta' \in \Theta^*(\xi,\gamma)\} \quad \textit{for all } \theta \textit{ and } \theta' \textit{ such that } \theta \in \Theta(\theta') \tag{1.3.33}$$

in the class of all confidence sets of level γ.

Conversely, the minimal probability in (1.3.33) *corresponds to a set $X(\theta,\gamma)$ that generates a UMP test.*

PROOF. Equality (1.3.32) yields

$$\mathsf{P}_\theta\{\theta \in \Theta^*(\xi,\gamma)\} = \mathsf{P}_\theta\{\xi \in X(\theta,\gamma)\} \geq \gamma,$$

whence the first statement of the theorem follows. To prove the second statement of the theorem we consider another confidence set $\widetilde{\Theta}^*(\xi,\gamma)$ and the corresponding subset $\widetilde{X}(\theta,\gamma)$ in X. Then

$$\mathsf{P}_\theta\{\xi \in \widetilde{X}(\theta,\gamma)\} = \mathsf{P}_\theta\{\theta \in \widetilde{\Theta}^*(\theta,\gamma)\} \geq \gamma.$$

Since $X(\theta_0,\gamma)$ is the acceptance set for a UMP test,

$$\mathsf{P}_\theta\{\xi \in \widetilde{X}(\theta_0,\gamma)\} \geq \mathsf{P}_\theta\{\xi \in X(\theta_0,\gamma)\}$$

for all $\theta \in \Theta(\theta_0)$. Thus

$$\mathsf{P}_\theta\{\theta_0 \in \widetilde{\Theta}^*(\xi,\gamma)\} \geq \mathsf{P}_\theta\{\theta_0 \in \Theta^*(\xi,\gamma)\}$$

for all $\theta \in \Theta(\theta_0)$. □

DEFINITION 1.3.9. Confidence sets for which the probability in (1.3.33) is minimal given (1.3.29) are called *uniformly most precise confidence sets* of level γ with respect to alternatives θ' such that $\theta \in \Theta(\theta')$.

Consider in more detail the notion introduced above for a particular case of a one-dimensional parameter. The following result holds in this case.

THEOREM 1.3.8. *Let the set $X(\theta, \gamma)$ for a UMP test described in Theorem* 1.3.7 *be of the form*

$$c_1(\theta, \gamma) \le T(x) \le c_2(\theta, \gamma)$$

where $c_i(\theta, y)$ are monotone and continuous in θ. If the functions $c_i(\theta, y)$ increase in θ, then a uniformly most precise confidence set of level γ with respect to alternatives θ' such that $\theta \in \Theta(\theta')$ is the interval

$$c_2^{-1}(T(x), \gamma) \le \theta \le c_1^{-1}(T(x), \gamma)$$

where $c_i^{-1}(t, y)$ is a solution in θ of the equation $c_i(\theta, y) = t$.

The proof of Theorem 1.3.8 is obvious and omitted.

Consider in more detail *one-sided confidence intervals for a one-dimensional parameter* θ, namely we consider the intervals $(\underline{\theta}(\xi, \gamma), \infty)$ and $(-\infty, \overline{\theta}(\xi, \gamma))$. We restrict our consideration to the case of a *lower confidence bound* $\underline{\theta}(\xi, \gamma)$ for which

$$\mathsf{P}_\theta\{\underline{\theta}(\xi, \gamma) \le \theta\} \ge \gamma, \tag{1.3.34}$$

since an *upper confidence bound* $\overline{\theta}(\xi, \gamma)$ is considered similarly.

DEFINITION 1.3.10. A bound $\underline{\theta} = \underline{\theta}(\xi, \gamma)$ for which the probability $\mathsf{P}_\theta\{\underline{\theta} \le \theta'\}$ is minimal for all $\theta' < \theta$ is called a *uniformly most precise lower confidence bound of level* γ.

Below we consider another definition of an optimal confidence interval. Let $L(\theta, \underline{\theta})$ be a loss arising if θ is underestimated, so that for all fixed θ the function $L(\theta, \underline{\theta})$ is defined as follows: $L(\theta, \underline{\theta}) = 0$ for $\underline{\theta} \ge \theta$ and $L(\theta, \underline{\theta}) \ge 0$ for $\underline{\theta} < \theta$. We also assume that $L(\theta, \underline{\theta})$ is continuously increasing if $\underline{\theta}$ goes away from θ and that $\mathsf{E}_\theta L(\theta, \underline{\theta}) < \infty$ for all θ. Given (1.3.34) our goal is to minimize $\mathsf{E}_\theta L(\theta, \underline{\theta})$.

The following auxiliary result establishes a relationship between two notions of optimality.

LEMMA 1.3.4. *Given* (1.3.34) *a uniformly most precise lower bound $\underline{\theta}$ minimizes $\mathsf{E}_\theta L(\theta, \underline{\theta})$ for any loss function $L(\theta, \underline{\theta})$ satisfying the above conditions.*

PROOF. Let $\underline{\theta}'$ be an arbitrary lower bound. Since the increments of the loss function $L(\theta, u)$ in u in the domain $u < \theta$ are negative, we get

$$\mathsf{E}_\theta L(\theta, \underline{\theta}) = \int_{-\infty}^{\theta} L(\theta, u)\, d_u \mathsf{P}_\theta(\underline{\theta} < u) = -\int_{-\infty}^{\theta} \mathsf{P}_\theta(\underline{\theta} < u)\, d_u L(\theta, u)$$

$$\le -\int_{-\infty}^{\theta} \mathsf{P}_\theta(\underline{\theta}' < u)\, d_u L(\theta, u) = \mathsf{E}_\theta L(\theta, \underline{\theta}')$$

where d_u is the differential with respect to the variable u. □

It is natural to call the number $\mathsf{E}_\theta L(\theta, \underline{\theta})$ a *risk of the underestimation of the parameter* θ. Therefore Lemma 1.3.4 implies that a uniformly most precise lower

bound $\underline{\theta}$ minimizes the risk of the underestimation of the parameter θ. This together with Theorems 1.3.7 and 1.3.8 allows one to construct a uniformly most precise one-sided interval in an explicit form for families with monotone likelihood ratio.

THEOREM 1.3.9. *Let a family* $\{\mathsf{P}_\theta, \theta \in \Theta\}$ *have the monotone likelihood ratio with respect to a statistic* $T(x)$ *whose distribution function* $G_\theta(t) = \mathsf{P}_\theta\{T(\xi) < t\}$ *is continuous in* θ *and* t. *Then the distribution of the statistic* $T(x)$ *monotonically and continuously depends on the parameter* θ, *that is,* $G_\theta(t)$ *continuously decreases if* θ *increases* (*see relation* (5.4.1) *in* [**38**]). *If* $b(t,\gamma)$ *is a solution in* θ *of the equation* $G_\theta(t) = \gamma$, *then a uniformly most precise lower bound* $\underline{\theta}(\xi,\gamma)$ *of level* γ *is*

$$\underline{\theta}(\xi,\gamma) = b(T(\xi),\gamma).$$

PROOF. We put $\Theta(\theta) = \{t: t > \theta\}$ in Theorems 1.3.7 and 1.3.8. According to Theorem 1.3.3 there exists a nonrandomized UMP test for distinguishing the hypotheses $H_1: \theta = \theta_0$ and $H_2: \theta > \theta_0$. Moreover the acceptance set for the hypothesis H_1 is $X(\theta_0,\gamma) = \{x: T(x) < c\}$ where the constant $c = c(\theta_0,\gamma) = G_{\theta_0}^{-1}(\gamma)$ is such that

$$\mathsf{P}_{\theta_0}\{T(\xi) < c(\theta_0,\gamma)\} = \gamma.$$

By the assumptions of the theorem we have

$$\mathsf{P}_\theta\{T(\xi) \geq c\} > 1 - \gamma = \mathsf{P}_{\theta_0}\{T(\xi) \geq c\}$$

for all $\theta > \theta_0$. The latter relation means that $c(\theta_0,\gamma) < c(\theta,\gamma)$ for $\theta_0 < \theta$, that is, the function $c(\theta,\gamma)$ increases in θ. The continuity of $c(\theta,\gamma) = G_\theta^{-1}(\gamma)$ in θ follows from that of $G_\theta(t)$.

Thus the conditions of Theorems 1.3.7 and 1.3.8 hold for $c_2(\theta,\gamma) = c(\theta,\gamma)$ and therefore a uniformly most precise confidence interval is $(c^{-1}(T(\xi);\gamma),\infty)$ where obviously $c^{-1}(t,\gamma) = b(t,\gamma)$. □

A uniformly most precise upper bound $\overline{\theta}(\xi,\gamma)$ can be constructed by the same method if the assumptions of Theorem 1.3.9 hold.

Now let $\underline{\theta}(\xi,\gamma_1) < \overline{\theta}(\xi,\gamma_2)$ where $\underline{\theta}(\xi,\gamma_1)$ and $\overline{\theta}(\xi,\gamma_2)$ are lower and upper bounds of levels γ_1 and γ_2, respectively. Let γ_1 and γ_2 be such that the events $\{\underline{\theta}(\xi,\gamma_1) > \theta\}$ and $\{\overline{\theta}(\xi,\gamma_2) < \theta\}$ are disjoint. Then

$$\mathsf{P}_\theta\{\underline{\theta}(\xi,\gamma_1) < \theta < \overline{\theta}(\xi,\gamma_2)\} \geq \gamma_1 + \gamma_2 - 1,$$

that is, $(\underline{\theta}(\xi,\gamma_1), \overline{\theta}(\xi,\gamma_2))$ is a confidence interval of level $\gamma_1 + \gamma_2 - 1$.

Let $L_1(\theta,\underline{\theta})$ and $L_2(\theta,\overline{\theta})$ be the loss functions due to the underestimation of the parameter θ for bounds $\underline{\theta}$ and $\overline{\theta}$, respectively. Assume that $L_1(\theta,\underline{\theta})$ and $L_2(\theta,\overline{\theta})$ satisfy the conditions indicated above.

LEMMA 1.3.5. *Let* $L(\theta,\underline{\theta},\overline{\theta}) = L(\theta,\underline{\theta}) + L(\theta,\overline{\theta})$. *Then the confidence interval* $(\underline{\theta},\overline{\theta})$ *whose end points are uniformly most precise lower and upper bounds minimizes* $\mathsf{E}_\theta L(\theta,\underline{\theta},\overline{\theta})$ *under the conditions*

$$\mathsf{P}_\theta\{\underline{\theta} > \theta\} \leq 1 - \gamma_1, \qquad \mathsf{P}_\theta\{\overline{\theta} < \theta\} \leq 1 - \gamma_2.$$

This result is an obvious corollary of Lemma 1.3.4. Using Theorem 1.3.9 and Lemma 1.3.5 one can construct optimal intervals in an explicit form for families with the monotone likelihood ratio.

To conclude this section we show how to construct confidence sets and intervals with the help of unbiased tests.

As before let a set $\Theta(\theta)$ correspond to every θ such that $\theta \notin \Theta(\theta)$.

DEFINITION 1.3.11. A confidence set $\Theta^*(\xi, \gamma)$ of level γ for a parameter θ is called *unbiased with respect to the alternative hypothesis* θ' such that $\theta \in \Theta(\theta')$ if

$$\mathsf{P}_\theta\{\theta' \in \Theta^*(\xi, \gamma)\} \leq \gamma \quad \text{for all } \theta \text{ and } \theta' \text{ such that } \theta \in \Theta(\theta'). \tag{1.3.35}$$

The set $\Theta^*(\xi, \gamma)$ is called *unbiased* if (1.3.35) holds for all $\theta' \neq \theta$.

If a confidence set is unbiased, then the *probability that it contains a wrong value* θ' *of the parameter is less than or equal to the probability that it contains the true value of the parameter.*

DEFINITION 1.3.12. If conditions (1.3.29) and (1.3.35) hold, then a confidence set for which the probability (1.3.33) is minimal is called a *uniformly most precise unbiased confidence set* of level γ against the alternatives θ' such that $\theta \in \Theta(\theta')$.

THEOREM 1.3.10. 1) *Since inclusions* (1.3.31) *are equivalent, unbiased nonrandomized tests generate unbiased confidence sets, and vice versa.*

2) *If* $X(\theta_0, \gamma)$ *for any* $\theta_0 \in \Theta$ *is the acceptance set of the null hypothesis*

$$H_1: \theta = \theta_0$$

for a nonrandomized UMP test against the alternative $H_2: \theta \in \Theta(\theta_0)$, *then the corresponding set* $\Theta^*(\xi, \gamma)$ *is a uniformly most precise unbiased confidence set, and vice versa.*

PROOF. The method of proof is the same as that of Theorem 1.3.7. An additional reasoning is that if a test is unbiased, then so is the corresponding confidence set, and vice versa. Indeed relations (1.3.29) and (1.3.35) are equivalent to inequalities

$$\sup_{\theta \in \Theta(\theta_0)} \mathsf{P}_\theta\{\xi \in X(\theta_0, \gamma)\} \leq \gamma \leq \mathsf{P}_{\theta_0}\{\xi \in X(\theta_0, \gamma)\}.$$

If δ is a nonrandomized test (that is, $\delta(x) = 0$ for $x \in X(\theta_0, \gamma)$), then

$$\mathsf{E}_\theta \delta(\xi) = 1 - \mathsf{P}_\theta\{\xi \in X(\theta_0, \gamma)\}, \qquad \inf_{\theta \in \Theta(\theta_0)} \mathsf{E}_\theta \delta(\xi) \geq 1 - \gamma \geq \mathsf{E}_{\theta_0} \delta(\xi).$$

These conditions obviously mean that the test is unbiased and this property is equivalent to (1.3.35). □

Applying Theorem 1.3.10 one can construct a uniformly most precise unbiased confidence interval for a parameter of the exponential family. The method of this construction is the same as above.

EXAMPLE 1.3.6. Let $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables with the density

$$\theta^{-1} I(x > 0) \exp(-x/\theta), \qquad 0 < \theta < \infty.$$

The statistic $T_n(x) = \sum_{i=1}^n x_i$, $x = (x_1, \ldots, x_n)$, is a minimal essential statistic and moreover the distribution of $T_n(\xi)/(2\theta)$ is $\chi^2(2n)$. Denote by $\chi^2_\gamma(2n)$ the γ-quantile of the $\chi^2(2n)$ law. Then

$$\mathsf{P}_\theta\left\{\theta \geq T_n(\xi)/(2\chi^2_\gamma(2n))\right\} = \gamma,$$

that is, $\underline{\theta} = T_n(\xi)/(2\chi^2_\gamma(2n))$ is a lower confidence bound of level γ for the parameter θ. Further let $\zeta_{n,1} = \min_{1\le i\le n}\xi_i$ be the minimal order statistic. Then the distribution of $n\zeta_{n,1}/(2\theta)$ is $\chi^2(2)$. This implies that $\widetilde{\underline{\theta}} = n\zeta_{n,1}/(2\chi^2_\gamma(2))$ also is a lower bound of level γ for the parameter θ.

Which of these two bounds is better? To answer this question note that the distribution of ξ has the monotone likelihood ratio with respect to the statistic $T_n(x)$. Thus the UMP test of level α for distinguishing the hypotheses $H_1\colon\theta = \theta_0$ and $H_2\colon\theta > \theta_0$ is determined by the acceptance set of the hypothesis H_1 which is of the form $X(\theta_0, 1-\alpha) = \{x\colon T_n(x) \le 2\theta_0\chi^2_{1-\alpha}(2n)\}$. Thus a uniformly most precise lower bound of level γ for the parameter θ is $\underline{\theta} = T_n(\xi)/(2\chi^2_\gamma(2n))$. Therefore

$$\mathsf{P}_\theta\{\theta' \ge \underline{\theta}\} < \mathsf{P}_\theta\{\theta' \ge \widetilde{\underline{\theta}}\} \quad \text{for all } \theta \text{ and } \theta' \text{ such that } \theta > \theta'.$$

This implies that the bound $\underline{\theta}$ is better, since it is a uniformly most precise lower bound.

More details on confidence sets and intervals and on an application of statistical tests to construct confidence sets as well as various examples can be found in [**54**].

Bayes tests for distinguishing a finite number of composite hypotheses. The last topic of this section is the problem of distinguishing N composite hypotheses for $N \ge 2$. As before let the distribution of an observation ξ belong to a family $\mathscr{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ and let $\mathscr{P} = \bigcup_{i=1}^N \mathscr{P}_i$ where $\mathscr{P}_i = \{\mathsf{P}_\theta\colon\theta \in \Theta_i\}$, $i = 1, 2, \dots, N$; $\Theta_i \cap \Theta_j = \varnothing$, $i \ne j$; $\bigcup_{i=1}^N \Theta_i = \Theta$. Assume that at least one of the sets $\Theta_1, \Theta_2, \dots, \Theta_N$ contains at least two points. Let the hypothesis H_i be that the distribution of ξ belongs to the set $\mathscr{P}_i$. We write in this case $H_i\colon\theta \in \Theta_i$. Consider a randomized test δ whose values are $1, 2, \dots, N$ and the corresponding probabilities are $q_i^\delta(x) = \mathsf{P}\{\delta = i/\xi = x\}$. The event $\{\delta = i\}$ means that the hypothesis H_i is accepted, and $q_i^\delta(x)$ is the conditional probability of accepting the hypothesis H_i according to the test δ given $\xi = x$.

Let $A_i(t)$, $i = 1, 2, \dots, N$, be nonnegative functions defined on Θ that are measurable with respect to some σ-algebra of subsets of Θ. Let A be a random variable treated as the loss; it is equal to $A_i(t)$ if the hypothesis H_i is accepted and the parameter θ is t. Further we assume that a probability measure $\mathbf{Q}$ (a priori measure) is given on Θ, so that the parameter can be treated to be random with the distribution $\mathbf{Q}$. Then we define the mean loss or *risk of the test* δ (see (1.3.4) and (1.3.5)):

$$R(\delta) = \mathsf{E}^\delta A = \int_\Theta \mathsf{E}_t^\delta A\, \mathbf{Q}(dt) \tag{1.3.36}$$

where

$$\mathsf{E}_t^\delta A = \sum_{i=1}^N A_i(t)\mathsf{P}_t\{\delta(\xi) = i\}. \tag{1.3.37}$$

It is obvious that

$$\mathsf{P}_t\{\delta(\xi) = i\} = \int_X q_i^\delta(x)\, \mathsf{P}_t(dx). \tag{1.3.38}$$

Combining (1.3.36)–(1.3.38) we obtain

$$R(\delta) = \int_\Theta \sum_{i=1}^N A_i(t) \int_X q_i^\delta(x) \mathsf{P}_t(dx)\, \mathbf{Q}(dt). \tag{1.3.39}$$

According to Definition 1.3.3, a test $\delta_{A,Q}$ is Bayes with respect to a loss A and a priori measure $\mathbf{Q}$ if it minimizes the risk (1.3.39). Let us find the test $\delta_{A,Q}$ in an explicit form.

Let the family $\mathscr{P}$ be dominated by some σ-finite measure μ and let

$$p(x;t) = d\mathsf{P}_t/d\mu(x)$$

be the density of the measure P_t with respect to μ. Then the risk $R(\delta)$ can be written as

$$R(\delta) = \int_X \sum_{i=1}^N q_i^\delta(x) \int_\Theta A_i(t) p(x;t)\, \mathbf{Q}(dt)\, \mu(dx). \tag{1.3.40}$$

Put

$$R_i(x) = \int_\Theta A_i(t) p(x;t)\, \mathbf{Q}(dt), \qquad i = 1, 2, \ldots, N,\ x \in X. \tag{1.3.41}$$

It follows from (1.3.40) that the test δ minimizes the risk if

$$q_i^\delta(x) = \begin{cases} 1, & \text{if } \ R_i(x) = \min_{1 \le j \le N} R_j(x), \\ 0, & \text{otherwise,} \end{cases} \tag{1.3.42}$$

where $R_i(x)$ are defined by (1.3.41). If $\min_{1 \le j \le N} R_j(x)$ is attained for several indices $i_1, \ldots, i_k$, then one can proceed in the same way as in the case of simple hypotheses (see Section 1.2). Thus equality (1.3.42) defines a Bayes test $\delta_{A,Q}$.

More details about Bayes tests for distinguishing composite hypotheses can be found in [**9**, **54**].

EXAMPLE 1.3.7. Let the distribution of ξ be $\mathcal{N}(\theta, 1)$ and $\theta \in \Theta = (-\infty, \infty)$. Let $\Theta = \Theta_1 \cup \Theta_2 \cup \Theta_3$ where

$$\Theta_1 = (-\infty, -1), \qquad \Theta_2 = [-1, 1], \qquad \Theta_3 = (1, \infty).$$

Consider the loss functions

$$A_1(t) = I(t \ge -1), \qquad A_2(t) = I(|t| \ge 1), \qquad A_3(t) = I(t \le 1).$$

Let the a priori distribution $\mathbf{Q}$ of the parameter θ be $\mathcal{N}(0, \tau^2)$. Then the distribution of ξ is $\mathcal{N}(0, 1 + \tau^2)$, whence

$$p(x;t)\, \mathbf{Q}(dt) = n\left(x; 0, 1 + \tau^2\right) n\left(t; x\frac{\tau^2}{1+\tau^2}, \frac{\tau^4}{1+\tau^2}\right) dt$$

where $n(x; a, b^2)$ is the density of the $\mathcal{N}(a, b^2)$ law. Thus in view of (1.3.41) we get

$$\begin{aligned} R_1(x) &= n\left(x; 0, 1 + \tau^2\right) \int_{-1}^\infty n\left(t; x\frac{\tau^2}{1+\tau^2}, \frac{\tau^4}{1+\tau^2}\right) dt \\ &= n\left(x; 0, 1 + \tau^2\right) \Phi\left(\frac{1 + x^*}{\sigma}\right) \end{aligned}$$

where

$$x^* = x\frac{\tau^2}{1+\tau^2}, \qquad \sigma^2 = \frac{\tau^4}{1+\tau^2}.$$

Similarly we get

$$R_2(x) = n\left(x; 0, 1+\tau^2\right) \int_{|x|>1} n\left(t; x^*, \sigma^2\right) dt$$
$$= n\left(x; 0, 1+\tau^2\right) \left[2 - \Phi\left(\frac{1+x^*}{\sigma}\right) - \Phi\left(\frac{1-x^*}{\sigma}\right)\right],$$
$$R_3(x) = n\left(x; 0, 1+\tau^2\right) \Phi\left(\frac{1-x^*}{\sigma}\right).$$

Let

$$R_j^*(x) = R_j(x)/n\left(x; 0, 1+\tau^2\right).$$

The function $R_2^*(x)$ is symmetric with respect to x^*; its minimum is 0 and it is attained at the point $x = 0$. Further

$$\frac{d}{dx^*} R_2^*(x) = -\frac{1}{\sigma}\left[\varphi\left(\frac{1+x^*}{\sigma}\right) - \varphi\left(\frac{1-x^*}{\sigma}\right)\right] \tag{1.3.43}$$

where $\varphi(x)$ is the density of the $\mathcal{N}(0,1)$ law.

For all $x^* < 0$ it holds that $|1+x^*| < 1-x^*$. Hence $\varphi((1+x^*)/\sigma) > \varphi((1-x^*)/\sigma)$ for all $x^* < 0$. Equality (1.3.43) implies that $R_2^*(x)$ is decreasing in $(-\infty, 0)$ and $\lim_{x\to-\infty} R_2^*(x) = 1$. The function $R_1^*(x)$ is increasing and

$$\lim_{x\to-\infty} R_1^*(x) = 0, \qquad \lim_{x\to\infty} R_1^*(x) = 1.$$

Thus there exists a unique point x' in the interval $(-\infty, 0)$ such that

$$R_1^*(x') = R_2^*(x').$$

Since the problem is symmetric, there exists a unique point $x'' = -x'$ in the interval $(0, \infty)$ such that $R_2^*(x'') = R_3^*(x'')$. Thus the partition $X = X_1 \cup X_2 \cup X_3$ where $X_1 = (-\infty, x')$, $X_2 = [x', x'']$, and $X_3 = (x'', \infty)$ determines a Bayes test. According to this test the hypothesis H_i, $i = 1, 2, 3$, is accepted if $x \in X_i$.

REMARK 1.3.6. More details about distinguishing composite hypotheses can be found in [**9, 34, 54**]. Distinguishing composite hypotheses can be viewed as a problem of the general theory of statistical decisions; see [**4, 9, 15, 52, 54**]. Asymptotic problems of distinguishing composite hypotheses for independent observations are considered in [**10**]. Tests of significance are studied in Chapter 3 below.

CHAPTER 2

Asymptotic Distinguishability of Simple Hypotheses

2.1. Statistical hypotheses and tests

Let ξ^t, $t \in \mathbf{R}_+$, be a family of observations assuming values in a measurable space $(X^t, \mathscr{B}^t)$ and let $\mathscr{P}^t = \{\mathsf{P}^t, \widetilde{\mathsf{P}}^t\}$ be a family of two probability measures defined on $(X^t, \mathscr{B}^t)$. Let H^t and $\widetilde{H}^t$ be two statistical hypotheses that the distribution of the observation ξ^t is generated by the measures P^t and $\widetilde{\mathsf{P}}^t$, respectively. Denote by δ_t a measurable mapping of the space $(X^t, \mathscr{B}^t)$ into the space $([0,1], \mathscr{B}([0,1]))$. The mapping δ_t (as well as the random variable $\delta(\xi^t)$ denoted by the same symbol δ_t) is called a *test for distinguishing the hypotheses H^t and $\widetilde{H}^t$ by the observation ξ^t*; here we treat $\delta_t(x)$ as the conditional probability of rejecting the hypothesis H^t (or, equivalently, of accepting the hypothesis $\widetilde{H}^t$) given $\xi^t = x$. Denote by Σ^t the set of all tests δ_t for distinguishing the hypotheses H^t and $\widetilde{H}^t$; for all $\delta_t \in \Sigma^t$ we introduce the *type* I *and type* II *error probabilities*

$$\alpha(\delta_t) = \mathsf{E}^t \delta_t \quad \text{and} \quad \beta(\delta_t) = \widetilde{\mathsf{E}}^t(1 - \delta_t), \tag{2.1.1}$$

respectively, where E^t and $\widetilde{\mathsf{E}}^t$ are expectations with respect to the distributions P^t and $\widetilde{\mathsf{P}}^t$, respectively. For all $\alpha \in [0,1]$ we denote by Σ_α^t the set of all tests δ_t of Σ^t such that $\alpha(\delta_t) \le \alpha$.

Let $\mathbf{Q}^t = (\mathsf{P}^t + \widetilde{\mathsf{P}}^t)/2$ be another probability measure on the measurable space $(X^t, \mathscr{B}^t)$ and let $\mathfrak{z}_t = d\mathsf{P}^t/d\mathbf{Q}^t$ and $\tilde{\mathfrak{z}}_t = d\widetilde{\mathsf{P}}^t/d\mathbf{Q}^t$ be two finite versions of the Radon–Nikodym derivatives of the measures P^t and $\widetilde{\mathsf{P}}^t$, respectively, with respect to the measure $\mathbf{Q}^t$. Consider the likelihood ratios

$$z_t = \tilde{\mathfrak{z}}_t/\mathfrak{z}_t, \qquad \widetilde{z}_t = \mathfrak{z}_t/\tilde{\mathfrak{z}}_t. \tag{2.1.2}$$

Here we agree that $0/0 = 0$ and the likelihood ratios are well defined for all t. Note that $\mathfrak{z}_t + \tilde{\mathfrak{z}}_t = 2$.

Put

$$\overline{\alpha}_t = \mathsf{P}^t(\tilde{\mathfrak{z}}_t > 0), \qquad \overline{\beta}_t = \widetilde{\mathsf{P}}^t(\mathfrak{z}_t > 0). \tag{2.1.3}$$

Then

$$\overline{\alpha}_t = \mathsf{P}^t(z_t > 0) = \mathsf{P}^t(\widetilde{z}_t < \infty), \tag{2.1.4}$$

$$\overline{\beta}_t = \widetilde{\mathsf{P}}^t(\widetilde{z}_t > 0) = \widetilde{\mathsf{P}}^t(z_t < \infty). \tag{2.1.5}$$

Equalities (2.1.3)–(2.1.5) and Lemma 1.1.9 imply the following Lebesgue decomposition of any of the measures P^t or $\widetilde{\mathsf{P}}^t$ with respect to the other one.

LEMMA 2.1.1. *For all* $A \in \mathscr{B}^t$ *it holds that*

$$\widetilde{\mathsf{P}}^t(A) = \int_A z_t \, d\mathsf{P}^t + \widetilde{\mathsf{P}}^t(A \cap \{z_t = \infty\}), \tag{2.1.6}$$

$$\mathsf{P}^t(A) = \int_A \widetilde{z}_t \, d\widetilde{\mathsf{P}}^t + \mathsf{P}^t(A \cap \{\widetilde{z}_t = \infty\}) \tag{2.1.7}$$

where z_t *and* $\widetilde{z}_t$ *are the likelihood ratios defined by equalities* (2.1.2).

Lemma 1.1.10 can also be rewritten in the following form.

LEMMA 2.1.2. *For all nonnegative measurable functions* η *defined on* $(X^t, \mathscr{B}^t)$ *it holds that*

$$\widetilde{\mathsf{E}}^t \eta = \mathsf{E}^t \eta z_t + \widetilde{\mathsf{E}}^t \eta I(z_t = \infty), \tag{2.1.8}$$

$$\mathsf{E}^t \eta = \widetilde{\mathsf{E}}^t \eta \widetilde{z}_t + \mathsf{E}^t \eta I(\widetilde{z}_t = \infty). \tag{2.1.9}$$

Consider the set

$$\mathfrak{N}^t = \{(\alpha(\delta_t), \beta(\delta_t)) : \delta_t \in \Sigma^t\} \tag{2.1.10}$$

where $\alpha(\delta_t)$ and $\beta(\delta_t)$ are type I and type II error probabilities of the test $\delta_t \in \Sigma^t$ defined by (2.1.1).

The properties of the set $\mathfrak{N}^t$ are studied in Section 1.1. In particular, the set $\mathfrak{N}^t$ is convex, closed, and symmetric about the point $(1/2, 1/2)$, $\mathfrak{N}^t$ contains the points $(0,1)$ and $(1,0)$, and $\mathfrak{N}^t \subseteq [0,1] \times [0,1]$. An example of the set $\mathfrak{N}^t$ is shown in Figure 2.1.1.

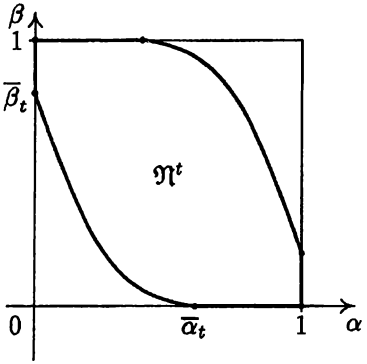

FIGURE 2.1.1

Now we introduce the *likelihood ratio test* by

$$\delta_t^{c,\varepsilon} = I(z_t > c) + \varepsilon I(z_t = c) \tag{2.1.11}$$

where $c \in [0, \infty]$ and $\varepsilon \in [0,1]$ are parameters of the test. Let $(c_t(\alpha), \varepsilon_t(\alpha))$ be some solution of the equation $\alpha(\delta_t^{c,\varepsilon}) = \alpha$ with respect to (c, ε).

A likelihood ratio test $\delta_t^{c_t(\alpha), \varepsilon_t(\alpha)}$ with $\varepsilon_t(0) = 1$ is called the *Neyman–Pearson test of level* α *for distinguishing the hypotheses* H^t *and* $\widetilde{H}^t$. In what follows we denote this test by $\delta_t^{+,\alpha}$. The functions $c_t(\alpha)$ and $\varepsilon_t(\alpha)$ as well as the type II error probability $\beta(\delta_t^{+,\alpha})$ can be obtained by applying results of Section 1.1.

According to the Neyman–Pearson fundamental lemma, the point $(\alpha, \beta(\delta_t^{+,\alpha}))$ belongs to the boundary of the set $\mathfrak{N}^t$ for all $\alpha \in [0,1]$. In various cases where the level α depends on t, one can study the behavior of the set $\mathfrak{N}^t$ as $t \to \infty$ instead of that of $\beta(\delta_t^{+,\alpha})$ as $t \to \infty$.

Below we consider the Neyman–Pearson test δ_t^{+,α_t} of level α_t that depends, generally speaking, on t. We will use the notation $c_t = c_t(\alpha_t)$ and $\varepsilon_t = \varepsilon_t(\alpha_t)$. Moreover we put $\Lambda_t = \ln z_t$ and $d_t = \ln c_t$ assuming that $\ln 0 = -\infty$. Then the test δ_t^{+,α_t} can be rewritten as

$$\delta_t^{+,\alpha_t} = I(\Lambda_t > d_t) + \varepsilon_t I(\Lambda_t = d_t). \tag{2.1.12}$$

Consider the *Bayes test δ_t^π with respect to the a priori distribution $(\pi, 1-\pi)$ for distinguishing the hypotheses H^t and $\widetilde{H}^t$* that minimizes the error probability $e_\pi(\delta_t)$ (see (1.1.32)) in the class Σ^t of all tests δ_t.

Denote by δ_t^* the *minimax test for distinguishing the hypotheses H^t and $\widetilde{H}^t$* that minimizes $\alpha(\delta_t) \vee \beta(\delta_t)$ in the class Σ^t of all tests δ_t. It follows from Theorem 1.2.4 that the likelihood ratio test $\delta_t^{c,\varepsilon}$ is minimax if $\alpha(\delta_t^{c,\varepsilon}) = \beta(\delta_t^{c,\varepsilon})$. According to (2.1.11) and (2.1.12) the Neyman–Pearson test is minimax if $\beta(\delta_t^{+,\alpha_t}) = \alpha_t$. Moreover we learned in Section 1.2 that the Neyman–Pearson test δ_t^{+,α_t} coincides with some Bayes test $\delta_t^{\pi_t}$, and the probability π_t depends, generally speaking, on t.

Below we study the asymptotic behavior (as $t \to \infty$) of the Neyman–Pearson test δ_t^{+,α_t}, the minimax δ_t^* test, and the Bayes $\delta_t^{\pi_t}$ test. The asymptotic behavior of these tests depends on the behavior of the set $\mathfrak{N}^t$ which in turn is determined by the behavior of the measures P^t and $\widetilde{\mathsf{P}}^t$ as $t \to \infty$ (see Section 1.1). In the next section we consider all the possible types of the asymptotic behavior of the set $\mathfrak{N}^t$ as $t \to \infty$.

2.2. Types of the asymptotic distinguishability of families of hypotheses. The characterization of types

Consider two families of statistical hypotheses (H^t) and $(\widetilde{H}^t)$ and the two corresponding families of probability measures (P^t) and $(\widetilde{\mathsf{P}}^t)$. Here and throughout this section the symbols (H^t) and $(\widetilde{H}^t)$ stand for $(H^t)_{t\in\mathbf{R}_+}$ and $(\widetilde{H}^t)_{t\in\mathbf{R}_+}$, respectively. For other families of measures, random variables, etc. we follow the same notation to make them shorter.

Below we define the types of the asymptotic distinguishability of families of hypotheses (H^t) and $(\widetilde{H}^t)$ as $t \to \infty$. Our approach is based on the asymptotic behavior of sets $\mathfrak{N}^t$ as $t \to \infty$.

The distance in variance between measures. The Kakutani–Hellinger distance and Hellinger integrals. To state the main results on the characterization of types we need the following notions.

DEFINITION 2.2.1. Let P^t and $\widetilde{\mathsf{P}}^t$ be two measures. The full variation of the charge $\widetilde{\mathsf{P}}^t - \mathsf{P}^t$ is called *the distance in variance between measures P^t and $\widetilde{\mathsf{P}}^t$* and is denoted by $\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\|$, namely

$$\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| = \mathsf{E}_Q^t |\tilde{\mathfrak{z}}_t - \mathfrak{z}_t| \tag{2.2.1}$$

where E_Q^t stands for the expectation with respect to the measure $\mathbf{Q}^t$.

DEFINITION 2.2.2. Let

$$\rho^2(\mathsf{P}^t, \widetilde{\mathsf{P}}^t) = \frac{1}{2}\mathsf{E}_Q^t \left(\tilde{\jmath}_t^{1/2} - \jmath_t^{1/2}\right)^2. \tag{2.2.2}$$

The nonnegative number $\rho(\mathsf{P}^t, \widetilde{\mathsf{P}}^t)$ is called the *Kakutani–Hellinger distance between the measures* P^t *and* $\widetilde{\mathsf{P}}^t$.

DEFINITION 2.2.3. Let

$$H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) = \mathsf{E}_Q^t \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon}. \tag{2.2.3}$$

Here we put

$$\tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} = \begin{cases} 0, & \varepsilon < 0,\ \tilde{\jmath}_t = 0, \text{ and } \jmath_t = 0, \\ \infty, & \varepsilon < 0,\ \tilde{\jmath}_t = 0, \text{ and } \jmath_t > 0, \\ \jmath_t I(\tilde{\jmath}_t > 0), & \varepsilon = 0, \\ \tilde{\jmath}_t I(\jmath_t > 0), & \varepsilon = 1, \\ 0, & \varepsilon > 1,\ \jmath_t = 0, \text{ and } \tilde{\jmath}_t = 0, \\ \infty, & \varepsilon > 1,\ \jmath_t = 0, \text{ and } \tilde{\jmath}_t > 0. \end{cases}$$

Then $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ is well defined for all ε and t. The number $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ is called the *Hellinger integral of order* $\varepsilon \in \mathbf{R} = (-\infty, \infty)$ *for the measures* $\widetilde{\mathsf{P}}^t$ *and* P^t. The number $H(1/2; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ is called the *Hellinger integral for the measures* $\widetilde{\mathsf{P}}^t$ *and* P^t and is denoted by $H(\widetilde{\mathsf{P}}^t, \mathsf{P}^t)$.

Properties of $\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\|$, $\rho(\widetilde{\mathsf{P}}^t, \mathsf{P}^t)$, and $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ defined by (2.2.1), (2.2.2), and (2.2.3), respectively, can be found in [**28, 33, 35, 47**]. In particular, neither $\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\|$ nor $\rho(\widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ nor $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ depend on the dominating measure $\mathbf{Q}^t$.

Below we give an auxiliary result on some relationships between these notions.

LEMMA 2.2.1. *We have*

$$2\left(1 - H\left(\widetilde{\mathsf{P}}^t, \mathsf{P}^t\right)\right) \le \|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| \le \sqrt{8\left(1 - H\left(\widetilde{\mathsf{P}}^t, \mathsf{P}^t\right)\right)}, \tag{2.2.4}$$

$$\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| \le 2\sqrt{1 - H^2\left(\widetilde{\mathsf{P}}^t, \mathsf{P}^t\right)}. \tag{2.2.5}$$

In particular

$$2\rho^2(\mathsf{P}^t, \widetilde{\mathsf{P}}^t) \le \|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| \le \sqrt{8}\,\rho(\mathsf{P}^t, \widetilde{\mathsf{P}}^t). \tag{2.2.6}$$

PROOF. By the definitions of the Kakutani–Hellinger distance and Hellinger integral

$$\rho^2(\mathsf{P}^t, \widetilde{\mathsf{P}}^t) = 1 - H(\widetilde{\mathsf{P}}^t, \mathsf{P}^t). \tag{2.2.7}$$

Thus inequalities (2.2.6) follow from (2.2.4). Let us prove inequalities (2.2.4) and (2.2.5).

Since $\jmath_t + \tilde{\jmath}_t = 2$, it follows from the Jensen inequality that

$$\begin{aligned} \frac{1}{2}\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| &= \frac{1}{2}\mathsf{E}_Q^t|\tilde{\jmath}_t - \jmath_t| = \mathsf{E}_Q^t|1 - \jmath_t| \le \sqrt{\mathsf{E}_Q^t(1 - \jmath_t)^2} \\ &= \sqrt{1 - \mathsf{E}_Q^t \jmath_t(2 - \jmath_t)} = \sqrt{1 - \mathsf{E}_Q^t \jmath_t \tilde{\jmath}_t}. \end{aligned}$$

It follows from the Cauchy–Bunyakovskiĭ inequality that

$$\mathsf{E}_Q^t\sqrt{\mathfrak{z}_t\tilde{\mathfrak{z}}_t} \le \sqrt{\mathsf{E}_{Q\mathfrak{z}_t\tilde{\mathfrak{z}}_t}^t},$$

and thus

$$\frac{1}{2}\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| \le \sqrt{1-\mathsf{E}_{Q\mathfrak{z}_t\tilde{\mathfrak{z}}_t}^t} \le \sqrt{1-\left(\mathsf{E}_Q^t\sqrt{\mathfrak{z}_t\tilde{\mathfrak{z}}_t}\right)^2} = \sqrt{1-H^2(\widetilde{\mathsf{P}}^t,\mathsf{P}^t)},$$

that is, (2.2.5) is proved.

Since $H(\widetilde{\mathsf{P}}^t,\mathsf{P}^t) \le 1$, we have

$$1-H^2(\widetilde{\mathsf{P}}^t,\mathsf{P}^t) = \left(1-H(\widetilde{\mathsf{P}}^t,\mathsf{P}^t)\right)\left(1+H(\widetilde{\mathsf{P}}^t,\mathsf{P}^t)\right) \le 2\left(1-H(\widetilde{\mathsf{P}}^t,\mathsf{P}^t)\right).$$

This estimate together with (2.2.5) implies the second inequality in (2.2.4).

Since $(a-b)^2 \le |a^2-b^2|$ for $a>0$ and $b>0$, we get

$$\frac{1}{2}\left(\sqrt{z}-\sqrt{2-z}\right)^2 \le |z-1|, \qquad 0 \le z \le 2. \tag{2.2.8}$$

Using (2.2.7), equality $\mathfrak{z}_t+\tilde{\mathfrak{z}}_t=2$, and inequality (2.2.8) we obtain

$$\begin{aligned}1-H(\widetilde{\mathsf{P}}^t,P^t) = \rho^2(\mathsf{P}^t,\widetilde{\mathsf{P}}^t) &= \frac{1}{2}\mathsf{E}_Q^t(\sqrt{\mathfrak{z}_t}-\sqrt{1-\mathfrak{z}_t})^2\\ &\le \mathsf{E}_Q^t|\mathfrak{z}_t-1| = \frac{1}{2}\mathsf{E}_Q^t|\mathfrak{z}_t-\tilde{\mathfrak{z}}_t| = \frac{1}{2}\|\mathsf{P}^t-\widetilde{\mathsf{P}}^t\|,\end{aligned}$$

that is, the first inequality in (2.2.4) is also proved. □

Let

$$\|\mathsf{P}^t\wedge\widetilde{\mathsf{P}}^t\| = \inf\left\{\alpha(\delta_t)+\beta(\delta_t)\colon \delta_t\in\Sigma^t\right\}. \tag{2.2.9}$$

It is clear that $\|\mathsf{P}^t\wedge\widetilde{\mathsf{P}}^t\|$ is the doubled error of the Bayes test with respect to the a priori distribution $(1/2,1/2)$.

LEMMA 2.2.2. *We have*

$$\|\mathsf{P}^t\wedge\widetilde{\mathsf{P}}^t\| = 1-\frac{1}{2}\|\widetilde{\mathsf{P}}^t-\mathsf{P}^t\|. \tag{2.2.10}$$

PROOF. Since the Bayes test with respect to the a priori distribution $(1/2,1/2)$ can be represented as $\delta_t^{1,1} = I(z_t\ge 1)$, relation (2.2.9) implies

$$\begin{aligned}\|\mathsf{P}^t\wedge\widetilde{\mathsf{P}}^t\| &= \mathsf{E}^t I(z_t\ge 1)+\widetilde{\mathsf{E}}^t(1-I(z_t\ge 1))\\ &= 1+\mathsf{E}_{Q\mathfrak{z}_t}^t I(z_t\ge 1)-\mathsf{E}_{Q\tilde{\mathfrak{z}}_t}^t I(z_t\ge 1)\\ &= 1-\mathsf{E}_Q^t(\tilde{\mathfrak{z}}_t-\mathfrak{z}_t)I(z_t\ge 1).\end{aligned} \tag{2.2.11}$$

Since $\mathsf{E}_Q^t(\tilde{\mathfrak{z}}_t-\mathfrak{z}_t)=0$, we obtain from (2.2.1) that

$$\begin{aligned}\|\widetilde{\mathsf{P}}^t-\mathsf{P}^t\| &= \mathsf{E}_Q^t(\tilde{\mathfrak{z}}_t-\mathfrak{z}_t)I(z_t\ge 1)+\mathsf{E}_Q^t(\mathfrak{z}_t-\tilde{\mathfrak{z}}_t)I(z_t<1)\\ &= 2\mathsf{E}_Q^t(\tilde{\mathfrak{z}}_t-\mathfrak{z}_t)I(z_t\ge 1).\end{aligned} \tag{2.2.12}$$

Now (2.2.11) and (2.2.12) imply equality (2.2.10). □

REMARK 2.2.1. Similarly to (2.2.11) we prove that

$$\|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\| = \mathsf{E}^t I(z_t \geq 1) + \widetilde{\mathsf{E}}^t I(z_t < 1) = \mathsf{E}^t_Q \mathfrak{z}_t I(z_t \geq 1) + \mathsf{E}^t_Q \tilde{\mathfrak{z}}_t I(z_t < 1)$$
$$= \mathsf{E}^t_Q (\mathfrak{z}_t \wedge \tilde{\mathfrak{z}}_t)$$

where $a \wedge b$ stands for the minimum of two numbers a and b. The latter equality makes the notation $\|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\|$ clear.

The complete asymptotic distinguishability. First we give some necessary definitions.

DEFINITION 2.2.4. Families of hypotheses (H^t) and $(\widetilde{H}^t)$ are called *completely asymptotically distinguishable* (denoted by $(H^t) \Delta (\widetilde{H}^t)$) if there exist a sequence $t_n \uparrow \infty$, $n \to \infty$, and tests $\delta_{t_n} \in \Sigma^{t_n}$ such that

$$\alpha(\delta_{t_n}) \to 0 \quad \text{and} \quad \beta(\delta_{t_n}) \to 0 \quad \text{as } n \to \infty. \tag{2.2.13}$$

DEFINITION 2.2.5. Families of measures (P^t) and $(\widetilde{\mathsf{P}}^t)$ are called *completely asymptotically separable* (denoted by $(\mathsf{P}^t) \Delta (\widetilde{\mathsf{P}}^t)$) if there exist a sequence $t_n \uparrow \infty$, $n \to \infty$, and sets $A_n \in \mathscr{B}^{t_n}$ such that

$$\mathsf{P}^{t_n}(A_n) \to 0 \quad \text{and} \quad \widetilde{\mathsf{P}}^{t_n}(A_n) \to 1 \quad \text{as } n \to \infty. \tag{2.2.14}$$

The following result contains a characterization of the complete asymptotic distinguishability $(H^t) \Delta (\widetilde{H}^t)$.

THEOREM 2.2.1. *The following statements are equivalent*:

a) $(H^t) \Delta (\widetilde{H}^t)$;
b) $(\mathsf{P}^t) \Delta (\widetilde{\mathsf{P}}^t)$;
c) $\limsup_{t\to\infty} \widetilde{\mathsf{P}}^t(z_t > N) = 1$ *for all* $N < \infty$;
d) $\limsup_{t\to\infty} \mathsf{P}^t(z_t < N) = 1$ *for all* $N > 0$;
e) $\liminf_{t\to\infty} \|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\| = 0$;
f) $\liminf_{t\to\infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) = 0$ *for all* $\varepsilon \in (0,1)$;
g) $\limsup_{t\to\infty} \|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| = 2$;
h) $\limsup_{t\to\infty} \rho(\mathsf{P}^t, \widetilde{\mathsf{P}}^t) = 1$.

PROOF. a) $\Rightarrow$ b). Let $t_n \uparrow \infty$ as $n \to \infty$ and let tests $\delta_{t_n} \in \Sigma^{t_n}$ be such that $\alpha(\delta_{t_n}) \to 0$ and $\beta(\delta_{t_n}) \to 0$. Put $A_n = I(\delta_{t_n} > \gamma)$ for $0 < \gamma < 1$. By the Chebyshev inequality we have

$$\mathsf{P}^{t_n}(A_n) = \mathsf{P}^{t_n}(\delta_{t_n} > \gamma) \leq \gamma^{-1} \alpha(\delta_{t_n}) \to 0.$$

Similarly we obtain

$$\widetilde{\mathsf{P}}^{t_n}(\delta_{t_n} \leq \gamma) \leq \widetilde{\mathsf{P}}^{t_n}(1 - \delta_{t_n} \geq 1 - \gamma) \leq (1-\gamma)^{-1} \beta(\delta_{t_n}) \to 0,$$

that is, $\widetilde{\mathsf{P}}^{t_n}(A_n) \to 1$. Thus the implication a) $\Rightarrow$ b) is proved.

b) $\Rightarrow$ d). Let $t_n \uparrow \infty$ as $n \to \infty$ and let sets $A_n \in \mathscr{B}^{t_n}$ be such that

$$\mathsf{P}^{t_n}(A_n) \to 0, \qquad \widetilde{\mathsf{P}}^{t_n}(A_n) \to 1. \tag{2.2.15}$$

By the Lebesgue decomposition (2.1.7) we have for all $A \in \mathscr{B}^t$ that

$$\begin{aligned} \mathsf{P}^t(A) &= \mathsf{P}^t(A \cap \{\widetilde{z}_t \le N\}) + \mathsf{P}^t(A \cap \{\widetilde{z}_t > N\}) \\ &= \int_{A \cap \{\widetilde{z}_t \le N\}} \widetilde{z}_t \, d\widetilde{\mathsf{P}}^t + \mathsf{P}^t(A \cap \{\widetilde{z}_t > N\}) \le N\widetilde{\mathsf{P}}^t(A) + \mathsf{P}^t(\widetilde{z}_t > N) \end{aligned} \tag{2.2.16}$$

for $0 < N < \infty$. Relations (2.2.15) and estimate (2.2.16) for $A = A_n^c = X^t \backslash A_n$ and $t = t_n$ imply that $\mathsf{P}^{t_n}(\widetilde{z}_{t_n} > N) \to 1$ for all $N \in (0, \infty)$, whence we obtain d), since

$$\mathsf{P}^t(\widetilde{z}_t > N) = \mathsf{P}^t(z_t < N^{-1}). \tag{2.2.17}$$

d) $\Rightarrow$ f). For all $\varepsilon, \lambda \in (0,1)$ and $\gamma > 0$ we have

$$\begin{aligned} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) &= \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t < \gamma) + \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t \ge \gamma, \jmath_t \le \lambda) \\ &\quad + \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t \ge \gamma, \jmath_t > \lambda) \\ &\le \gamma^{\varepsilon} \mathsf{E}^t_Q \jmath_t^{1-\varepsilon} + \lambda^{1-\varepsilon} \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} + \mathsf{E}^t_Q \jmath_t z_t^{\varepsilon} I(\tilde{\jmath}_t \ge \gamma, \jmath_t > \lambda) \\ &\le \gamma^{\varepsilon} + \lambda^{1-\varepsilon} + \mathsf{E}^t z_t^{\varepsilon} I(\tilde{\jmath}_t \ge \gamma) \\ &\le \gamma^{\varepsilon} + \lambda^{1-\varepsilon} + (\mathsf{E}^t z_t)^{\varepsilon} \big(\mathsf{E}^t I(\tilde{\jmath}_t \ge \gamma)\big)^{1-\varepsilon} \\ &\le \gamma^{\varepsilon} + \lambda^{1-\varepsilon} + [\mathsf{P}^t(\tilde{\jmath}_t \ge \gamma)]^{1-\varepsilon}, \end{aligned} \tag{2.2.18}$$

since $\mathsf{E}^t_Q \jmath_t^{1-\varepsilon} \le 1$ and $\mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \le 1$ by the Hölder inequality, and $\mathsf{E}^t z_t \le 1$ by (2.1.8) for $\eta = 1$.

Equality (2.2.17) and relation d) imply that $\limsup_{t\to\infty} \mathsf{P}^t(\widetilde{z}_t > N) = 1$. Since $\jmath_t + \tilde{\jmath}_t = 2$, we have $\widetilde{z}_t = 2/\tilde{\jmath}_t - 1$, whence

$$\liminf_{t\to\infty} \mathsf{P}^t(\tilde{\jmath}_t \ge \gamma) = 0 \tag{2.2.19}$$

for all $\gamma > 0$. It follows from (2.2.18) and (2.2.19) that

$$\liminf_{t\to\infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) \le \gamma^{\varepsilon} + \lambda^{1-\varepsilon}.$$

Since γ and λ are arbitrary, the latter inequality proves f).

f) $\Rightarrow$ h). Follows from (2.2.7).

h) $\Rightarrow$ g). This implication follows from the first inequality in (2.2.6) and

$$\|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| \le 2$$

(the latter estimate holds in view of (2.2.1)).

g) $\Rightarrow$ e). Follows from inequality (2.2.10).

e) $\Rightarrow$ a). Relation (2.2.9) and Theorem 1.2.2 for $N = 2$, $A_{ij} = 1 - \delta_{ij}$, and $\pi_1 = \pi_2 = 1/2$ yield

$$\|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\| = \alpha(\delta_t^{1,1}) + \beta(\delta_t^{1,1}) \tag{2.2.20}$$

where $\delta_t^{1,1} = I(z_t \ge 1)$. Condition e) and equality (2.2.20) imply that there exists a sequence $t_n \uparrow \infty$, $n \to \infty$, such that $\alpha(\delta_{t_n}^{1,1}) \to 0$ and $\beta(\delta_{t_n}^{1,1}) \to 0$ as $n \to \infty$, that is, condition a) holds.

Therefore conditions a), b), d), e), f), g), and h) are equivalent. It remains to prove that conditions b) and c) are equivalent.

b) $\Rightarrow$ c). Let $t_n \to \infty$ as $n \to \infty$ and let sets $A_n \in \mathscr{B}^{t_n}$ be such that relations (2.2.14) hold. The Lebesgue decomposition (2.1.6) for all $A \in \mathscr{B}^t$ and $0 < N < \infty$ imply similarly to (2.2.16) that

$$\widetilde{\mathsf{P}}^t(A) \le N\mathsf{P}^t(A) + \widetilde{\mathsf{P}}^t(z_t > N).$$

This together with (2.2.14) yields c).

c) $\Rightarrow$ b). Condition c) implies that there exists a sequence $t_n \uparrow \infty$, $n \to \infty$, such that

$$\widetilde{\mathsf{P}}^{t_n}(z_{t_n} > n) \ge 1 - n^{-1}.$$

Thus $\widetilde{\mathsf{P}}^{t_n}(z_{t_n} > n) \to 1$ as $n \to \infty$. On the other hand, the Chebyshev inequality implies that

$$\mathsf{P}^{t_n}(z_{t_n} > n) \le n^{-1}\mathsf{E}^{t_n} z_{t_n} \le n^{-1},$$

since $\mathsf{E}^t z_t \le 1$ in view of (2.1.8). This implies that $\mathsf{P}^{t_n}(z_{t_n} > n) \to 0$ as $n \to \infty$ and condition b) is proved. □

REMARK 2.2.2. If $(H^t) \Delta (\widetilde{H}^t)$, then there exist a sequence $t_n \uparrow \infty$, $n \to \infty$, and tests $\delta_{t_n} \in \Sigma^{t_n}$ such that conditions (2.2.13) hold. Then, obviously,

$$\mathfrak{N}^{t_n} \to [0,1] \times [0,1]$$

as $t \to \infty$ where the set $\mathfrak{N}^t$ is defined by (2.1.10). On the other hand, if

$$(\mathsf{P}^t) \Delta (\widetilde{\mathsf{P}}^t),$$

then there exist a sequence $t_n \uparrow \infty$, $n \to \infty$, and sets $A_n \in \mathscr{B}^{t_n}$ such that relations (2.2.14) hold. These relations mean that the sequences of measures P^{t_n} and $\widetilde{\mathsf{P}}^{t_n}$, $n = 1, 2, \dots$, are *asymptotically singular* (cf. (1.1.9)).

Now we define the counterparts of the notions of the complete asymptotic distinguishability $(H^t) \Delta (\widetilde{H}^t)$ and complete asymptotic separability $(\mathsf{P}^t) \Delta (\widetilde{\mathsf{P}}^t)$.

DEFINITION 2.2.6. We say that *families of hypotheses* (H^t) *and* $(\widetilde{H}^t)$ *are not completely asymptotically distinguishable* (denoted by $(H^t) \overline{\Delta} (\widetilde{H}^t)$) if there is no sequence $t_n \uparrow \infty$, $n \to \infty$, and tests $\delta_{t_n} \in \Sigma^{t_n}$ such that relations (2.2.13) hold.

DEFINITION 2.2.7. We say that *families of measures* (P^t) *and* $(\widetilde{\mathsf{P}}^t)$ *are not completely asymptotically separable* (denoted by $(\mathsf{P}^t) \overline{\Delta} (\widetilde{\mathsf{P}}^t)$) if there is no sequence $t_n \uparrow \infty$, $n \to \infty$, and sets $A_n \in \mathscr{B}^{t_n}$ such that relations (2.2.14) hold.

REMARK 2.2.3. If $(H^t) \overline{\Delta} (\widetilde{H}^t)$, then according to Definition 2.2.6

$$\liminf_{n\to\infty} \beta(\delta_{t_n}) > 0$$

for all sequences $t_n \uparrow \infty$, $n \to \infty$, and all tests $\delta_{t_n} \in \Sigma^{t_n}$ such that $\alpha(\delta_{t_n}) \to 0$ as $n \to \infty$. A similar remark regarding the notion $(\mathsf{P}^t) \overline{\Delta} (\widetilde{\mathsf{P}}^t)$ also holds. Therefore we have the following dichotomy: either $(H^t) \Delta (\widetilde{H}^t)$ or $(H^t) \overline{\Delta} (\widetilde{H}^t)$ (respectively, either $(\mathsf{P}^t) \Delta (\widetilde{\mathsf{P}}^t)$ or $(\mathsf{P}^t) \overline{\Delta} (\widetilde{\mathsf{P}}^t)$). Since Theorem 2.2.1 provides the necessary and sufficient conditions for $(H^t) \Delta (\widetilde{H}^t)$, it can be used to characterize the notion $(H^t) \overline{\Delta} (\widetilde{H}^t)$. For example, $(H^t) \overline{\Delta} (\widetilde{H}^t) \iff \liminf_{t\to\infty} \|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\| > 0$.

The complete asymptotic indistinguishability. Now we consider an asymptotic analog of the indistinguishability of hypotheses (cf. (1.1.8)).

DEFINITION 2.2.8. We say that families of hypotheses (H^t) and $(\widetilde{H}^t)$ are *completely asymptotically indistinguishable* (denoted by $(H^t) \cong (\widetilde{H}^t)$) if

$$\lim_{t\to\infty} \alpha(\delta_t) = \alpha \Rightarrow \lim_{t\to\infty} \beta(\delta_t) = 1 - \alpha \tag{2.2.21}$$

for all $\alpha \in [0,1]$ and all families (δ_t) of tests $\delta_t \in \Sigma^t$ such that the limit $\lim_{t\to\infty} \alpha(\delta_t)$ exists.

DEFINITION 2.2.9. We say that families of measures (P^t) and $(\widetilde{\mathsf{P}}^t)$ are *completely asymptotically inseparable* (denoted by $(\mathsf{P}^t) \cong (\widetilde{\mathsf{P}}^t)$) if

$$\lim_{t\to\infty} \mathsf{P}^t(A_t) = \alpha \Rightarrow \lim_{t\to\infty} \widetilde{\mathsf{P}}^t(A_t) = \alpha \tag{2.2.22}$$

for all $\alpha \in [0,1]$ and families (A_t) of sets $A_t \in \mathscr{B}^t$ such that the limit $\lim_{t\to\infty} \mathsf{P}^t(A_t)$ exists.

The following result contains a characterization of the complete asymptotic indistinguishability of families of hypotheses.

THEOREM 2.2.2. *The following statements are equivalent*:

a) $(H^t) \cong (\widetilde{H}^t)$;
b) $(\mathsf{P}^t) \cong (\widetilde{\mathsf{P}}^t)$;
c) $\lim_{t\to\infty} \mathsf{P}^t(|\Lambda_t| > \gamma) = 0$ *for all* $\gamma > 0$;
d) $\lim_{t\to\infty} \widetilde{\mathsf{P}}^t(|\Lambda_t| > \gamma) = 0$ *for all* $\gamma > 0$;
e) $\lim_{t\to\infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) = 1$ *for all* $\varepsilon \in (0,1)$;
f) $\lim_{t\to\infty} \rho(\mathsf{P}^t, \widetilde{\mathsf{P}}^t) = 0$;
g) $\lim_{t\to\infty} \|\widetilde{\mathsf{P}}^t - \mathsf{P}^t\| = 0$;
h) $\lim_{t\to\infty} \|\mathsf{P}^t \wedge \widetilde{\mathsf{P}}^t\| = 1$.

PROOF. a) $\Rightarrow$ b). Let α be an arbitrary number of the interval $[0,1]$ and let (A_t) be an arbitrary family of sets $A_t \in \mathscr{B}^t$ such that $\mathsf{P}^t(A_t) \to \alpha$ as $t \to \infty$. The test $\delta_t = I(A_t)$ is such that $\alpha(\delta_t) = \mathsf{P}^t(A_t) \to \alpha$. Then condition a) implies that $\widetilde{\mathsf{P}}^t(A_t) = 1 - \beta(\delta_t) \to \alpha$ as $t \to \infty$, that is, the implication (2.2.22) is proved.

b) $\Rightarrow$ c). We prove this implication by contradiction. Assume that condition b) holds and condition c) does not hold. Then there exist a number $\gamma_0 > 0$ and a sequence (t_n) such that $t_n \to \infty$ and $\mathsf{P}^{t_n}(|\Lambda_{t_n}| > \gamma_0) \to a > 0$ as $n \to \infty$. Thus there are sequences $(m) \subset (n)$ and $(k) \subset (n)$ such that either $\mathsf{P}^{t_m}(\Lambda_{t_m} > \gamma_0) \to b$ as $m \to \infty$ for some constant $b > 0$, or $\mathsf{P}^{t_k}(\Lambda_{t_k} < -\gamma_0) \to c$ as $k \to \infty$ for some constant $c > 0$.

Let $\mathsf{P}^{t_m}(\Lambda_{t_m} > \gamma_0) \to b > 0$ as $m \to \infty$. Then condition b) implies that $\widetilde{\mathsf{P}}^{t_m}(\Lambda_{t_m} > \gamma_0) \to b$ as $m \to \infty$. From the Lebesgue decomposition (2.1.6) we obtain

$$\widetilde{\mathsf{P}}^t(\Lambda_t > \gamma_0) = \int_{(\Lambda_t > \gamma_0)} \exp(\Lambda_t)\, d\mathsf{P}^t + \widetilde{\mathsf{P}}^t(\Lambda_t = \infty) \geq e^{\gamma_0} \mathsf{P}^t(\Lambda_t > \gamma_0).$$

Passing to the limit in this inequality along the sequence (t_m) we get $b \geq e^{\gamma_0} b$. Since $b > 0$ and $\gamma_0 > 0$, we obtain the contradiction $b > b$. Now we let

$$\mathsf{P}^{t_k}(\Lambda_{t_k} < -\gamma_0) \to c > 0$$

as $k \to \infty$ and arrive at a similar contradiction. These contradictions prove the implication b) $\Rightarrow$ c).

c) $\Rightarrow$ d). Using the Lebesgue decomposition (2.1.6) we obtain

$$\widetilde{\mathsf{P}}^t(|\Lambda_t| \le \gamma) \ge \widetilde{\mathsf{P}}^t(|\Lambda_t| \le \gamma') = \int_{(|\Lambda_t| \le \gamma')} \exp(\Lambda_t)\, d\mathsf{P}^t \ge e^{-\gamma'} \mathsf{P}^t(|\Lambda_t| \le \gamma')$$

for all $\gamma > 0$ and $\gamma' \in (0, \gamma)$, whence $\liminf_{t\to\infty} \widetilde{\mathsf{P}}^t(|\Lambda_t| \le \gamma) \ge e^{-\gamma'}$ by condition c). Passing to the limit as $\gamma' \to 0$ we prove that $\widetilde{\mathsf{P}}^t(|\Lambda_t| \le \gamma) \to 1$ as $t \to \infty$ for all $\gamma > 0$, that is, condition d) holds.

d) $\Rightarrow$ e). Let $\varepsilon \in (0,1)$ and $\gamma > 0$. Since $(|\Lambda_t| \le \gamma) \subset (\tilde{\jmath}_t > 0,\ \jmath_t > 0)$, it holds that

$$\begin{aligned} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) &\ge \mathsf{E}_Q^t\, \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon}\, I(|\Lambda_t| \le \gamma) \\ &= \mathsf{E}_Q^t\, \tilde{\jmath}_t z_t^{\varepsilon-1}\, I(|\Lambda_t| \le \gamma) \ge e^{(\varepsilon-1)\gamma} \widetilde{\mathsf{P}}^t(|\Lambda_t| \le \gamma). \end{aligned}$$

In view of d) this implies that

$$\liminf_{t\to\infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) \ge e^{(\varepsilon-1)\gamma} \tag{2.2.23}$$

for all $\varepsilon \in (0,1)$ and $\gamma > 0$. Passing to the limit in (2.2.23) as $\gamma \to 0$ and taking into account the inequality $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) \le 1$ we obtain condition e).

e) $\Rightarrow$ f). Follows from equality (2.2.7).

f) $\Rightarrow$ g). Follows from inequalities (2.2.6).

g) $\Rightarrow$ h). Follows from equality (2.2.10).

h) $\Rightarrow$ a). Let α be an arbitrary number of the interval $[0,1]$ and let $(\bar{\delta}_t)$ be an arbitrary family of tests $\bar{\delta}_t \in \Sigma^t$ such that $\alpha(\bar{\delta}_t) \to \alpha$ as $t \to \infty$. By Lemma 1.1.3 the set $\mathfrak{N}^t$ is symmetric about the point $(1/2, 1/2)$, so that it contains both points (α, β) and $(1-\alpha, 1-\beta)$. Thus

$$\inf\{\alpha(\delta_t) + \beta(\delta_t) \colon \delta_t \in \Sigma^t\} = 2 - \sup\{\alpha(\delta_t) + \beta(\delta_t) \colon \delta_t \in \Sigma^t\}. \tag{2.2.24}$$

Now we obtain from h) that

$$\lim_{t\to\infty} \sup\{\alpha(\delta_t) + \beta(\delta_t) \colon \delta_t \in \Sigma^t\} = 1. \tag{2.2.25}$$

Applying the inequality

$$\begin{aligned} \inf\{\alpha(\delta_t) + \beta(\delta_t) \colon \delta_t \in \Sigma^t\} &\le \alpha(\bar{\delta}_t) + \beta(\bar{\delta}_t) \\ &\le \sup\{\alpha(\delta_t) + \beta(\delta_t) \colon \delta_t \in \Sigma^t\}, \end{aligned} \tag{2.2.26}$$

condition h), and relation (2.2.25) we prove that $\beta(\bar{\delta}_t) \to 1 - \alpha$ as $t \to \infty$, that is, the implication (2.2.21) is proved. □

Definitions 2.2.8 and 2.2.9 are nonsymmetric with respect to $\alpha(\delta_t)$ and $\beta(\delta_t)$. Nevertheless using (2.2.24)–(2.2.26) one can prove the following result showing that (2.2.21) and (2.2.22) are, in fact, equivalent.

LEMMA 2.2.3. *If* $(H^t) \cong (\widetilde{H}^t)$, *then*

$$\lim_{t\to\infty} \beta(\delta_t) = \beta \;\Rightarrow\; \lim_{t\to\infty} \alpha(\delta_t) = 1 - \beta$$

for all numbers $\beta \in [0,1]$ *and all families* (δ_t) *of tests* $\delta_t \in \Sigma^t$ *such that the limit* $\lim_{t\to\infty} \beta(\delta_t)$ *exists.*

If $(\mathsf{P}^t) \cong (\widetilde{\mathsf{P}}^t)$, *then*

$$\lim_{t\to\infty} \widetilde{\mathsf{P}}^t(A_t) = \alpha \;\Rightarrow\; \lim_{t\to\infty} \mathsf{P}^t(A_t) = \alpha$$

for all numbers $\alpha \in [0,1]$ *and for all families* (A_t) *of sets* $A_t \in \mathscr{B}^t$ *such that the limit* $\lim_{t\to\infty} \widetilde{\mathsf{P}}^t(A_t)$ *exists.*

REMARK 2.2.4. Definition 2.2.8 in the case of $(H^t) \cong (\widetilde{H}^t)$ implies that

$$\mathfrak{N}^t \to \underline{\mathfrak{N}} \quad \text{as } t \to \infty,$$

that is, the set $\mathfrak{N}^t$ "approaches", as $t \to \infty$, the diagonal of the square

$$[0,1] \times [0,1]$$

joining its corners $(1,0)$ and $(0,1)$. The measures P^t and $\widetilde{\mathsf{P}}^t$ corresponding to the hypotheses H^t and $\widetilde{H}^t$, respectively, asymptotically coincide in this case (cf. (1.1.8)).

Contiguous families of hypotheses. Now we consider families of hypotheses (H^t) and $(\widetilde{H}^t)$ whose asymptotic behavior differs from the complete asymptotic indistinguishability $(H^t) \cong (\widetilde{H}^t)$ and complete asymptotic distinguishability $(H^t) \triangle (\widetilde{H}^t)$.

DEFINITION 2.2.10. We say that a *family of hypotheses* $(\widetilde{H}^t)$ *is contiguous to a family of hypotheses* (H^t) (denoted by $(\widetilde{H}^t) \lhd (H^t)$) if $\beta(\delta_t) \to 1$ as $t \to \infty$ for all tests $\delta_t \in \Sigma^t$ such that $\alpha(\delta_t) \to 0$ as $t \to \infty$. Otherwise, that is, if there exists a family (δ_t) of tests $\delta_t \in \Sigma^t$ such that

$$\lim_{t\to\infty} \alpha(\delta_t) = 0, \qquad \liminf_{t\to\infty} \beta(\delta_t) < 1,$$

we say that a *family of hypotheses* $(\widetilde{H}^t)$ *is noncontiguous to a family of hypotheses* (H_t) (denoted by $(\widetilde{H}^t) \overline{\lhd} (H_t)$).

DEFINITION 2.2.11. We say that a *family of measures* $(\widetilde{\mathsf{P}}^t)$ *is contiguous to a family* (P^t) (denoted by $(\widetilde{\mathsf{P}}^t) \lhd (\mathsf{P}^t)$) if $\widetilde{\mathsf{P}}^t(A_t) \to 0$ as $t \to \infty$ for all sets $A_t \in \mathscr{B}^t$ such that $\mathsf{P}^t(A_t) \to 0$ as $t \to \infty$. Otherwise, that is, if there exists a family (A_t) of sets $A_t \in \mathscr{B}^t$ such that

$$\lim_{t\to\infty} \mathsf{P}^t(A_t) = 0, \qquad \limsup_{t\to\infty} \widetilde{\mathsf{P}}^t(A_t) > 0,$$

we say that a *family of measures* $(\widetilde{\mathsf{P}}^t)$ *is noncontiguous to a family of measures* (P_t) (denoted by $(\widetilde{\mathsf{P}}^t) \overline{\lhd} (\mathsf{P}_t)$).

Let $(X^t, \mathscr{B}^t)$, $t \in \mathbf{R}_+$, be a family of measurable spaces, let $\mathbf{S}^t$ be a probability measure on $(X^t, \mathscr{B}^t)$, and let ζ_t, $t \in \mathbf{R}_+$, be a measurable function defined on $(X^t, \mathscr{B}^t)$ and assuming values in $\overline{\mathbf{R}} = [-\infty, \infty]$.

DEFINITION 2.2.12. We say that a *family* (ζ_t) *is dense with respect to a family of measures* $(\mathbf{S}^t)$ (denoted by $(\zeta_t | \mathbf{S}^t)$) if

$$\lim_{N\to\infty} \limsup_{t\to\infty} \mathbf{S}^t(|\zeta_t| > N) = 0.$$

DEFINITION 2.2.13. We say that a *family* (ζ_t) *is uniformly integrable with respect to a family of measures* $(\mathbf{S}^t)$ if

$$\lim_{N\to\infty} \sup_{t\in\mathbf{R}_+} \int I(|\zeta_t| > N)|\zeta_t|\, d\mathbf{S}^t = 0.$$

The characterization of $(\widetilde{H}^t) \lhd (H^t)$ is given in the following result.

THEOREM 2.2.3. *The following statements are equivalent*:

a) $(\widetilde{H}^t) \lhd (H^t)$;

b) $(\widetilde{\mathsf{P}}^t) \lhd (\mathsf{P}^t)$;

c) $\lim_{t\to\infty} \widetilde{\mathsf{P}}^t(z_t = \infty) = 0$ *and the family* (z_t) *is uniformly integrable with respect to the family of measures* (P^t);

d) $(z_t|\widetilde{\mathsf{P}}^t)$;

e) $(1/\mathfrak{z}_t|\widetilde{\mathsf{P}}^t)$;

f) $\lim_{\varepsilon\uparrow 1} \liminf_{t\to\infty} H(\varepsilon;\, \widetilde{\mathsf{P}}^t, \mathsf{P}^t) = 1$.

PROOF. a) $\Rightarrow$ b). Let $A_t \in \mathscr{B}^t$ and $\mathsf{P}^t(A_t) \to 0$ as $t \to \infty$. Then $\alpha(\bar{\delta}_t) \to 0$ as $t \to \infty$ for the test $\bar{\delta}_t = I(A_t)$. Condition a) implies that $\widetilde{\mathsf{P}}^t(A_t) = 1 - \beta(\bar{\delta}_t) \to 0$ as $t \to \infty$.

b) $\Rightarrow$ a). Let (δ_t) be an arbitrary family of tests such that $\alpha(\delta_t) \to 0$ as $t \to \infty$, and let ε be an arbitrary positive number. Put $A_t^\varepsilon = I(\delta_t \geq \varepsilon)$. Then

$$\mathsf{P}^t(A_t^\varepsilon) \leq \varepsilon^{-1} \int_{A_t^\varepsilon} \delta_t\, d\mathsf{P}^t \leq \varepsilon^{-1}\alpha(\delta_t),$$

whence it follows that $\mathsf{P}^t(A_t^\varepsilon) \to 0$ as $t \to \infty$ for all $\varepsilon > 0$. This together with condition b) implies that $\widetilde{\mathsf{P}}^t(A_t^\varepsilon) \to 0$ as $t \to \infty$ for all $\varepsilon > 0$. Since

$$1 - \beta(\delta_t) = \int_{A_t^\varepsilon} \delta_t\, d\widetilde{\mathsf{P}}^t + \int_{(A_t^\varepsilon)^c} \delta_t\, d\widetilde{\mathsf{P}}^t \leq \widetilde{\mathsf{P}}^t(A_t^\varepsilon) + \varepsilon$$

and ε is arbitrary, we deduce that $\beta(\delta_t) \to 1$ as $t \to \infty$.

b) $\Rightarrow$ c). Since $\mathsf{P}^t(z_t = \infty) = 0$, condition b) implies that $\widetilde{\mathsf{P}}^t(z_t = \infty) \to 0$ as $t \to \infty$. A family (z_t) is uniformly integrable with respect to (P^t) if and only if

$$\sup_{t\in\mathbf{R}_+} \int z_t\, d\mathsf{P}^t < \infty, \tag{2.2.27}$$

$$\text{if } \mathsf{P}^t(A_t) \to 0 \quad \text{for } A_t \in \mathscr{B}^t, \quad \text{then} \quad \int_{A_t} z_t\, d\mathsf{P}^t \to 0 \tag{2.2.28}$$

(see Lemma 2.6.2 in [47]). It follows from the Lebesgue decomposition that

$$\int_{A_t} z_t\, d\mathsf{P}^t \leq \widetilde{\mathsf{P}}^t(A_t) \leq 1.$$

This implies (2.2.27), while condition b) implies (2.2.28).

c) $\Rightarrow$ d). According to the Lebesgue decomposition

$$\widetilde{\mathsf{P}}^t(z_t > N) = \int_{(z_t>N)} z_t\, d\mathsf{P}^t + \widetilde{\mathsf{P}}^t(z_t = \infty).$$

This together with condition c) implies d).

d) $\Rightarrow$ b). Let $A_t \in \mathscr{B}^t$ be sets such that $\mathsf{P}^t(A_t) \to 0$ as $t \to \infty$. According to the Lebesgue decomposition

$$\begin{aligned}\widetilde{\mathsf{P}}^t(A_t) &= \widetilde{\mathsf{P}}^t(A_t \cap (z_t \le N)) = \widetilde{\mathsf{P}}^t(A_t \cap (z_t > N)) \\ &\le \int_{A_t \cap (z_t \le N)} z_t \, d\mathsf{P}^t + \widetilde{\mathsf{P}}^t(z_t > N) \le N\mathsf{P}^t(A_t) + \widetilde{\mathsf{P}}^t(z_t > N).\end{aligned}$$

Using the latter result and condition d) we prove that $\widetilde{\mathsf{P}}^t(A_t) \to 0$ as $t \to \infty$.

d) $\Leftrightarrow$ e). Follows from equality $z_t = 2/\mathfrak{z}_t - 1$.

e) $\Rightarrow$ f). Let $\gamma > 0$. Then for $\varepsilon \in (0,1)$

$$\begin{aligned}H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) &\ge \mathsf{E}_Q^t \tilde{\mathfrak{z}}_t \left(\frac{\mathfrak{z}_t}{\tilde{\mathfrak{z}}_t}\right)^{1-\varepsilon} I(\mathfrak{z}_t \ge \gamma, \tilde{\mathfrak{z}}_t > 0) = \widetilde{\mathsf{E}}^t(\mathfrak{z}_t/\tilde{\mathfrak{z}}_t)^{1-\varepsilon} I(\mathfrak{z}_t \ge \gamma) \\ &\ge (\gamma/2)^{1-\varepsilon} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma),\end{aligned}$$

since $\mathfrak{z}_t + \tilde{\mathfrak{z}}_t = 2$. Thus for all $\gamma > 0$

$$\lim_{\varepsilon \uparrow 1} \liminf_{t \to \infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) \ge \liminf_{\varepsilon \uparrow 1} \left(\frac{\gamma}{2}\right)^{1-\varepsilon} \liminf_{t \to \infty} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma) \ge \liminf_{t \to \infty} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma).$$

Condition e) implies that $\lim_{\gamma \downarrow 0} \lim_{t \to \infty} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma) = 1$, therefore the latter result and inequality $H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) \le 1$ prove condition f).

f) $\Rightarrow$ e). For all $\varepsilon, \lambda \in (0,1)$ and $\gamma > 0$ we obtain similarly to (2.2.18) that

$$\begin{aligned}H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) &\le \gamma^{1-\varepsilon} + \lambda^{\varepsilon} + \widetilde{\mathsf{E}}^t(\mathfrak{z}_t/\tilde{\mathfrak{z}}_t)^{1-\varepsilon} I(\mathfrak{z}_t \ge \gamma, \tilde{\mathfrak{z}}_t \ge \lambda) \\ &\le \gamma^{1-\varepsilon} + \lambda^{\varepsilon} + (2/\lambda)^{1-\varepsilon} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma).\end{aligned}$$

Then for all ε, $\lambda \in (0,1)$ we get

$$\liminf_{\gamma \downarrow 0} \liminf_{t \to \infty} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma) \ge \left(\frac{\lambda}{2}\right)^{1-\varepsilon} \liminf_{t \to \infty} H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) - \frac{\lambda}{2^{1-\varepsilon}}.$$

Passing to the limit in this inequality as $\varepsilon \uparrow 1$ and using condition f), then passing to the limit as $\lambda \downarrow 0$ we prove that

$$\liminf_{\gamma \downarrow 0} \liminf_{t \to \infty} \widetilde{\mathsf{P}}^t(\mathfrak{z}_t \ge \gamma) \ge 1,$$

whence condition e) follows. □

REMARK 2.2.5. If $(\widetilde{H}^t) \lhd (H^t)$ and $\mathfrak{N}^t \to \mathfrak{N}^\infty$ as $t \to \infty$, then according to Definition 2.2.10 the limit set $\mathfrak{N}^\infty$ does not contain any point of the interval of the straight line joining the points $(0,0)$ and $(0,1)$, except for the point $(0,1)$.

DEFINITION 2.2.14. If $(H^t) \lhd (\widetilde{H}_t)$ and $(\widetilde{H}^t) \lhd (H_t)$, then the families of hypotheses (H^t) and $(\widetilde{H}^t)$ are called *mutually contiguous* (denoted by $(H^t) \lhd\rhd (\widetilde{H}_t)$). If $(\widetilde{H}^t) \overline{\lhd} (H_t)$ and $(H^t) \overline{\lhd} (\widetilde{H}_t)$, then they are called *mutually noncontiguous* (denoted by $(H^t) \overline{\lhd\rhd} (\widetilde{H}_t)$). If either $(\widetilde{H}^t) \overline{\lhd} (H_t)$ or $(H^t) \overline{\lhd} (\widetilde{H}_t)$, then we say that the *families of hypotheses* (H^t) *and* $(\widetilde{H}^t)$ *are not mutually contiguous* (denoted by $(H^t) \overline{\lhd\rhd} (\widetilde{H}_t)$).

DEFINITION 2.2.15. If $(\widetilde{\mathsf{P}}^t) \lhd (\mathsf{P}_t)$ and $(\mathsf{P}^t) \lhd (\widetilde{\mathsf{P}}_t)$, then the families of measures (P^t) and $(\widetilde{\mathsf{P}}^t)$ are called *mutually contiguous* (denoted by $(\mathsf{P}^t) \lhd\rhd (\widetilde{\mathsf{P}}_t)$). If $(\widetilde{\mathsf{P}}^t) \overline{\lhd} (\mathsf{P}_t)$ and $(\mathsf{P}^t) \overline{\lhd} (\widetilde{\mathsf{P}}_t)$, then they are called *mutually noncontiguous* (denoted by $(\mathsf{P}^t) \overline{\lhd\rhd} (\widetilde{\mathsf{P}}_t)$). If either $(\widetilde{\mathsf{P}}^t) \overline{\lhd} (\mathsf{P}_t)$ or $(\mathsf{P}^t) \overline{\lhd} (\widetilde{\mathsf{P}}_t)$, then we say that the *families of measures* (P^t) *and* $(\widetilde{\mathsf{P}}^t)$ *are not mutually contiguous* (denoted by $(\mathsf{P}^t) \overline{\lhd\rhd} (\widetilde{\mathsf{P}}_t)$).

REMARK 2.2.6. Theorem 2.2.3 implies the characterization of all types

$$(\widetilde{H}^t) \overline{\lhd} (H^t), \qquad (H^t) \lhd\rhd (\widetilde{H}^t), \qquad (H^t) \overline{\lhd\rhd} (\widetilde{H}^t), \quad \text{and} \quad (H^t) \overline{\lhd\rhd} (\widetilde{H}^t).$$

For example,

$$(\widetilde{H}^t) \overline{\lhd} (H^t) \iff \lim_{\varepsilon \uparrow 1} \liminf_{t \to \infty} H\big(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t\big) < 1.$$

Further results on the contiguous families can be found in [**21, 22, 37, 45**].

The whole range of types of the asymptotic distinguishability. Using the notions of the asymptotic distinguishability of families of hypotheses we obtain the whole range of types of the asymptotic distinguishability of families of hypotheses (H^t) and $(\widetilde{H}^t)$. We will use the following conditions:

a_0) $(H^t) \cong (\widetilde{H}^t)$;
a) $(H^t) \lhd\rhd (\widetilde{H}^t)$;
b) $(H^t) \lhd (\widetilde{H}^t)$, $(\widetilde{H}^t) \overline{\lhd} (H^t)$;
c) $(H^t) \overline{\lhd} (\widetilde{H}^t)$, $(\widetilde{H}^t) \lhd (H^t)$;
d) $(H^t) \overline{\lhd\rhd} (\widetilde{H}^t)$, $(H^t) \overline{\Delta} (\widetilde{H}^t)$;
e) $(H^t) \Delta (\widetilde{H}^t)$.

DEFINITION 2.2.16. We say that the *the asymptotic distinguishability of families of hypotheses* (H^t) *and* $(\widetilde{H}^t)$ *is of type* $\mathbf{a}_0$ (respectively, of type **a**, **b**, **c**, **d**, or **e**), if condition a_0) (respectively a), b), c), d), or e)) holds.

Note that the types **a**, **b**, **c**, **d**, and **e** are disjoint and form the *whole range of types of the asymptotic distinguishability of families of hypotheses*. Since $\mathrm{a}_0) \Rightarrow$ a), type $\mathbf{a}_0$ is a subtype of type **a**. If the type $\mathbf{a}_1$ is defined as a subtype of the type **a** for which condition a_0) does not hold, then the types $\mathbf{a}_0$, $\mathbf{a}_1$, **b**, **c**, **d**, and **e** still form the whole range of disjoint types of the asymptotic distinguishability of families of hypotheses (H^t) and $(\widetilde{H}^t)$.

A characterization of types **e** and $\mathbf{a}_0$ is given in Theorems 2.2.1 and 2.2.2, respectively. A characterization of other types can easily be obtained by combining Theorems 2.2.1–2.2.3 and taking into account Remarks 2.2.3 and 2.2.6. We do not give this characterization and leave it to the reader.

EXAMPLE 2.2.1. Let an observation be the vector $\xi^{(n)} = (\xi_{n1}, \xi_{n2}, \dots, \xi_{nn})$ where $\xi_{n1}, \xi_{n2}, \dots, \xi_{nn}$ are independent random variables such that the distribution of ξ_{ni} is $\mathcal{N}(a_{ni}, 1)$ under the hypothesis H^n or $\mathcal{N}(\widetilde{a}_{ni}, 1)$ under the hypothesis $\widetilde{H}^n$. Then the likelihood ratio $z_n(x)$, $x \in \mathbf{R}^n$, is the density of the measure $\widetilde{\mathsf{P}}^n$ corresponding to the hypothesis $\widetilde{H}^n$ with respect to the measure P^t corresponding to the hypothesis H^n. The likelihood ratio is given by

$$z_n(x) = \exp\left(\sum_{i=1}^n (\widetilde{a}_{ni} - a_{ni}) x_i - \frac{1}{2} \sum_{i=1}^n \big(\widetilde{a}_{ni}^2 - a_{ni}^2\big)\right)$$

where $x = (x_1, x_2, \ldots, x_n)$. Put $\Lambda_n = \Lambda_n(\xi^{(n)})$ for $\Lambda_n(x) = \ln z_n(x)$. It is clear that

$$\mathcal{L}\left(() \Lambda_n | H^n\right) = \mathcal{N}\left(-\frac{1}{2}v_n^2, v_n^2\right), \tag{2.2.29}$$

$$\mathcal{L}\left(() \Lambda_n | \widetilde{H}^n\right) = \mathcal{N}\left(\frac{1}{2}v_n^2, v_n^2\right) \tag{2.2.30}$$

where

$$v_n^2 = \sum_{i=1}^{n} (\widetilde{a}_{ni} - a_{ni})^2 \tag{2.2.31}$$

and $\mathcal{L}\left(() \Lambda_n | H^n\right)$ is the distribution of Λ_n under the hypothesis H^n. From equality (2.2.29) we derive that

$$\mathsf{P}^n(\Lambda_n < N) = \Phi\left(\frac{v_n}{2} + \frac{N}{v_n}\right), \tag{2.2.32}$$

whence it follows by Theorem 2.2.1 that

$$(H^n) \triangle (\widetilde{H}^n) \iff \limsup_{n\to\infty} v_n = \infty. \tag{2.2.33}$$

Using equality (2.2.30) we obtain

$$\widetilde{\mathsf{P}}^n(\Lambda_n > N) = \Phi\left(\frac{v_n}{2} - \frac{N}{v_n}\right), \tag{2.2.34}$$

whence

$$\limsup_{n\to\infty} v_n < \infty \implies (\widetilde{H}^n) \lhd (H^n) \tag{2.2.35}$$

in view of Theorem 2.2.3. By contradiction, we derive from (2.2.33) that

$$(\widetilde{H}^n) \lhd (H^n) \implies \limsup_{n\to\infty} v_n < \infty. \tag{2.2.36}$$

Analogously, using (2.2.32) and Theorem 2.2.3 with the hypotheses H^n and $\widetilde{H}^n$ interchanged we obtain

$$(H^n) \lhd (\widetilde{H}^n) \iff \limsup_{n\to\infty} v_n < \infty. \tag{2.2.37}$$

Combining (2.2.35)–(2.2.37) we prove that

$$(H^n) \lhd\rhd (\widetilde{H}^n) \iff \limsup_{n\to\infty} v_n < \infty. \tag{2.2.38}$$

It follows from (2.2.33) and (2.2.38) that either $(H^n) \triangle (\widetilde{H}^n)$ or $(H^n) \lhd\rhd (\widetilde{H}^n)$. In other words, either the distinguishability is of type **a** or of type **e**. Moreover using Theorem 2.2.2 and relations (2.2.29) and (2.2.30) one can show that

$$(H^n) \cong (\widetilde{H}^n) \iff \limsup_{n\to\infty} v_n = 0.$$

Therefore either the distinguishability is of type $\mathbf{a}_0$ or of type $\mathbf{a}_1$ or of type $\mathbf{e}$. Namely

$$\text{type } \mathbf{a}_0 \iff \limsup_{n\to\infty} v_n = 0,$$
$$\text{type } \mathbf{a}_1 \iff 0 < \limsup_{n\to\infty} v_n < \infty,$$
$$\text{type } \mathbf{e} \iff \limsup_{n\to\infty} v_n = \infty.$$

Further results on types of the asymptotic distinguishability of families of hypotheses and various examples can be found in [**37**].

2.3. Complete asymptotic distinguishability under the strong law of large numbers

Consider the Neyman–Pearson test δ_t^{+,α_t} of level $\alpha_t \in [0,1]$ defined by (2.1.12) and assume that the complete asymptotic distinguishability $(H^t) \Delta (\widetilde{H}^t)$ holds. By Theorem 2.2.1, we have

$$(2.3.1) \qquad (H^t) \Delta (\widetilde{H}^t) \iff \liminf_{t\to\infty} \left[\alpha(\delta_t^{c,1}) + \beta(\delta_t^{c,1})\right] = 0 \quad \text{for all } c \in (0,\infty)$$

where $\delta_t^{c,1} = I(z_t \geq c)$ is the likelihood ratio test defined by (2.1.11).

This implies that if $(H^t) \Delta (\widetilde{H}^t)$, for any $c \in (0,\infty)$ there exists a sequence (t_n) for the test δ_t^{+,α_t} with $\alpha_t = \alpha(\delta_t^{c,1})$ such that $t_n \to \infty$, $\alpha_{t_n} \to 0$, and $\beta(\delta_t^{+,\alpha_{t_n}}) \to 0$ as $n \to \infty$. By equivalence (2.3.1), obtaining more refined properties of the function $\beta(\delta_{t_n}^{+,\alpha})$ requires that we consider a more stringent constraint than $(H^t) \Delta (\widetilde{H}^t)$.

Relative entropy and the law of large numbers. Consider the following condition:

$\Lambda 1$. $\lim_{t\to\infty} \mathsf{P}^t(|\chi_t^{-1}\Lambda_t + 1| > \gamma) = 0$ for any $\gamma > 0$ where χ_t is a nonrandom positive function such that $\chi_t \to \infty$ as $t \to \infty$.

It is easy to see that, in view of Theorem 2.2.1,

$$\Lambda 1 \Rightarrow (H^t) \Delta (\widetilde{H}^t),$$

that is, the complete asymptotic distinguishability holds under condition $\Lambda 1$.

DEFINITION 2.3.1. The number

$$(2.3.2) \qquad I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \mathsf{E}_Q^t \mathfrak{z}_t \ln(\mathfrak{z}_t/\tilde{\mathfrak{z}}_t)$$

is called *the relative entropy of a measure* P^t *with respect to a measure* $\widetilde{\mathsf{P}}^t$. We agree that $\mathfrak{z}_t \ln(\mathfrak{z}_t/\tilde{\mathfrak{z}}_t)$ equals 0 if $\mathfrak{z}_t = 0$ and equals ∞ if $\tilde{\mathfrak{z}}_t = 0$. Then the relative entropy is well defined for all t.

The relative entropy $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t)$ is often called the *Kullback–Leibler divergence*, or *distance* or *deviation* [**7, 9, 11, 33**], or the *Kullback–Leibler information* defined for the measures P^t and $\widetilde{\mathsf{P}}^t$ [**1, 33**].

LEMMA 2.3.1. *We have*

$$(2.3.3) \qquad I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \mathsf{E}^t \ln \tilde{z}_t = -\mathsf{E}^t \ln z_t$$

where z_t and $\widetilde{z}_t$ are the likelihood ratios defined by (2.1.2). *Moreover*

$$I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \widetilde{\mathsf{E}}^t \widetilde{z}_t \ln \widetilde{z}_t \tag{2.3.4}$$

if $\mathsf{P}^t \ll \widetilde{\mathsf{P}}^t$ *and* $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \infty$ *if* $\mathsf{P}^t \not\ll \widetilde{\mathsf{P}}^t$.

PROOF. The first equality in (2.3.3) follows immediately from definitions (2.3.2) and (2.1.2). The second equality in (2.3.3) holds, since $\widetilde{z}_t = z_t^{-1}$ (P^t-a.s.).

If $\mathsf{P}^t \ll \widetilde{\mathsf{P}}^t$, then $\widetilde{z}_t = d\mathsf{P}^t/d\widetilde{\mathsf{P}}^t$ ($\widetilde{\mathsf{P}}^t$-a.s.). Thus equality (2.3.4) follows from the first equality in (2.3.3). If $\mathsf{P}^t \not\ll \widetilde{\mathsf{P}}^t$, then $\mathsf{P}^t(\widetilde{\jmath}_t = 0) > 0$. Therefore

$$\mathsf{P}^t(\jmath_t > 0, \widetilde{\jmath}_t = 0) > 0$$

implying that $\mathbf{Q}^t(\jmath_t > 0, \widetilde{\jmath}_t = 0) > 0$. Then definition (2.3.2) gives

$$I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \infty. \qquad \square$$

LEMMA 2.3.2. *The relative entropy $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t)$ of a measure P^t with respect to a measure $\widetilde{\mathsf{P}}^t$ is nonnegative. Moreover $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = 0$ if and only if $\mathsf{P}^t = \widetilde{\mathsf{P}}^t$.*

PROOF. By Lemma 2.3.1, it is sufficient to consider the case $\mathsf{P}^t \ll \widetilde{\mathsf{P}}^t$ in view of equality (2.3.4). Put $\varphi(t) = t \ln t$. Equality (2.3.4) and the Jensen inequality imply that

$$I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) = \widetilde{\mathsf{E}}^t \varphi(\widetilde{z}_t) \geq \varphi(\widetilde{\mathsf{E}}^t \widetilde{z}_t) = \varphi(1) = 0$$

where the inequality becomes an equality if and only if $\widetilde{z}_t = \text{const} = a$ ($\widetilde{\mathsf{P}}^t$-a.s.). Note that $a = \widetilde{\mathsf{E}}^t \widetilde{z}_t = 1$ in this case. This means that if the equality holds, we have for any $A \in \mathscr{B}^t$

$$\mathsf{P}^t(A) = \int_A \widetilde{z}_t d\widetilde{\mathsf{P}}^t = \int_A d\widetilde{\mathsf{P}}^t = \widetilde{\mathsf{P}}^t(A),$$

that is, $\mathsf{P}^t = \widetilde{\mathsf{P}}^t$. $\square$

REMARK 2.3.1. Condition $\Lambda 1$ is known as *relative stability* of Λ_t as $t \to \infty$ and is the most natural and general form of the law of large numbers [**20**]. If the relative entropy $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t)$ is finite for any $t \in \mathbf{R}_+$ and $I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t) \to \infty$ as $t \to \infty$, then, by putting $\chi_t = I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t)$, condition $\Lambda 1$ can be transformed into the following form resembling the law of large numbers:

$$\lim_{t\to\infty} \mathsf{P}^t \left(\left| \frac{\Lambda_t - \mathsf{E}^t \Lambda_t}{\chi_t} \right| > \varepsilon \right) = 0 \quad \text{for all } \varepsilon > 0$$

where, by (2.3.3), we have $\mathsf{E}^t \Lambda_t = -I(\mathsf{P}^t|\widetilde{\mathsf{P}}^t)$.

The following result describes the behavior of the points $(\overline{\alpha}_t, 0)$ and $(0, \overline{\beta}_t)$ of the set $\mathfrak{N}^t$ under the law of large numbers.

LEMMA 2.3.3. *The following relations hold*:

$$\Lambda 1 \Rightarrow \lim_{t\to\infty} \overline{\alpha}_t = 1, \tag{2.3.5}$$

$$\Lambda 1 \Rightarrow \liminf_{t\to\infty} \chi_t^{-1} \ln \overline{\beta}_t \geq -1. \tag{2.3.6}$$

PROOF. Implication (2.3.5) follows from (2.1.4) and the estimate

$$\mathsf{P}^t(\Lambda_t < -a\chi_t) \geq \mathsf{P}^t(\Lambda_t = -\infty) = \mathsf{P}^t(z_t = 0)$$

where $1 < a < \infty$. By the Lebesgue decomposition (2.1.7) and in view of equality (2.1.5), we obtain for any $a \in (1, \infty)$ that

$$\mathsf{P}^t(\Lambda_t \geq -a\chi_t) \leq e^{a\chi_t}\mathsf{E}^t e^{\Lambda_t}\chi(\Lambda_t \geq -a\chi_t) \leq e^{a\chi_t}\mathsf{E}^t z_t = e^{a\chi_t}\widetilde{\mathsf{P}}^t(z_t < \infty) = e^{a\chi_t}\overline{\beta}_t$$

which together with condition $\Lambda 1$ yields

$$\liminf_{t\to\infty} \chi_t^{-1} \ln \overline{\beta}_t \geq -a.$$

Approaching the limit as $a \to 1$, we obtain implication (2.3.6). □

REMARK 2.3.2. By equality (1.1.50), implication (2.3.6) is equivalent to

$$\Lambda 1 \Rightarrow \liminf_{t\to\infty} \chi_t^{-1} \ln \beta(\delta_t^{+,0}) \geq -1.$$

Behavior of the Neyman–Pearson test under the law of large numbers. Introduce the following conditions:

$\alpha 1)$ $\liminf_{t\to\infty} \alpha_t > 0$; $\alpha 2)$ $\limsup_{t\to\infty} \alpha_t < 1$;

$d1)$ $\limsup_{t\to\infty} \chi_t^{-1} d_t \leq -1$; $d2)$ $\liminf_{t\to\infty} \chi_t^{-1} d_t \geq -1$;

$\beta 1)$ $\limsup_{t\to\infty} \chi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \leq -1$; $\beta 2)$ $\liminf_{t\to\infty} \chi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \geq -1$,

where χ_t is the normalizing term appearing in the law of large numbers $\Lambda 1$.

The following result establishes relationships between the behavior of α_t, d_t, and $\beta(\delta_t^{+,\alpha_t})$ under condition $\Lambda 1$.

THEOREM 2.3.1. *The following implications hold*:

$$\Lambda 1, \alpha 1 \Rightarrow d1 \Rightarrow \beta 1, \tag{2.3.7}$$

$$\Lambda 1, \alpha 2 \Rightarrow \beta 2 \Rightarrow d2. \tag{2.3.8}$$

PROOF. The implication $\Lambda 1, \alpha 1 \Rightarrow d1$ is proved by contradiction. Representation (2.1.12) implies

$$\alpha_t = \mathsf{P}^t(Y_t > y_t) + \varepsilon_t \mathsf{P}^t(Y_t = y_t) \tag{2.3.9}$$

where $Y_t = \chi_t^{-1}\Lambda_t$ and $y_t = \chi_t^{-1} d_t$.

Let (t_n) be a sequence such that $t_n \to \infty$ and $y_{t_n} \to \overline{y} = \limsup_{t\to\infty} y_t$ as $n \to \infty$. Assume that $\underline{\alpha} = \liminf_{t\to\infty} \alpha_t > 0$ and condition $\Lambda 1$ holds, but $\overline{y} > -1$. Then, by condition $\Lambda 1$, we obtain that $\mathsf{P}^t(Y_{t_n} \geq y_{t_n}) \to 0$ as $n \to \infty$. On the other hand, the inequality $\underline{\alpha} > 0$ and equality (2.3.9) imply that

$$\lim_{n\to\infty} \mathsf{P}^{t_n}(Y_{t_n} \geq y_{t_n}) \geq \liminf_{n\to\infty} \alpha_{t_n} \geq \liminf_{t\to\infty} \alpha_t = \underline{\alpha} > 0.$$

The contradiction we have obtained proves the implication $\Lambda 1, \alpha 1 \Rightarrow d1$.

Using equality (2.1.8) we obtain that

$$\begin{aligned}\beta(\delta_t^{+,\alpha_t}) &= \widetilde{\mathsf{E}}^t(1-\delta_t^{+,\alpha_t}) = \mathsf{E}^t(1-\delta_t^{+,\alpha_t})z_t + \widetilde{\mathsf{E}}^t(1-\delta_t^{+,\alpha_t})\,I(z_t=\infty)\\ &= \mathsf{E}^t(1-\delta_t^{+,\alpha_t})z_t \le e^{d_t},\end{aligned} \tag{2.3.10}$$

proving the implications $d1 \Rightarrow \beta 1$ and $\beta 2 \Rightarrow d2$.

Now we prove the implication $\Lambda 1, \alpha 2 \Rightarrow \beta 2$. Let $a > 1$ and $\gamma > 0$ be arbitrary numbers and let $S_t = \{\chi_t^{-1}\Lambda_t \ge -a\}$. In view of condition $\Lambda 1$, there exists a number $t_0 = t_0(a,\gamma)$ depending on a and γ such that $\mathsf{P}^t(S_t^c) < \gamma$ for all $t > t_0$. Therefore we obtain from (2.1.8) for $t > t_0$ that

$$\begin{aligned}\beta(\delta_t^{+,\alpha_t}) &= \mathsf{E}^t\, z_t(1-\delta_t^{+,\alpha_t}) \ge \mathsf{E}^t\, I(S_t)\, z_t(1-\delta_t^{+,\alpha_t})\\ &\ge e^{-a\chi_t}\mathsf{E}^t\, I(S_t)(1-\delta_t^{+,\alpha_t}) \ge e^{-a\chi_t}(1-\alpha_t-\gamma).\end{aligned} \tag{2.3.11}$$

In view of condition $\alpha 2$ and since $a > 1$ and $\gamma > 0$ are arbitrary, relation (2.3.11) proves the implication $\Lambda 1, \alpha 2 \Rightarrow \beta 2$. □

Implications (2.3.7) and (2.3.8) give the following result.

COROLLARY 2.3.1. *If the law of large numbers* $\Lambda 1$ *holds and the level* α_t *satisfies conditions* $\alpha 1$ *and* $\alpha 2$, *then*

$$\lim_{t\to\infty} \chi_t^{-1} \ln \beta(\delta_t^{+,\alpha_t}) = -1. \tag{2.3.12}$$

Corollary 2.3.1 implies that if the law of large numbers $\Lambda 1$ holds for the Neyman–Pearson test δ_t^{+,α_t} whose level α_t tends to a limit $\alpha \in (0,1)$ as $t \to \infty$, then the rate of decay of the type II error probability $\beta(\delta_t^{+,\alpha_t})$ does not depend on α. More specifically, (2.3.12) means that

$$\beta(\delta_t^{+,\alpha_t}) = \exp(-b\chi_t(1+o(1))), \qquad t \to \infty, \tag{2.3.13}$$

where $b = 1$ for all $\alpha \in (0,1)$.

Independent observations. Stein's lemma. Assume that an observation is the vector $\xi^{(n)} = (\xi_{n1}, \xi_{n2}, \dots, \xi_{nn})$, $n = 1, 2, \dots$, where $\xi_{n1}, \xi_{n2}, \dots, \xi_{nn}$ are independent random variables. The hypothesis H^n is that ξ_{ni} has a distribution P_{ni} with the density $p_{ni}(x)$ with respect to the Lebesgue measure, while the hypothesis $\widetilde{H}^n$ is that the distribution of ξ_{ni} is $\widetilde{\mathsf{P}}_{ni}$ with the density $\widetilde{p}_{ni}(x)$. Then the distribution of the vector $\xi^{(n)}$ under the hypothesis H^n is $\mathsf{P}^n = \mathsf{P}_{n1} \times \mathsf{P}_{n2} \times \cdots \times \mathsf{P}_{nn}$, and the density of this distribution with respect to the Lebesgue measure is

$$p_n(x) = \prod_{i=1}^n p_{ni}(x_i), \qquad x = (x_1, \dots, x_n).$$

Similarly, the distribution of the vector $\xi^{(n)}$ under the hypothesis $\widetilde{H}^n$ is

$$\widetilde{\mathsf{P}}^n = \widetilde{\mathsf{P}}_{n1} \times \widetilde{\mathsf{P}}_{n2} \times \cdots \times \widetilde{\mathsf{P}}_{nn}$$

and the density with respect to the Lebesgue measure equals

$$\widetilde{p}_n(x) = \prod_{i=1}^n \widetilde{p}_{ni}(x_i), \qquad x = (x_1, \dots, x_n).$$

The likelihood ratio $z_n(x)$ for the measures $\widetilde{\mathsf{P}}^n$ and P^n is

$$z_n(x) = \prod_{i=1}^{n} z_{ni}(x_i), \qquad z_{ni}(x_i) = \frac{\widetilde{p}_{ni}(x_i)}{p_{ni}(x_i)}, \quad x = (x_1, \dots, x_n), \tag{2.3.14}$$

where we agree that $0/0 = 0$. It is clear that the relative entropies $I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n)$ and $I(\mathsf{P}_{ni}|\widetilde{\mathsf{P}}_{ni})$, $i = 1, 2, \dots, n$, are related as follows:

$$I\left(\mathsf{P}^n|\widetilde{\mathsf{P}}^n\right) = \sum_{i=1}^{n} I\left(\mathsf{P}_{ni}|\widetilde{\mathsf{P}}_{ni}\right). \tag{2.3.15}$$

Let

$$\Lambda_n(x) = \ln z_n(x), \qquad \lambda_{ni}(x_i) = \ln z_{ni}(x_i), \quad i = 1, 2, \dots, n, \tag{2.3.16}$$

where we agree that $\ln 0 = -\infty$. If

$$\Lambda_n = \Lambda_n\left(\xi^{(n)}\right) \quad \text{and} \quad \lambda_{ni} = \lambda_{ni}(\xi_{ni}), \ i = 1, 2, \dots, n,$$

then $\Lambda_n = \sum_{i=1}^{n} \lambda_{ni}$. It follows from (2.3.3) that

$$I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) = -\mathsf{E}^n\Lambda_n, \qquad I(\mathsf{P}_{ni}|\widetilde{\mathsf{P}}_{ni}) = -\mathsf{E}_{ni}\lambda_{ni} \tag{2.3.17}$$

where E^n and E_{ni} are mathematical expectations with respect to the distributions P^n and P_{ni}, respectively.

Corollary 2.3.1 can be stated in the following form.

COROLLARY 2.3.2. *Let* $I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) < \infty$ *for all* $n = 1, 2, \dots$, *and let*

$$I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) \to \infty \quad \text{as } n \to \infty.$$

If the law of large numbers $\Lambda 1$ *holds for* Λ_n, $n = 1, 2, \dots$, *with* $\chi_n = I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n)$ *and the level* α_n *satisfies conditions* $\alpha 1$ *and* $\alpha 2$, *then*

$$\lim_{n\to\infty} \chi_n^{-1} \ln \beta(\delta_n^{+,\alpha_n}) = -1. \tag{2.3.18}$$

Now let $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ where $\xi_1, \xi_2, \dots, \xi_n$ are independent identically distributed random variables such that the distribution of ξ_i under the hypothesis H^n is P with the density $p(x)$ with respect to the Lebesgue measure, while the distribution of ξ_i under the hypothesis $\widetilde{H}^n$ is $\widetilde{\mathsf{P}}$ with the density $\widetilde{p}(x)$. We assume that the distribution of ξ_i is independent of n both under H^n and under $\widetilde{H}^n$. Then (2.3.14) and (2.3.16) can be rewritten as

$$z_n(x) = \prod_{i=1}^{n} z(x_i), \qquad z(x_i) = \frac{\widetilde{p}(x_i)}{p(x_i)}, \tag{2.3.19}$$

$$\Lambda_n(x) = \ln z_n(x), \qquad \lambda(x_i) = \ln z(x_i) \tag{2.3.20}$$

where $x = (x_1, \dots, x_n)$. Put $\Lambda_n = \Lambda_n(\xi^{(n)})$ and $\lambda_i = \lambda(\xi_i)$, $i = 1, 2, \dots, n$. Then $\Lambda_n = \sum_{i=1}^{n} \lambda_i$ and (2.3.15) and (2.3.17) imply

$$I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) = nI(\mathsf{P}|\widetilde{\mathsf{P}}), \qquad I(\mathsf{P}|\widetilde{\mathsf{P}}) = -\mathsf{E}\lambda_1. \tag{2.3.21}$$

Therefore the following classical result follows from Corollary 2.3.2.

COROLLARY 2.3.3. *Let* $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ *where* $\xi_1, \dots, \xi_n$ *are independent identically distributed random variables both under the hypothesis* H^n *and under the hypothesis* $\widetilde{H}^n$. *Assume that the distributions of* ξ_n *are independent of* n *and the relative entropy* $I(\mathsf{P}|\widetilde{\mathsf{P}})$ *is positive and finite. If* $\alpha_n \to \alpha \in (0,1)$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty} \frac{1}{n} \ln \beta(\delta_n^{+,\alpha_n}) = -I(\mathsf{P}|\widetilde{\mathsf{P}}). \tag{2.3.22}$$

PROOF. It follows from (2.3.19) and (2.3.20) that $\Lambda_n = \sum_{i=1}^n \lambda_i$ where the random variables $\lambda_1, \lambda_2, \dots, \lambda_n$ are independent and identically distributed under H^n with mean $\mathsf{E}\lambda_i = -I(\mathsf{P}|\widetilde{\mathsf{P}})$. By the Khintchine law of large numbers for sums of independent identically distributed random variables [**20**], we obtain that condition $\Lambda 1$ holds with $\chi_n = nI(\mathsf{P}|\widetilde{\mathsf{P}})$. Since $\alpha_n \to \alpha \in (0,1)$ as $n \to \infty$, the assumptions of Corollary 2.3.2 hold. Therefore (2.3.22) follows from (2.3.18). □

Corollary 2.3.3 is proved by Rao [**43**]. Corollary 2.3.3 with $\alpha_n = \alpha \in (0,1)$ for all n is called the *Stein lemma* (see [**1, 33**]).

EXAMPLE 2.3.1. Let an observation be $\xi^{(n)} = (\xi_{n1}, \xi_{n2}, \dots, \xi_{nn})$ where ξ_{n1}, $\xi_{n2}, \dots, \xi_{nn}$ are independent random variables such that the distribution of ξ_{ni} is $\mathcal{N}(a_{ni}, 1)$ under the hypothesis H^n and $\mathcal{N}(\widetilde{a}_{ni}, 1)$ under the hypothesis $\widetilde{H}^n$ (see Example 2.2.1). Using the notation of Example 2.2.1 we obtain from (2.2.29)–(2.2.31) that

$$\Lambda_n = v_n \eta - \frac{1}{2} v_n^2, \tag{2.3.23}$$

$$\Lambda_n = v_n \widetilde{\eta} + \frac{1}{2} v_n^2 \tag{2.3.24}$$

where

$$\mathcal{L}(\eta|H^n) = \mathcal{N}(0,1), \qquad \mathcal{L}\left(\widetilde{\eta}|\widetilde{H}^n\right) = \mathcal{N}(0,1). \tag{2.3.25}$$

Assume that the complete asymptotic distinguishability $(H^n) \triangle (\widetilde{H}^n)$ holds. Considering (2.2.33) we assume without loss of generality that $v_n \to \infty$ as $n \to \infty$. It follows from (2.3.23)–(2.3.25) that

$$I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) = I(\widetilde{\mathsf{P}}^n|\mathsf{P}^n) = \frac{1}{2} v_n^2. \tag{2.3.26}$$

Since $v_n \to \infty$ as $n \to \infty$, relation (2.3.23) implies that the law of large numbers holds with $\chi_n = v_n^2/2$. Therefore Corollary 2.3.2 holds if conditions $\alpha 1$ and $\alpha 2$ are satisfied. Note that (2.2.32) yields

$$\alpha(\delta_n^{+,\alpha_n}) = \mathsf{P}^n(\Lambda_n > d_n) = 1 - \Phi\left(\frac{v_n}{2} + \frac{d_n}{v_n}\right), \tag{2.3.27}$$

whence

$$d_n = -\frac{1}{2} v_n^2 + v_n t_{1-\alpha_n} \tag{2.3.28}$$

where t_p is the p-quantile of the distribution $\mathcal{N}(0,1)$. This implies that conditions $\alpha 1$ and $\alpha 2$ hold if the parameter d_n of the test δ_n^{+,α_n} is such that

$$\liminf_{n\to\infty}\left(\frac{d_n}{v_n}+\frac{v_n}{2}\right)>-\infty,\qquad \limsup_{n\to\infty}\left(\frac{d_n}{v_n}+\frac{v_n}{2}\right)<\infty. \tag{2.3.29}$$

Therefore Corollary 2.3.2 implies that

$$\beta(\delta_n^{+,\alpha_n})=\exp\left\{-\frac{1}{2}v_n^2(1+o(1))\right\} \tag{2.3.30}$$

as $n\to\infty$ if conditions (2.3.29) hold.

Moreover $\alpha_n\to\alpha\in(0,1)$ as $n\to\infty$ if conditions (2.3.27) and (2.3.28) hold and if

$$d_n=-\frac{1}{2}v_n^2+v_nt_{1-\alpha}+o(v_n). \tag{2.3.31}$$

It follows from (2.2.34) and (2.3.28) that

$$\beta(\delta_n^{+,\alpha_n})=\widetilde{\mathrm{P}}^n(\Lambda_n<d_n)=1-\Phi(v_n-t_{1-\alpha_n}). \tag{2.3.32}$$

By the well-known asymptotic expansion

$$1-\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-z^2/2}\,dz=\frac{1}{x\sqrt{2\pi}}e^{-x^2/2}(1+o(1)) \tag{2.3.33}$$

as $x\to\infty$ (see [**34**]) we obtain (2.3.30) from equality (2.3.32) if condition (2.3.31) holds. If we put $\alpha_n=\alpha\in(0,1)$ for all n (as in the Stein lemma), then (2.3.32) and (2.3.33) imply a stronger result:

$$\beta(\delta_n^{+,\alpha})=\frac{1}{\sqrt{2\pi}v_n}\exp\left(-\frac{1}{2}v_n^2+t_{1-\alpha}v_n-\frac{1}{2}t_{1-\alpha}^2+o(1)\right) \tag{2.3.34}$$

or, in other words,

$$\ln\beta(\delta_n^{+,\alpha})=-\frac{1}{2}v_n^2+t_{1-\alpha}v_n-\ln v_n-\frac{1}{2}t_{1-\alpha}^2-\ln\sqrt{2\pi}+o(1). \tag{2.3.35}$$

It is clear from asymptotic expansion (2.3.35) that the dependence on α in the behavior of $\beta(\delta_t^{+,\alpha})$ shows up in the second term of the asymptotic expansion of $\ln\beta(\delta_n^{+,\alpha})$ only, while the first term is independent of α (cf. (2.3.13)).

Large deviations. Implications (2.3.7) and (2.3.8) show that under condition $\Lambda 1$, relations $\beta 1$ and $\beta 2$ require conditions $\alpha 1$ and $\alpha 2$ to be satisfied; the latter conditions prohibit the level α_t to approach 0 and 1, respectively. However, if we impose a more restrictive condition on the likelihood ratio z_t, relations $\beta 1$ and $\beta 2$ can also be obtained for levels α_t that tend to 0 or 1 as $t\to\infty$, but rather slowly. To be more specific, let us introduce the following conditions:

$\alpha 1'$) $\lim_{t\to\infty}\chi_t^{-1}\ln\alpha_t=0$;

$\alpha 2'$) $\lim_{t\to\infty}\chi_t^{-1}\ln(1-\alpha_t)=0$;

$\Lambda 2$) $\limsup_{\varepsilon\downarrow 0}\limsup_{t\to\infty}\varepsilon^{-1}\chi_t^{-1}\ln H_t(\varepsilon)\le -1$;

$\Lambda 3$) $\liminf_{\varepsilon\uparrow 0}\liminf_{t\to\infty}\varepsilon^{-1}\chi_t^{-1}\ln H_t(\varepsilon)\ge -1$

where χ_t is the normalization occurring in condition $\Lambda 1$ and $H_t(\varepsilon) = H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t)$ is the Hellinger integral of order ε for the measures $\widetilde{\mathsf{P}}^t$ and P^t.

Observe that the definition of $H_t(\varepsilon)$ gives

$$H_t(\varepsilon) = H(\varepsilon; \widetilde{\mathsf{P}}^t, \mathsf{P}^t) = \begin{cases} \mathsf{P}^t(\tilde{\jmath}_t > 0), & \varepsilon = 0, \\ \widetilde{\mathsf{P}}^t(\jmath_t > 0), & \varepsilon = 1, \\ \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon}, & \varepsilon \neq 0,\ \varepsilon \neq 1. \end{cases} \tag{2.3.36}$$

Introduce the following notation:

$$\varepsilon^t_- = \inf\{\varepsilon: H_t(\varepsilon) > -\infty\}, \qquad \varepsilon^t_+ = \sup\{\varepsilon: H_t(\varepsilon) < \infty\}. \tag{2.3.37}$$

It is clear that $\varepsilon^t_- \leq 0$ and $\varepsilon^t_+ \geq 1$, since $H_t(\varepsilon) \leq 1$ for $\varepsilon \in [0,1]$. The following result gives a useful representation for the Hellinger integral $H_t(\varepsilon)$ in terms of the likelihood ratios z_t and $\tilde{z}_t$.

LEMMA 2.3.4. *For any $\varepsilon \in (\varepsilon^t_-, \varepsilon^t_+)$ different from 0 and 1,*

$$H_t(\varepsilon) = \mathsf{E}^t z_t^{\varepsilon} = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon}. \tag{2.3.38}$$

PROOF. If $0 < \varepsilon < 1$, then

$$H_t(\varepsilon) = \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\jmath_t > 0) = \mathsf{E}^t z_t^{\varepsilon} I(\jmath_t > 0) = \mathsf{E}^t z_t^{\varepsilon},$$
$$H_t(\varepsilon) = \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0) = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0) = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon}$$

proving equalities (2.3.38) for $0 < \varepsilon < 1$.

Now let $\varepsilon^t_- < 0$ and $\varepsilon \in (\varepsilon^t_-, 0)$. Since $H_t(\varepsilon) < \infty$ by (2.3.37), we have $\mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t = 0, \jmath_t > 0) < \infty$. This implies that $\mathbf{Q}(\tilde{\jmath}_t = 0, \jmath_t > 0) = 0$ and therefore $\mathsf{P}^t(\tilde{\jmath}_t = 0)$. Thus

$$H_t(\varepsilon) = \mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0) = \mathsf{E}^t (\jmath_t/\tilde{\jmath}_t)^{-\varepsilon} = \mathsf{E}^t z_t^{\varepsilon},$$
$$H_t(\varepsilon) = \mathsf{E}^t_Q \tilde{\jmath}_t \tilde{\jmath}_t^{\varepsilon-1} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0) = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0) = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon}.$$

Therefore equalities (2.3.38) are proved for $\varepsilon \in (\varepsilon^t_-, 0)$ and $\varepsilon^t_- < 0$.

Finally let $\varepsilon^t_+ > 1$ and $\varepsilon \in (1, \varepsilon^t_+)$. Since $H_t(\varepsilon) < \infty$, we also have

$$\mathsf{E}^t_Q \tilde{\jmath}_t^{\varepsilon} \jmath_t^{1-\varepsilon} I(\tilde{\jmath}_t > 0, \jmath_t = 0) < \infty.$$

Then $\mathbf{Q}(\tilde{\jmath}_t > 0, \jmath_t = 0) = 0$ and hence $\widetilde{\mathsf{P}}^t(\jmath_t = 0)$. Therefore

$$H_t(\varepsilon) = \widetilde{\mathsf{E}}^t_Q \tilde{\jmath}_t^{\varepsilon-1} \jmath_t^{1-\varepsilon} I(\jmath_t > 0) = \widetilde{\mathsf{E}}^t (\tilde{\jmath}_t/\jmath_t)^{\varepsilon-1} = \widetilde{\mathsf{E}}^t \tilde{z}_t^{1-\varepsilon},$$
$$H_t(\varepsilon) = \mathsf{E}^t_Q \jmath_t \tilde{\jmath}_t^{\varepsilon} \jmath_t^{-\varepsilon} I(\jmath_t > 0) = \mathsf{E}^t z_t^{\varepsilon} I(\jmath_t > 0) = \mathsf{E}^t z_t^{\varepsilon}.$$

This proves equalities (2.3.38) for $\varepsilon \in (1, \varepsilon^t_+)$ and $\varepsilon^t_+ > 1$. □

REMARK 2.3.3. If $H_t(\varepsilon_-^t) < \infty$, then the same argument as that used to prove Lemma 2.3.4 shows that equality (2.3.38) holds for $\varepsilon = \varepsilon_-^t$, too. If $H_t(\varepsilon_+^t) < \infty$, then equality (2.3.38) holds for $\varepsilon = \varepsilon_+^t$, too. Put $\mathsf{E}^t z_t^\varepsilon = \mathsf{P}^t(\tilde{\jmath}_t > 0)$ in the case $\varepsilon = 0$ and $\mathsf{P}^t(\tilde{\jmath}_t = 0) > 0$ and put $\widetilde{\mathsf{E}}^t \tilde{z}_t^\varepsilon = \widetilde{\mathsf{P}}^t(\jmath_t > 0)$ in the case $\varepsilon = 0$ and $\widetilde{\mathsf{P}}^t(\jmath_t = 0) > 0$. Then equality (2.3.38) holds for any $\varepsilon \in (-\infty, \infty)$. Moreover we get for any $\varepsilon \in (-\infty, \infty)$

$$H_t(\varepsilon) = \mathsf{E}^t e^{\varepsilon \Lambda_t}, \tag{2.3.39}$$

that is, $H_t(\varepsilon)$ is the moment generating function of the random variable Λ_t which is, generally speaking, an extended random variable since $\mathsf{P}^t(\Lambda_t = -\infty)$ can be positive.

REMARK 2.3.4. If $\varepsilon_-^t < 0$, then the proof of Lemma 2.3.4 shows that

$$\mathsf{P}^t(\tilde{\jmath}_t = 0) = 0.$$

By (1.1.18), this means that $\mathsf{P}^t \ll \widetilde{\mathsf{P}}^t$. Moreover $H_t(0) = 1$ by (2.3.36) in this case. If $\varepsilon_+^t > 1$, then it has also been shown in the proof of Lemma 2.3.4 that $\widetilde{\mathsf{P}}^t(\jmath_t = 0) = 0$. By (1.1.17), this means that $\widetilde{\mathsf{P}}^t \ll \mathsf{P}^t$. Relation (2.3.36) implies in this case that $H_t(1) = 1$.

REMARK 2.3.5. Condition $\Lambda 3$ implies that there exist numbers $\varepsilon_0 < 0$ and $t_0 < \infty$ such that $H_t(\varepsilon) < \infty$ for all $\varepsilon \in (\varepsilon_0, 0)$ and $t > t_0$. Then we have by Remark 2.3.4 that

$$\Lambda 3 \Rightarrow \mathsf{P}^t \ll \widetilde{\mathsf{P}}^t \quad \text{for all } t > t_0.$$

In view of (1.1.18), this implies

$$\Lambda 3 \Rightarrow \overline{\alpha}_t = 1 \quad \text{for all } t > t_0.$$

The following result establishes a relationship between conditions $\Lambda 2$, $\alpha 1$, $\alpha 2$ and conditions $\Lambda 2$, $\Lambda 3$, $\alpha 1'$, $\alpha 2'$.

LEMMA 2.3.5. *We have*

$$\alpha 1 \Rightarrow \alpha 1'; \qquad \alpha 2 \Rightarrow \alpha 2'; \tag{2.3.40}$$

$$\Lambda 2, \Lambda 3 \Rightarrow \Lambda 1. \tag{2.3.41}$$

PROOF. Implications (2.3.40) are obvious. To prove implication (2.3.41) let $\gamma > 0$ be an arbitrary number. Then, by condition $\Lambda 2$, there exists a positive number $\varepsilon_0 = \varepsilon_0(\gamma)$ such that

$$\limsup_{t \to \infty} \chi_t^{-1} \ln H_t(\varepsilon) \le -\varepsilon + \frac{1}{2}\varepsilon\gamma \tag{2.3.42}$$

for all $\varepsilon \in (0, \varepsilon_0)$. Fix some $\varepsilon \in (0, \varepsilon_0)$. Then, in view of (2.3.42), there exists a number $t_0 = t_0(\varepsilon, \gamma)$ such that for all $t > t_0$

$$\chi_t^{-1} \ln H_t(\varepsilon) \le -\varepsilon + \frac{1}{2}\varepsilon\gamma + \frac{1}{4}\varepsilon\gamma,$$

whence we obtain for $t > t_0$ that

$$\begin{aligned} \mathsf{P}^t\left(\chi_t^{-1}\Lambda_t > -1+\gamma\right) &= \mathsf{P}^t\left(z_t^\varepsilon > e^{-\varepsilon(1-\gamma)\chi_t}\right) \le e^{\varepsilon(1-\gamma)\chi_t}\mathsf{E}^t\, z_t^\varepsilon \\ &= e^{\varepsilon(1-\gamma)\chi_t} H_t(\varepsilon) \le \exp\left(-\frac{1}{4}\varepsilon\gamma\chi_t\right) \end{aligned} \tag{2.3.43}$$

by the Chebyshev inequality and equality (2.3.38) for $\varepsilon \in (0,1)$. Following a similar argument, we obtain from condition $\Lambda 3$ that for any $\gamma > 0$ there exists $\varepsilon_1 < 0$ such that

$$\mathsf{P}^t\left(\chi_t^{-1}\Lambda_t < -1-\gamma\right) \le \exp\left(\frac{1}{4}\varepsilon\gamma\chi_t\right) \tag{2.3.44}$$

for all $\varepsilon \in (\varepsilon_1, 0)$ and all $t > t_1(\varepsilon,\gamma)$. Bounds (2.3.43) and (2.3.44) imply that condition $\Lambda 1$ holds. Thus implication (2.3.41) is proved. □

Conditions $\Lambda 2$ and $\Lambda 3$ are related to a theorem on large deviations for Λ_t as $t \to \infty$. To state it we introduce the following condition.

Λ^*. For any $\varepsilon \in (-\infty, \infty)$, the limit

$$\lim_{t\to\infty} \varphi_t^{-1} \ln H_t(\varepsilon) = \varkappa(\varepsilon) \tag{2.3.45}$$

exists where $\varphi_t \to \infty$ as $t \to \infty$ and $\varkappa(\varepsilon)$ is a proper convex function differentiable in the interval $(\varepsilon_-, \varepsilon_+)$ where

$$\varepsilon_- = \inf\{\varepsilon\colon \varkappa(\varepsilon) < \infty\}, \qquad \varepsilon_+ = \sup\{\varepsilon\colon \varkappa(\varepsilon) < \infty\}. \tag{2.3.46}$$

It is clear that $\varepsilon_- \le 0$ and $\varepsilon_+ \ge 1$. Let

$$\begin{aligned} \gamma_0 &= \varkappa'(0), \qquad \gamma_1 = \varkappa'(1), \\ \gamma_- &= \lim_{\varepsilon\downarrow\varepsilon_-} \varkappa'(\varepsilon), \qquad \gamma_+ = \lim_{\varepsilon\uparrow\varepsilon_+} \varkappa'(\varepsilon). \end{aligned} \tag{2.3.47}$$

Note that γ_0 is defined for $\varepsilon_- < 0$ only and γ_1 is defined for $\varepsilon_+ > 1$. If condition Λ^* holds with $\varepsilon_- < 0$, then it is clear that conditions $\Lambda 2$ and $\Lambda 3$ hold with $\chi_t = -\gamma_0\varphi_t$. It follows from the properties of the function $H_t(\varepsilon)$ that $\gamma_0 < 0$. In what follows, we will reveal a tighter relationship between conditions $\Lambda 2$, $\Lambda 3$ and condition Λ^* providing a theorem on large deviations for Λ_t.

The following result gives upper and lower bounds for $\beta(\delta_t^{+,\alpha})$ for all $\alpha \in (0,1)$ enabling us to obtain $\beta 1$ and $\beta 2$ if conditions $\alpha 1'$ and $\alpha 2'$ hold.

THEOREM 2.3.2. *For all $\alpha \in (0,1)$ and all $t \in \mathbf{R}_+$*

$$\beta(\delta_t^{+,\alpha}) \ge (1-\alpha)^{\varepsilon/(\varepsilon-1)}(H_t(1-\varepsilon))^{1/(1-\varepsilon)}, \qquad \varepsilon > 1, \tag{2.3.48}$$

$$\beta(\delta_t^{+,\alpha}) \le (1-\varepsilon)(\varepsilon/\alpha)^{\varepsilon/(1-\varepsilon)}(H_t(1-\varepsilon))^{1/(1-\varepsilon)}, \qquad 0 < \varepsilon < 1. \tag{2.3.49}$$

PROOF. If $H_t(1-\varepsilon) = \infty$ for $\varepsilon > 1$, then estimate (2.3.48) is trivial. Therefore we assume that $H_t(1-\varepsilon) < \infty$ for $\varepsilon > 1$. By the definition of a Bayes test, we have for any $c \in (0,\infty)$

$$\begin{aligned} c\alpha + \beta(\delta_t^{+,\alpha}) &\ge \inf\{c\alpha(\delta_t) + \beta(\delta_t)\colon \delta \in \Sigma^t\} = c\alpha\left(\delta_t^{c,1}\right) + \beta\left(\delta_t^{c,1}\right) \\ &= 1 - \mathsf{E}_Q^t\,(\tilde{\jmath}_t - c\jmath_t)^+ \end{aligned} \tag{2.3.50}$$

where $a^+ = a \vee 0$. By the fundamental Neyman–Pearson lemma and by the definition of a Bayes test, there exists a constant $c \in (0, \infty)$ such that inequality (2.3.50) becomes an equality. Therefore

$$\beta(\delta_t^{+,\alpha}) = 1 - \inf\left\{c\alpha + \mathsf{E}_Q^t(\tilde{\jmath}_t - c\jmath_t)^+; c \geq 0\right\}. \tag{2.3.51}$$

The condition $H_t(1-\varepsilon) < \infty$ for $\varepsilon > 1$ implies that $\mathbf{Q}(\tilde{\jmath}_t = 0) = 0$. Hence

$$\mathsf{E}_Q^t(\tilde{\jmath}_t - c\jmath_t)^+ = \widetilde{\mathsf{E}}^t(1 \vee cz_t) - c. \tag{2.3.52}$$

Since $1 \vee z \leq az^\varepsilon + 1$ for $\varepsilon > 1$, $z \geq 0$, and $a = \varepsilon^{-\varepsilon}(\varepsilon - 1)^{\varepsilon-1}$, relations (2.3.51) and (2.3.52) imply that

$$\beta(\delta_t^{+,\alpha}) \geq \sup\left\{(1-\alpha)c - ac^\varepsilon H_t(1-\varepsilon); c \geq 0\right\} \tag{2.3.53}$$

by (2.3.38) for $\varepsilon > 1$. The upper bound (2.3.53) is attained at

$$c = c^* = (1-\alpha)^{1/(\varepsilon-1)}(a\varepsilon H_t(1-\varepsilon))^{1/(1-\varepsilon)}. \tag{2.3.54}$$

Therefore (2.3.53) and (2.3.54) imply estimate (2.3.48).

Now we prove estimate (2.3.49). As in the proof of equality (2.3.52) we get

$$\mathsf{E}_Q^t(\tilde{\jmath}_t - c\jmath_t)^+ = 1 - \widetilde{\mathsf{E}}^t(1 \wedge c\tilde{z}_t). \tag{2.3.55}$$

Since

$$z \wedge 1 \leq z^\varepsilon, \qquad z \geq 0,\ 0 < \varepsilon < 1, \tag{2.3.56}$$

we obtain from (2.3.51), (2.3.55), and (2.3.38) for $0 < \varepsilon < 1$ that

$$\beta(\delta_t^{+,\alpha}) \leq \sup\left\{c^\varepsilon H_t(1-\varepsilon) - c\alpha; c \geq 0\right\}. \tag{2.3.57}$$

The upper bound in (2.3.57) is attained at

$$c = c^* = \alpha^{1/(\varepsilon-1)}(\varepsilon H_t(1-\varepsilon))^{1/(1-\varepsilon)}. \tag{2.3.58}$$

Therefore (2.3.57) and (2.3.58) imply estimate (2.3.49). □

COROLLARY 2.3.4. *For all $\varepsilon, \alpha \in (0,1)$ and $t \in \mathbf{R}_+$*

$$H_t(\varepsilon) \geq \beta^\varepsilon(\delta_t^{+,\alpha})\alpha^{1-\varepsilon}e^{h(\varepsilon)}$$

where $h(\varepsilon) = -\varepsilon\ln\varepsilon - (1-\varepsilon)\ln(1-\varepsilon)$ is the Shannon entropy of the distribution of the random variable taking two values with probabilities ε and $1-\varepsilon$.

REMARK 2.3.6. If $\overline{\alpha}_t < 1$, then $H_t(\varepsilon) = \infty$ for all $\varepsilon < 0$. Therefore, in the case of $\overline{\alpha}_t < 1$, estimate (2.3.48) acquires the trivial form $\beta(\delta_t^{+,\alpha}) \geq 0$ for all $\alpha \in (0,1)$. We also note that (1.1.50) implies $\beta(\delta_t^{+,\alpha}) = 0$ for $\alpha \in [\overline{\alpha}_t, 1]$. Estimate (2.3.49) is rather rough for $\alpha \in [\overline{\alpha}_t, 1]$, however $\overline{\alpha}_t \to 1$ as $t \to \infty$ by (2.3.5) and if condition $\Lambda 1$ holds. Moreover if condition $\Lambda 3$ holds, then Remark 2.3.5 yields that there exists $t_0 < \infty$ such that $\overline{\alpha}_t = 1$ for all $t > t_0$ (see Theorem 2.7.2 in [**37**] for improved estimates (2.3.48) and (2.3.49)).

THEOREM 2.3.3. *The following implications hold:*

$$\Lambda 2,\ \alpha 1' \Rightarrow d1 \Rightarrow \beta 1, \tag{2.3.59}$$

$$\Lambda 3,\ \alpha 2' \Rightarrow \beta 2 \Rightarrow d2. \tag{2.3.60}$$

PROOF. According to Theorem 2.3.1, it is sufficient to prove the first implications in (2.3.59) and (2.3.60). The first equality in (2.3.50) implies that for all $c > 0$

$$c\alpha\left(\delta_t^{c,1}\right) + \beta\left(\delta_t^{c,1}\right) \le c^{\varepsilon}\widetilde{\mathrm{E}}^t\widetilde{z}_t^{\varepsilon}, \qquad 0 < \varepsilon < 1,$$

in view of (2.3.55) and (2.3.56). Thus

$$e^{d_t}\alpha_t + \beta\left(\delta_t^{+,\alpha_t}\right) \le e^{\varepsilon d_t}H_t(1-\varepsilon)$$

for $0 < \varepsilon < 1$ by the first equality in (2.3.50) and by equality (2.3.38). Therefore we have for $0 < \varepsilon < 1$

$$d_t \le (1-\varepsilon)^{-1}\ln H_t(1-\varepsilon) - (1-\varepsilon)^{-1}\ln\alpha_t,$$

whence the implication $\Lambda 2$, $\alpha 1' \Rightarrow \beta 1$ follows.

The implication $\Lambda 3$, $\alpha 2' \Rightarrow \beta 2$ follows from estimate (2.3.48). □

REMARK 2.3.7. The implication $\Lambda 2$, $\alpha 1' \Rightarrow \beta 1$ can be obtained directly from estimate (2.3.49).

Theorem 2.3.3 implies the following well-known result of Krafft and Plachky (see [**37**]).

COROLLARY 2.3.5. *Let the assumptions of Corollary* 2.3.3 *hold. If conditions* $\alpha 1'$ *and* $\alpha 2'$ *are satisfied with* $\chi_n = n$, *then* (2.3.22) *holds.*

EXAMPLE 2.3.2. This is a continuation of Example 2.3.1. Relations (2.3.23)–(2.3.25) imply for all $t \in \mathbf{R}_+$ and $\varepsilon \in (-\infty, \infty)$ that

$$H_n(\varepsilon) = \exp\left\{-\frac{\varepsilon(1-\varepsilon)}{2}v_n^2\right\}, \tag{2.3.61}$$

whence

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}(\varepsilon\chi_n)^{-1}\ln H_n(\varepsilon) = -1$$

for $\chi_n = 2^{-1}v_n^2$, that is, conditions $\Lambda 2$ and $\Lambda 3$ hold. Theorem 2.3.3 implies (2.3.30) for the test δ_n^{+,α_n} of level α_n satisfying conditions $\alpha 1'$ and $\alpha 2'$.

The following result of independent interest holds under the assumptions of the latter example.

THEOREM 2.3.4. *If* $v_n \to \infty$ *as* $n \to \infty$, *then*

1) *if* $\alpha_n \to 0$ *as* $n \to \infty$, *then*

$$\alpha 1' \iff d1 \iff \beta 1 \iff z_{1-\alpha_n} = o(v_n); \tag{2.3.62}$$

2) *if* $\alpha_n \to 1$ *as* $n \to \infty$, *then*

$$\alpha 2' \iff d2 \iff \beta 2 \iff z_{1-\alpha_n} = o(v_n) \tag{2.3.63}$$

where z_p *is a p-quantile of the distribution* $\mathcal{N}(0,1)$ *and* $\chi_n = 2^{-1}v_n^2$.

PROOF. Let $\alpha_n \to 0$ as $n \to \infty$. Then it follows from (2.3.33) that

$$\ln \alpha_n = -2^{-1} z_{1-\alpha_n}^2 (1 + o(1)), \qquad n \to \infty. \tag{2.3.64}$$

Therefore

$$\alpha 1' \iff z_{1-\alpha_n} = o(v_n). \tag{2.3.65}$$

Since $v_n \to \infty$ as $n \to \infty$, conditions $\Lambda 1$, $\Lambda 2$, and $\Lambda 3$ hold for $\chi_n = 2^{-1} v_n^2$. Therefore, in view of Theorems 2.3.1 and 2.3.3, relation (2.3.62) follows from the chain of implications

$$\beta 1 \Rightarrow d1 \Rightarrow \alpha 1'. \tag{2.3.66}$$

Since $z_{1-\alpha_n} \to \infty$ as $n \to \infty$, it follows from (2.3.28) that $d1 \Rightarrow z_{1-a_n} = o(v_n)$, whence the implication $d1 \Rightarrow \alpha 1'$ follows in view of (2.3.65).

If $\beta 1$ holds, we have by (2.3.32) that $v_n - z_{1-\alpha_n} \to \infty$ as $n \to \infty$. Applying (2.3.33) we get

$$\lim_{n\to\infty} (v_n - z_{1-\alpha_n})^{-2} \ln \beta(\delta_n^{+,\alpha_n}) = -2^{-1}.$$

Thus $z_{1-a_n} = o(v_n)$ by $\beta 1$. Hence the implication $\beta 1 \Rightarrow d1$ follows from (2.3.65) and the equivalence $d1 \iff \alpha 1'$ proved above. Therefore implication (2.3.66) is proved.

The proof of (2.3.63) is similar to that of (2.3.62) and thus is omitted. □

REMARK 2.3.8. Since condition $\Lambda 2$ holds, we see from statement 1) in Theorem 2.3.4 that conditions $\alpha 1'$ and $d1$ cannot be weakened to prove the chain of implications $\Lambda 2$, $\alpha 1' \Rightarrow d1 \Rightarrow \beta 1$ in Theorem 2.3.3. Since condition $\Lambda 3$ also holds, we obtain from statement 2) in Theorem 2.3.4 that conditions $\alpha 2'$ and $\beta 2$ cannot be weakened in the chain of implications $\Lambda 3$, $\alpha 2' \Rightarrow \beta 2 \Rightarrow d2$ in Theorem 2.3.3. Moreover, it follows from Theorem 2.3.4 that

$$(\alpha 1', \alpha 2') \iff (d1, d2) \iff (\beta 1, \beta 2) \iff z_{1-\alpha_n} = o(v_n),$$

that is, relation (2.3.30) is equivalent to conditions $\alpha 1'$ and $\alpha 2'$.

Rates of decay of probabilities of error of the Neyman–Pearson, Bayes, and minimax tests under condition Λ^*. Throughout this section we assume that condition Λ^* holds. In this case, the Neyman–Pearson test $\delta_t^{+,\alpha}$ can be used even if the level α_t tends to zero as $t \to \infty$ faster than is allowed by condition $\alpha 1'$. In particular, if $\varphi_t^{-1} \ln \alpha_t \to -a$ for some positive number a, then one can prove that $\varphi_t^{-1} \ln \beta(\delta_t^{+,\alpha}) \to -b(a)$ where $b(a)$ is a positive function of a. In order to provide an exact statement, we need the following result on large deviations of Λ_t. Below we use the following notation (see also (2.3.45)–(2.3.47)):

$$\Gamma_0 = \gamma_0 I(\varepsilon_- < 0) + \gamma_- I(\varepsilon_- = 0), \tag{2.3.67}$$

$$\Gamma_1 = \gamma_1 I(\varepsilon_+ > 1) + \gamma_+ I(\varepsilon_+ = 1). \tag{2.3.68}$$

THEOREM 2.3.5. *Let condition Λ^* be satisfied. If*

$$\Gamma_0 < \gamma_+,$$

then for all $\gamma \in (\Gamma_0, \gamma_+)$

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \mathsf{P}^t\left(\varphi_t^{-1} \Lambda_t > \gamma\right) = \lim_{t\to\infty} \varphi_t^{-1} \ln \mathsf{P}^t\left(\varphi_t^{-1} \Lambda_t \geq \gamma\right) = -I(\gamma). \tag{2.3.69}$$

Further if $\gamma_- < \Gamma_1$, *then for all* $\gamma \in (\gamma_-, \Gamma_1)$

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \widetilde{\mathsf{P}}^t \left(\varphi_t^{-1}\Lambda_t < \gamma\right) = \lim_{t\to\infty} \varphi_t^{-1} \ln \widetilde{\mathsf{P}}^t \left(\varphi_t^{-1}\Lambda_t \le \gamma\right) = -I(\gamma) + \gamma \tag{2.3.70}$$

where $I(\gamma) = \gamma\varepsilon(\gamma) - \varkappa(\varepsilon(\gamma))$ *and* $\varepsilon(\gamma)$ *is an arbitrary solution of the equation* $\varkappa'(\varepsilon) = \gamma$.

Theorem 2.3.5 can be deduced from Theorem 2.6.3 of [**17**] in view of equality (2.3.39). This proof can be found in [**39**].

REMARK 2.3.9. Applying the methods of convex analysis [**44**], we can readily obtain that the function $I(\gamma)$ is strictly convex in the interval (γ_-, γ_+) and has the unique minimum at $\gamma = \gamma_0$ if $\varepsilon_- < 0$. Moreover $I(\gamma_0) = 0$ in this case, while the minimum is attained at $\gamma = \gamma_-$ if $\varepsilon_- = 0$.

First we consider the Bayes test δ_t^π with respect to the a priori distribution $(\pi, \widetilde{\pi})$, $\pi + \widetilde{\pi} = 1$, and the loss $A_{ij} = 1 - \delta_{ij}$ (see Section 2.1). It follows from Section 1.2 that we can put $\delta_t^\pi = \delta_t^{c,\varepsilon}$ where $c = \pi/\widetilde{\pi}$ and $\varepsilon \in [0, 1]$ is an arbitrary number.

THEOREM 2.3.6. *Let condition* Λ^* *hold with* $\Gamma_0 < 0 < \Gamma_1$. *Then*

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \alpha(\delta_t^\pi) = \lim_{t\to\infty} \varphi_t^{-1} \ln \beta(\delta_t^\pi) = -I(0), \tag{2.3.71}$$

$$\lim_{t\to\infty} \varphi_t^{-1} \ln e_\pi(\delta_t^\pi) = -I(0) \tag{2.3.72}$$

where e_π *is the probability of error of the test* δ_t^π (*see* (1.1.32)).

PROOF. It is clear that for any Bayes test δ_t^π

$$\mathsf{P}^t(\Lambda_t > \ln(\pi/\widetilde{\pi})) \le \alpha(\delta_t^\pi) \le \mathsf{P}^t(\Lambda_t \ge \ln(\pi/\widetilde{\pi})), \tag{2.3.73}$$

$$\widetilde{\mathsf{P}}^t(\Lambda_t < \ln(\pi/\widetilde{\pi})) \le \beta(\delta_t^\pi) \le \widetilde{\mathsf{P}}^t(\Lambda_t \le \ln(\pi/\widetilde{\pi})). \tag{2.3.74}$$

Then (2.3.69), (2.3.70), (2.3.73), and (2.3.74) imply (2.3.71). Now the equality

$$e_\pi(\delta_t^\pi) = \pi\alpha(\delta_t^\pi) + \widetilde{\pi}\beta(\delta_t^\pi)$$

and (2.3.71) imply (2.3.72). □

REMARK 2.3.10. By (2.3.67) and (2.3.68), the condition $\Gamma_0 < 0 < \Gamma_1$ implies that

$$\Gamma_0 < \gamma_+ \quad \text{and} \quad \gamma_- < \Gamma_1,$$

which enables us to apply relation (2.3.69) for $\gamma \in (\Gamma_0, \gamma_+)$ and relation (2.3.70) for $\gamma \in (\gamma_-, \Gamma_1)$. On the other hand,

$$\varphi_t^{-1} \ln(\pi/\widetilde{\pi}) \to \gamma = 0 \in (\Gamma_0, \Gamma_1) \quad \text{as } t \to \infty.$$

Now let δ_t^* be the minimax test for distinguishing the hypotheses H^t and $\widetilde{H}^t$ (see Section 2.1). According to the results of Section 1.2, $e(\delta_t^*)$ is the probability of error for the test δ_t^* (see (1.2.19)).

THEOREM 2.3.7. *Let condition Λ^* hold with $\Gamma_0 < 0 < \Gamma_1$. Then*

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \alpha(\delta_t^*) = \lim_{t\to\infty} \varphi_t^{-1} \ln \beta(\delta_t^*) = -I(0), \tag{2.3.75}$$

$$\lim_{t\to\infty} \varphi_t^{-1} \ln e_\pi(\delta_t^*) = -I(0). \tag{2.3.76}$$

PROOF. Let $(\pi, \widetilde{\pi})$ be the a priori distribution of the hypotheses, $\pi + \widetilde{\pi} = 1$. Then the definitions of Bayes tests and minimax tests imply that

$$e_\pi(\delta_t^\pi) = \min_{\delta_t} e_\pi(\delta_t) \le \min_{\delta_t}(\alpha(\delta_t) \vee \beta(\delta_t)) = e(\delta_t^*),$$
$$e_\pi(\delta_t^\pi) \ge (\pi \wedge \widetilde{\pi})(\alpha(\delta_t^\pi) \vee \beta(\delta_t^\pi)) \ge (\pi \wedge \widetilde{\pi}) e(\delta_t^*),$$

that is,

$$(\pi \wedge \widetilde{\pi}) e(\delta_t^*) \le e_\pi(\delta_t^*) \le e(\delta_t^*). \tag{2.3.77}$$

Combining (2.3.72) and (2.3.77), we obtain (2.3.76). Relation (2.3.75) follows from (2.3.76), since $\alpha(\delta_t^*) = \beta(\delta_t^*) = e(\delta_t^*)$ (see Theorem 1.2.4). □

The following result describes a relationship between the rates of decay of the level α_t and the type II error probability $\beta(\delta_t^{+,\alpha_t})$ for the Neyman–Pearson test under condition Λ^*.

THEOREM 2.3.8. *Let condition Λ^* hold with $\Gamma_0 < \Gamma_1$. Then*

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \alpha_t = -a \iff \lim_{t\to\infty} \varphi_t^{-1} \ln \beta_t(\delta_t^{+,\alpha_t}) = -b(a) \tag{2.3.78}$$

for any $a \in (I(\Gamma_0), I(\Gamma_1))$ where

$$b(a) = a - \gamma(a) \in \big(I(\Gamma_1) - \Gamma_1, I(\Gamma_0) - \Gamma_0\big)$$

and $\gamma(a)$ is a unique solution of the equation $I(\gamma) = a$.

PROOF. We prove the implication $\Rightarrow$ in (2.3.78). Assume that $\varphi_t^{-1} \ln \alpha_t \to -a$ as $t \to \infty$ for $a \in (I(\Gamma_0), I(\Gamma_1))$. We have

$$\alpha_t = \mathsf{P}^t(Y_t > y_t) + \varepsilon \mathsf{P}^t(Y_t = y_t)$$

where $Y_t = \varphi_t^{-1} \Lambda_t$ and $y_t = \varphi_t^{-1} d_t$. Our current goal is to show that $y_t \to \gamma(a)$ as $t \to \infty$. Put

$$\underline{y} = \liminf_{t\to\infty} y_t, \qquad \overline{y} = \limsup_{t\to\infty} y_t.$$

Then it is sufficient to prove that $\underline{y} = \gamma(a)$ and $\overline{y} = \gamma(a)$.

First we show that $\underline{y} = \gamma(a)$. Assume for contradiction that $\underline{y} \ne \gamma(a)$. Below we use the obvious estimates

$$\mathsf{P}^t(Y_t > y_t) \le \alpha_t \le \mathsf{P}^t(Y_t \ge y_t). \tag{2.3.79}$$

By the definition of $\underline{y}$, there exists a sequence (t_n) such that $t_n \to \infty$ and $y_{t_n} \to \underline{y}$ as $n \to \infty$. Then, if $\underline{y}$ is finite, for any $y' < \underline{y}$ and $y'' > \underline{y}$ there exists $n_0 = n_0(y', y'')$ such that $y' < y_{t_n} < y''$ for all $n > n_0$. If $\underline{y} = +\infty$, then for any $y' < \underline{y} = \infty$ there exists $n_0 = n_0(y')$ such that $y_{t_n} > y'$ for all $n > n_0$, while if $\underline{y} = -\infty$, then for any $y'' > \underline{y} = -\infty$ there exists $n_0 = n_0(y'')$ such that $y_{t_n} < y''$ for all $n > n_0$.

First let $\underline{y} \le \Gamma_0$ and $y'' \in (\Gamma_0, \gamma(a))$. Since $y_{t_n} < y''$ for all $n > n_0$, we have by Theorem 2.3.5

$$\liminf_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} > y_{t_n}) \ge \lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} > y'') = -I(y'').$$

Therefore (2.3.79) implies

$$\lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \alpha^{t_n} \ge -I(y'') > -I(\gamma(a)) = -a,$$

since $\Gamma_0 < y'' < \gamma(a)$. By Remark 2.3.9, the function $I(\gamma)$ is strictly increasing on the interval (Γ_0, γ_+), giving a contradiction.

Now assume that $\underline{y} \in (\Gamma_0, \gamma(a)) \cup (\gamma(a), \gamma_+)$ and that y' and y'' are such that $(y', y'') \subset (\Gamma_0, \gamma(a)) \cup (\gamma(a), \gamma_+)$. Then by Theorem 2.3.5 we have

$$\liminf_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} > y_{t_n}) \ge \lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} > y'') = -I(y''),$$
$$\limsup_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} \ge y_{t_n}) \le \lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} \ge y') = -I(y').$$

Since y' and y'' are arbitrary and since the function $I(\gamma)$ is continuous on the interval (γ_-, γ_+), we obtain in view of inequalities (2.3.79) that

$$\lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \alpha_{t_n} = -I(\underline{y}).$$

Since $\underline{y} \ne \gamma(a)$ by assumption and since $I(\gamma)$ is strictly increasing on (Γ_0, γ_+), this again gives a contradiction.

Finally, let $\underline{y} > \gamma_+$ and $y' \in (\gamma(a), \gamma_+)$. Since $y_{t_n} > y'$ for all $n > n_0$, we apply Theorem 2.3.5 once more to obtain

$$\limsup_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} \ge y_{t_n}) \le \lim_{n\to\infty} \varphi_{t_n}^{-1} \ln \mathsf{P}^{t_n}(Y_{t_n} \ge y') = -I(y').$$

Since $\gamma(a) < y' < \gamma_+$, we have $I(y') > I(\gamma(a)) = a$, giving a contradiction by (2.3.79).

The above contradictions show that $\underline{y} = \gamma(a)$. By a similar argument, we can prove that $\overline{y} = \gamma(a)$. Therefore $y_t \to \gamma(a)$ as $t \to \infty$.

Let $\varepsilon > 0$ be arbitrary and let $t = t_0(\varepsilon)$ be such that $|y_t - \gamma(a)| < \varepsilon$ for all $t > t_0$. Assume that $\varepsilon > 0$ is small enough in order that $(\gamma(a) - \varepsilon, \gamma(a) + \varepsilon) \subset (\Gamma_0, \Gamma_1)$. It is clear that the following inequalities hold for all $t > t_0$:

$$\widetilde{\mathsf{P}}^t\left(\varphi_t^{-1}\Lambda_t < \gamma(a) - \varepsilon\right) \le \beta(\delta_t^{+,\alpha_t}) \le \widetilde{\mathsf{P}}^t\left(\varphi_t^{-1}\Lambda_t \le \gamma(a) + \varepsilon\right),$$

whence

$$\liminf_{t\to\infty} \varphi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \ge -I(\gamma(a) - \varepsilon) + \gamma(a) - \varepsilon,$$
$$\limsup_{t\to\infty} \varphi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \le -I(\gamma(a) + \varepsilon) + \gamma(a) + \varepsilon$$

by Theorem 2.3.5. Approaching the limit in these inequalities as $\varepsilon \to 0$ we obtain that the upper and lower bounds in these inequalities coincide and

$$\lim_{t\to\infty} \varphi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) = -a + \gamma(a) = -b(a),$$

since the function $I(\gamma)$ is continuous in the interval (γ_-, γ_+) and $I(\gamma(a)) = a$.

Therefore the implication $\Rightarrow$ in (2.3.78) is proved. The inverse implication $\Leftarrow$ in (2.3.78) can be proved along similar lines. $\square$

REMARK 2.3.11. The behavior of $\beta(\delta_t^{+,\alpha_t})$ in the case where $\varphi_t^{-1}\ln\alpha_t \to -a$ for $a \notin (I(\Gamma_0), I(\Gamma_1))$ is studied in [**39**]. In particular, the results in [**39**] show that for any $a \in [0, I(\Gamma_0)]$ we have

$$\lim_{t\to\infty} \varphi_t^{-1}\ln = a_t = -a \Rightarrow \limsup_{t\to\infty} \varphi_t^{-1}\ln\beta\left(\delta_t^{+,\alpha_t}\right) \le \Gamma_0 - I(\Gamma_0). \tag{2.3.80}$$

Let $\varepsilon_- < 0$. Then $\Gamma_0 = \gamma_0$ and $I(\Gamma_0) = 0$. Therefore (2.3.80) becomes of the form

$$\lim_{t\to\infty} \varphi_t^{-1}\ln = a = 0 \Rightarrow \limsup_{t\to\infty} \varphi_t^{-1}\ln\beta\left(\delta_t^{+,\alpha_t}\right) \le \gamma_0. \tag{2.3.81}$$

As we have already observed, in this case conditions $\Lambda 2$ and $\Lambda 3$ hold with

$$\chi_t = -\gamma_0\varphi_t.$$

Therefore implication (2.3.81) is equivalent to the implication $\alpha 1' \Rightarrow \beta 1$. If condition $\alpha 2'$ holds, then we obtain from the implication $\alpha 2' \Rightarrow \beta 2$ that

$$\alpha 1', \alpha 2' \Rightarrow \lim_{t\to\infty} \varphi_t^{-1}\ln\beta\left(\delta_t^{+,\alpha_t}\right) = -\gamma_0$$

if condition Λ^* holds with $\varepsilon_- < 0$ and $\Gamma_0 < \Gamma_1$.

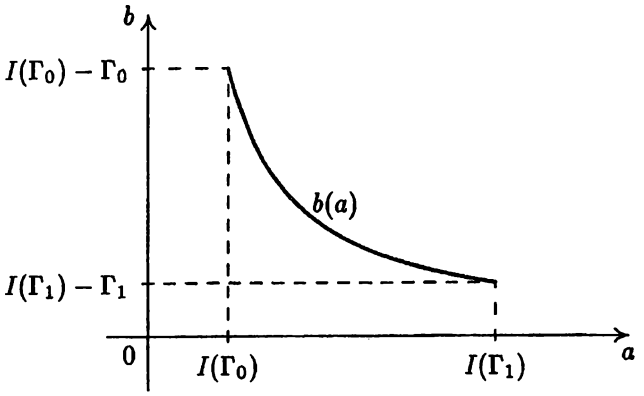

FIGURE 2.3.1

REMARK 2.3.12. The function $b(a)$ is shown in Figure 2.3.1. According to Remark 2.3.11 we have $I(\Gamma_0) = 0$ and $I(\Gamma_0) - \Gamma_0 = -\gamma_0$ if $\varepsilon_- < 0$. If $\varepsilon_- = 0$, then $\Gamma_0 = \gamma_- = \varkappa'(0+)$ and $I(\Gamma_0) - \Gamma_0 = -\varkappa(0+) - \varkappa'(0+)$. If $\varepsilon_+ > 1$, then $\Gamma_1 = \gamma_1$, $I(\gamma_1) = \gamma_1$, and $I(\Gamma_1) - \Gamma_1 = 0$. If $\varepsilon_+ = 1$, then $\Gamma_1 = \gamma_+ = \varkappa'(1-)$ and $I(\Gamma_1) - \Gamma_1 = -\varkappa(1-)$.

EXAMPLE 2.3.3. This is a continuation of Example 2.3.2. Relation (2.3.61) implies condition Λ^* with $\varphi_n = v_n^2$, $\varkappa(\varepsilon) = -2^{-1}\varepsilon(1-\varepsilon)$, $\varepsilon_- = -\infty$, and $\varepsilon_+ = \infty$. Hence $\gamma_- = -\infty$, $\gamma_+ = \infty$, $\gamma_0 = -2^{-1}$, and $\gamma_1 = 2^{-1}$. It is easy to show that $I(\gamma) = 2^{-1}(\gamma + 2^{-1})^2$. It is clear that $\Gamma_0 < 0 < \Gamma_1$. Therefore Theorems 2.3.6–2.3.8 hold. In particular, Theorems 2.3.6 and 2.3.7 imply that

$$\varepsilon_\pi(\delta_n^\pi) = \exp\left\{-\frac{1}{8}v_n^2(1+o(1))\right\}, \qquad \varepsilon_\pi(\delta_n^*) = \exp\left\{-\frac{1}{8}v_n^2(1+o(1))\right\}.$$

Since $\gamma(a) = \sqrt{2a} - 2^{-1}$ is a solution of the equation $I(\gamma) = a$, we obtain

$$b(a) = 2^{-1}\left(\sqrt{2a} - 1\right)^2$$

for all

$$a \in \left(0, 2^{-1}\right) = \left(I(\Gamma_0), I(\Gamma_1)\right).$$

Therefore, in view of Theorem 2.3.8 we have that

$$\lim_{n\to\infty} v_n^{-2} \ln \alpha_n = -a \Longleftrightarrow \lim_{n\to\infty} v_n^{-2} \ln \beta\left(\delta_n^{+,\alpha_n}\right) = -2^{-1}\left(\sqrt{2a} - 1\right)^2$$

for all $a \in (0, 2^{-1})$

EXAMPLE 2.3.4. Let an observation be

$$\xi^{(n)} = (\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn})$$

where $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$ are independent both under H^n and $\widetilde{H}^n$. Assume that the random variable ξ_{ni} has the exponential distribution with the density

$$\lambda_{ni} \exp(-\lambda_{ni}(x - b_{ni}))I(x \geq b_{ni})$$

under the hypothesis H^n, and the exponential distribution with the density

$$\widetilde{\lambda}_{ni} \exp\left(-\widetilde{\lambda}_{ni}(x - \widetilde{b}_{ni})\right)I(x \geq b_{ni})$$

under the hypothesis $\widetilde{H}^n$. Here $\lambda_{ni} \in (0, \infty)$, $\widetilde{\lambda}_{ni} \in (0, \infty)$, $b_{ni} \in (-\infty, \infty)$, and $\widetilde{b}_{ni} \in (-\infty, \infty)$. It is easy to show that

$$\varepsilon_-^n = -\min_{1\leq i\leq n}\left[\left(\left(\frac{\widetilde{\lambda}_{ni}}{\lambda_{ni}} \vee 1\right) - 1\right)^{-1} I(\widetilde{b}_{ni} \leq b_{ni})\right], \tag{2.3.82}$$

$$\varepsilon_+^n = \min_{1\leq i\leq n}\left[1 - \left(\frac{\widetilde{\lambda}_{ni}}{\lambda_{ni}} \wedge 1\right) I(\widetilde{b}_{ni} \geq b_{ni})\right]^{-1}, \tag{2.3.83}$$

and that for all $\varepsilon \in (\varepsilon_-^n, \varepsilon_+^n)$

$$\begin{aligned} \ln H_n(\varepsilon) = \sum_{i=1}^n \left\{\varepsilon \ln \frac{\widetilde{\lambda}_{ni}}{\lambda_{ni}} - \ln\left[\varepsilon\left(\frac{\widetilde{\lambda}_{ni}}{\lambda_{ni}} - 1\right) + 1\right]\right\} \\ + \varepsilon \ln H_n(1) + (1 - \varepsilon) \ln H_n(0) \end{aligned} \tag{2.3.84}$$

where

$$\ln H_n(1) = -\sum_{i=1}^n \widetilde{\lambda}_{ni}(b_{ni} - \widetilde{b}_{ni})I(b_{ni} > \widetilde{b}_{ni}), \tag{2.3.85}$$

$$\ln H_n(0) = -\sum_{i=1}^n \lambda_{ni}(\widetilde{b}_{ni} - b_{ni})I(\widetilde{b}_{ni} > b_{ni}). \tag{2.3.86}$$

Relations (2.3.82)–(2.3.86) allow us to completely study the problem of the asymptotic distinguishability of the hypotheses H^n and $\widetilde{H}^n$, although we consider some particular cases only.

EXAMPLE 2.3.5. This is a continuation of Example 2.3.4. Let $\widetilde{b}_{ni} = b_{ni}$ for all $i = 1, 2, \dots, n$ and all $n = 1, 2, \dots$. Then it is clear that the measures P^n and $\widetilde{\mathsf{P}}^n$ are equivalent and therefore $\overline{\alpha}_n = \overline{\beta}_n = 1$ for all n. Assume that $\widetilde{\lambda}_{ni} = \lambda_{ni} = \rho_n$ for all $i = 1, 2, \dots, n$ and all $n = 1, 2, \dots$, and let $\rho_n \to \rho$ as $n \to \infty$. Relations (2.3.82)–(2.3.86) imply that

$$\ln H_n(\varepsilon) = n\big(\varepsilon \ln \rho_n - \ln[\varepsilon(\rho_n - 1) + 1]\big), \qquad \varepsilon \in \big(\varepsilon_-^n, \varepsilon_+^n\big),$$
$$\varepsilon_-^n = -((\rho_n \vee 1) - 1)^{-1}, \qquad \varepsilon_+^n = (1 - (\rho_n \wedge 1))^{-1}.$$

Now we find ρ for which condition Λ^* holds and evaluate the function $b(a)$ defined in Theorem 2.3.8.

If $\rho \in (0, \infty) \setminus \{1\}$, then condition Λ^* holds and

$$\varphi_n = n, \qquad \varepsilon_- = -((\rho \vee 1) - 1)^{-1}, \qquad \varepsilon_+ = (1 - (\rho \wedge 1))^{-1},$$
$$\varkappa(\varepsilon) = \varepsilon \ln \rho - \ln(\varepsilon(\rho - 1) + 1), \qquad \varepsilon \in (\varepsilon_-, \varepsilon_+).$$

In this case, $\gamma_- = \ln \rho$ and $\gamma_+ = \infty$ if $\rho \in (0, 1)$, while $\gamma_- = -\infty$ and $\gamma_+ = \ln \rho$ if $\rho \in (1, \infty)$. Moreover, we have for any ρ

$$\gamma_0 = -(\rho - 1 - \ln \rho), \qquad \gamma_1 = \rho^{-1} - 1 - \ln \rho^{-1}.$$

It is easy to show that for all $\gamma \in (\gamma_-, \gamma_+)$ we have

$$I(\gamma) = z(\gamma) - 1 - \ln z(\gamma)$$

where

$$z(\gamma) = (\ln \rho - \gamma)/(\rho - 1).$$

Observe that the condition $\Gamma_0 < 0 < \Gamma_1$ (and therefore the condition $\Gamma_0 < \Gamma_1$) holds, since $\Gamma_0 = \gamma_0 < 0$ and $\Gamma_1 = \gamma_1 > 0$. Therefore Theorems 2.3.6–2.3.8 hold and in Theorems 2.3.6 and 2.3.7 we have that

$$I(0) = \frac{\ln \rho}{\rho - 1} - 1 - \ln \frac{\ln \rho}{\rho - 1}.$$

If $\rho \in (0, 1)$, then $\gamma > \gamma_0$ for $z(\gamma) > 1$. Let z_a be a solution of the equation $z - 1 - \ln z = a$ for $z > 1$ and $a \in (0, \gamma_1)$. Then $\gamma(a) = \ln \rho + (1 - \rho) z_a$ and therefore

$$b(a) = a - \ln \rho - (1 - \rho) z_a.$$

If $\rho \in (1, \infty)$, then $\gamma > \gamma_0$ for $z(\gamma) \in (0, 1)$. Let $\widetilde{z}_a$ be a solution of the equation $z - 1 - \ln z = a$ for $z \in (0, 1)$ and $a \in (0, \gamma_1)$. Then $\gamma(a) = \ln \rho - (\rho - 1)\widetilde{z}_a$ and therefore

$$b(a) = a - \ln \rho + (\rho - 1)\widetilde{z}_a.$$

Now let $\rho = 1$. Assume that $n(\rho_n - 1)^2 \to \infty$ as $n \to \infty$. Then condition Λ^* holds with

$$\varphi_n = n(\rho_n - 1)^2, \qquad \varkappa(\varepsilon) = -\varepsilon(1 - \varepsilon)/2, \qquad \varepsilon_- = -\infty, \qquad \varepsilon_+ = \infty.$$

This implies that

$$\gamma_- = -\infty, \qquad \gamma_0 = -\frac{1}{2}, \qquad \gamma_1 = \frac{1}{2}, \qquad \gamma_+ = \infty, \qquad I(\gamma) = \frac{1}{2}\left(\gamma + \frac{1}{2}\right)^2.$$

Therefore we have $I(0) = 1/8$ in Theorems 2.3.6 and 2.3.7 and $b(a) = (1 - \sqrt{2a})^2/2$ in Theorem 2.3.8 (cf. Example 2.3.3).

If $\rho = 0$, then condition Λ^* holds with $\varphi_n = n \ln \rho_n^{-1}$, $\varepsilon_- = -\infty$, $\varepsilon_+ = 1$, and $\varkappa(\varepsilon) = -\varepsilon$. Since $\varkappa'(\varepsilon) = -1$ for all ε, we have $\gamma_- = \gamma_+ = -1$. Therefore the conditions $\Gamma_0 < \Gamma_1$ and $\Gamma_0 < 0 < \Gamma_1$ fail to hold and Theorems 2.3.6–2.3.8 do not apply. Since $\varepsilon_- < 0$, conditions $\Lambda 1$, $\Lambda 2$, and $\Lambda 3$ hold and therefore Theorems 2.3.1 and 2.3.3 apply.

If $\rho = \infty$, then condition Λ^* holds with $\varphi_n = n \ln \rho_n^{-1}$, $\varepsilon_- = 0$, $\varepsilon_+ = \infty$, and $\varkappa(\varepsilon) = \varepsilon - 1$. In this case, $\varkappa'(\varepsilon) = 1$ for all ε. Therefore $\gamma_- = \gamma_+ = 1$ and Theorems 2.3.6–2.3.8 do not apply. Observe that conditions $\Lambda 1$, $\Lambda 2$, and $\Lambda 3$ fail to hold in this case.

EXAMPLE 2.3.6. This is a continuation of Example 2.3.4. Let $\lambda_{ni} = \lambda$, $\widetilde{\lambda}_{ni} = \widetilde{\lambda}$, $b_{ni} = b$, and $\widetilde{b}_{ni} = \widetilde{b}$ for all $i = 1, 2, \dots, n$ and all $n = 1, 2, \dots$. Then condition Λ^* holds with $\varphi_n = n$,

$$\varepsilon_- = -((\rho \vee 1) - 1)^{-1} I(\widetilde{b} \le b), \qquad \varepsilon_+ = -\big[1 - (\rho \wedge 1) I(\widetilde{b} \ge b)\big]^{-1},$$

and

$$\varkappa(\varepsilon) = \varepsilon \ln \rho - \ln[\varepsilon(\rho - 1) + 1] + \varepsilon \varkappa(1) + (1 - \varepsilon)\varkappa(0)$$

for all $\varepsilon \in (\varepsilon_-, \varepsilon_+)$ where

$$\varkappa(1) = -\widetilde{\lambda}(b - \widetilde{b}) I(b > \widetilde{b}), \qquad \varkappa(0) = -\lambda(\widetilde{b} - b) I(\widetilde{b} > b), \qquad \rho = \frac{\widetilde{\lambda}}{\lambda}.$$

If $\rho = 1$ and $\widetilde{b} \ne b$, then $\gamma_- = \gamma_+$ and therefore Theorems 2.3.6–2.3.8 do not apply. In this case conditions $\Lambda 1$, $\Lambda 2$, and $\Lambda 3$ hold and therefore Theorems 2.3.1 and 2.3.3 apply if $\widetilde{b} < b$.

Let $\rho \ne 1$ and $\widetilde{b} > b$. Then

$$\varepsilon_- = 0, \qquad \varepsilon_+ = (1 - (\rho \wedge 1))^{-1}, \qquad \varkappa(\varepsilon) = \varepsilon \ln \rho - \ln[\varepsilon(\rho - 1) + 1] - (1 - \varepsilon)(\widetilde{b} - b)\lambda$$

in condition Λ^*. In this case, $\gamma_- = -(\rho - 1 - \ln \rho) + (\widetilde{b} - b)\lambda$ for all $\rho \in (0, \infty) \setminus \{1\}$, while $\gamma_+ = \infty$ if $\rho \in (0, 1)$, and $\gamma_+ = \ln \rho + (\widetilde{b} - b)\lambda$ if $\rho \in (1, \infty)$. Observe that $\Gamma_0 = \gamma_-$ and $\Gamma_1 = \gamma_1 = (\rho^{-1} - 1 - \ln \rho^{-1}) + (\widetilde{b} - b)\lambda$. The condition $\Gamma_0 < \Gamma_1$ clearly holds and therefore Theorem 2.3.8 applies. The condition $\Gamma_0 < 0 < \Gamma_1$ holds if $(\widetilde{b} - b)\lambda < \rho - 1 - \ln \rho$ and therefore Theorems 2.3.6 and 2.3.7 apply only in this case. It is easy to show that for all $\gamma \in (\gamma_-, \gamma_+)$

$$I(\gamma) = z(\gamma) - 1 - \ln z(\gamma) + (\widetilde{b} - b)\lambda$$

where $z(\gamma) = (\rho - 1)^{-1}(\ln \rho + (\widetilde{b} - b)\lambda - \gamma)$.

If $\rho \in (0, 1)$, then $\gamma > \gamma_-$ for $z(\gamma) > 1$. Let z_a be a solution of the equation $z - 1 - \ln z + (\widetilde{b} - b)\lambda = a$ with respect to $z \in (1, \infty)$ where

$$a \in (I(\Gamma_0), I(\Gamma_1)) = \big((\widetilde{b} - b)\lambda, \gamma_1\big).$$

Then $\gamma(a) = \ln \rho + (1 - \rho) z_a + (\widetilde{b} - b)\lambda$ and therefore

$$b(a) = a - \ln \rho - (1 - \rho) z_a - (\widetilde{b} - b)\lambda.$$

If $\rho \in (1, \infty)$, then $\gamma > \gamma_-$ for $z(\gamma) \in (0, 1)$. Let $\widetilde{z}_a$ be a solution of the equation $z - 1 - \ln z + (\widetilde{b} - b)\lambda = a$ with respect to $z \in (0, 1)$ where $a \in ((\widetilde{b} - b)\lambda, \gamma_1)$. Then

$\gamma(a) = \ln\rho - (\rho-1)\widetilde{z}_a + (\widetilde{b}-b)\lambda$ and therefore

$$b(a) = a - \ln\rho + (\rho-1)\widetilde{z}_a - (\widetilde{b}-b)\lambda.$$

If $\rho \neq 1$ and $\widetilde{b} < b$, then $\varepsilon_- = -((\rho\wedge 1)-1)^{-1}$, $\varepsilon_+ = 1$, and

$$\varkappa(\varepsilon) = \varepsilon\ln\rho - \ln[\varepsilon(\rho-1)+1] - \varepsilon(b-\widetilde{b})\lambda.$$

This case can be considered similarly to the case $\widetilde{b} > b$.

2.4. Complete asymptotic distinguishability under the weak convergence

Consider completely asymptotically distinguishable families of hypotheses (H^t) and $(\widetilde{H}^t)$ in two cases:

1) the law of large numbers (LLN) for Λ_t does not hold;
2) the law of large numbers for Λ_t holds and, moreover, $(\Lambda_t - \varphi_t)/\psi_t$ converges weakly with some φ_t and ψ_t.

The case where the LLN does not hold. Let $\mathcal{L}\left(\eta_t|\mathsf{P}^t\right)$ be the distribution of η_t with respect to the measure P^t and let the symbol $\xrightarrow{w}$ stand for the weak convergence of probability distributions. Introduce the following condition:

$\Lambda 4$. $\mathcal{L}\left(\psi_t^{-1}\Lambda_t|\mathsf{P}^t\right) \xrightarrow{w} L$ as $t\to\infty$ where ψ_t is a positive function such that $\psi_t \to \infty$ as $t\to\infty$ and L is a probability distribution on $\mathbf{R}$ whose distribution function is $L(x)$.

LEMMA 2.4.1. *If condition $\Lambda 4$ holds, then $\overline{\alpha}_t \to 1$ as $t\to\infty$ and $L(x) = 1$ for any $x > 0$. Moreover,*

$$(H^t) \triangle (\widetilde{H}^t) \iff L(0) = 1. \tag{2.4.1}$$

PROOF. For any $a \in (-\infty, 0)$, we have

$$\mathsf{P}^t\left(\psi_t^{-1}\Lambda_t < a\right) \geq \mathsf{P}^t(\Lambda_t = -\infty) = 1 - \overline{\alpha}_t.$$

If a is a point of continuity of the function $L(x)$, then $L(a) \geq \limsup_{t\to\infty}(1-\overline{\alpha}_t)$ by condition $\Lambda 4$. Passing to the limit over the points of continuity $a \to -\infty$, we obtain $\overline{\alpha}_t \to 1$ as $t\to\infty$.

Choose $a > 0$ and apply the Chebyshev inequality to obtain

$$\mathsf{P}^t\left(\psi_t^{-1}\Lambda_t > a\right) = \mathsf{P}^t\left(z_t > e^{a\psi_t}\right) \leq e^{-a\psi_t}\mathsf{E}^t z_t \leq e^{-a\psi_t}.$$

Passing to the limit as $t\to\infty$ in this inequality, we have $1 - L(a) = 0$ and therefore $L(a) = 1$ for any $a > 0$.

Further, in view of condition $\Lambda 4$, we have as $t\to\infty$

$$\mathsf{P}^t(\Lambda_t > N) = \mathsf{P}^t\left(\frac{\Lambda_t}{\psi_t} > \frac{N}{\psi_t}\right) \to \begin{cases} 1 - L(0), & N < 0, \\ 1 - L(0+), & N > 0. \end{cases}$$

Since $L(0+) = 1$, Theorem 2.2.1 gives the required equivalence (2.4.1). □

REMARK 2.4.1. Equivalence (2.4.1) means the following: If condition $\Lambda 4$ holds, then the complete asymptotic distinguishability $(H^t) \triangle (\widetilde{H}^t)$ holds if and only if the distribution function $L(x)$ is continuous at the point $x = 0$.

The next result, which is an analog of Theorem 2.3.1 (see also Corollary 2.3.1), describes the behavior of the Neyman–Pearson test δ_t^{+,α_t} as $t \to \infty$ under condition $\Lambda 4$.

THEOREM 2.4.1. *If condition $\Lambda 4$ holds, then for any $\alpha \in (0,1)$*

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \limsup_{t\to\infty} \frac{d_t}{\psi_t} \le \bar{l}_{1-\alpha} \Rightarrow \limsup_{t\to\infty} \frac{\ln \beta\left(\delta_t^{+,\alpha_t}\right)}{\psi_t} \le \bar{l}_{1-\alpha}, \tag{2.4.2}$$

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \liminf_{t\to\infty} \frac{\ln \beta\left(\delta_t^{+,\alpha_t}\right)}{\psi_t} \ge \underline{l}_{1-\alpha} \Rightarrow \liminf_{t\to\infty} \frac{d_t}{\psi_t} \ge \underline{l}_{1-\alpha} \tag{2.4.3}$$

where

$$\bar{l}_p = \inf\{u: L(u) > p\}, \qquad \underline{l}_p = \sup\{u: L(u) < p\} \tag{2.4.4}$$

for $p \in (0,1)$.

PROOF. We have

$$\alpha_t = \mathsf{P}^t(Y_t > y_t) + \varepsilon_t \mathsf{P}^t(Y_t = y_t) \tag{2.4.5}$$

where $Y_t = \psi_t^{-1}\Lambda_t$ and $y_t = \psi_t^{-1} d_t$.

First we prove the first implication in (2.4.2). Suppose that $\alpha_t \to \alpha$, but

$$\limsup_{t\to\infty} y_t = \bar{y} > u$$

where u is a point of continuity of $L(x)$ such that $L(u) > 1-\alpha$. Let t_n be a sequence such that $y_{t_n} \to \bar{y}$ as $n \to \infty$. Then

$$\alpha = \lim_{t\to\infty} \alpha_t \le \limsup_{n\to\infty} \mathsf{P}^{t_n}(Y_{t_n} \ge y_{t_n}) \le \limsup_{n\to\infty} \mathsf{P}^{t_n}(Y_{t_n} \ge u) = 1 - L(u) < \alpha.$$

This contradiction shows that $\limsup_{t\to\infty} y_t \le u$ for any point u of continuity of $L(x)$. This proves the first implication in (2.4.2).

The second implications in (2.4.2) and (2.4.3) follow from estimate (2.3.10).

Assume again that $\alpha_t \to \alpha$ as $t \to \infty$ and let $u < 0$ be a point of continuity of the function $L(x)$ such that $L(u) < 1 - \alpha$. By equality (2.1.8), we obtain

$$\begin{aligned} \beta\left(\delta_t^{+,\alpha_t}\right) &\ge \widetilde{\mathsf{E}}^t I(Y_t \ge u)\left(1 - \delta_t^{+,\alpha_t}\right) \ge \mathsf{E}^t I(Y_t \ge u)\left(1 - \delta_t^{+,\alpha_t}\right) z_t \\ &\ge e^{u\psi_t} \mathsf{E}^t I(Y_t \ge u)\left(1 - \delta_t^{+,\alpha_t}\right) \ge e^{u\psi_t}\left(\mathsf{P}^t(Y_t \ge u) - \alpha_t\right). \end{aligned}$$

This yields in view of condition $\Lambda 4$ that

$$\lim_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \ge u,$$

whence the first implication in (2.4.3) follows, since u is arbitrary. □

If the distribution function $L(x)$ in condition $\Lambda 4$ is continuous, then Theorem 2.4.1 can be made more precise. First we give a necessary definition.

DEFINITION 2.4.1. Let $p \in [0,1]$. Let S be a probability distribution and let $S(x)$ be its distribution function. Any number $x_p \in \mathbf{R}^1$ such that

$$S(x_p) \le p \le S(x_p + 0)$$

is called a *quantile of order* $p \in [0,1]$ or, simply, a *p-quantile* of a probability distribution S. Such a number x_p is also called a p-quantile of the distribution function $S(x)$. If for some $p \in [0,1]$ we have $S(x) > p$ for all $x \in \mathbf{R}^1$, then we put $x_p = -\infty$. If $S(x) < p$ for all $x \in \mathbf{R}^1$, then we put $x_p = +\infty$.

REMARK 2.4.2. If L is the probability distribution appearing in condition $\Lambda 4$, then it is clear that for $p \in (0,1)$, the p-quantile l_p of the distribution L and the quantities $\underline{l}_p$ and $\bar{l}_p$ defined by relations (2.4.4) are related in the following way:

$$\underline{l}_p \le l_p \le \bar{l}_p.$$

THEOREM 2.4.2. *Assume that condition* $\Lambda 4$ *holds and the function* $L(x)$ *is continuous and strictly increasing in the interval* $(\underline{l}, \bar{l})$ *where*

$$\underline{l} = \sup\{x: L(x) = 0\}, \qquad \bar{l} = \inf\{x: L(x) = 1\}. \tag{2.4.6}$$

(*We agree that* $\sup(\varnothing) = -\infty$ *and* $\inf(\varnothing) = \infty$.) *Then for any* $\alpha \in (0,1)$

$$\lim_{t\to\infty} \alpha_t = \alpha \Longleftrightarrow \lim_{t\to\infty} \frac{d_t}{\psi_t} = l_{1-\alpha} \Longleftrightarrow \lim_{t\to\infty} \frac{\ln \beta(\delta_t^{+,\alpha_t})}{\psi_t} = l_{1-\alpha} \tag{2.4.7}$$

where $p \in (0,1)$, $l_p = \underline{l}_p = \bar{l}_p$ *is a p-quantile of the distribution* L, *and where the quantities* $\underline{l}_p$ *and* $\bar{l}_p$ *are defined by* (2.4.4).

PROOF. Since the function $L(x)$ is continuous and strictly increasing in the interval $(\underline{l}_p, \bar{l}_p)$, we have $l_p = \underline{l}_p = \bar{l}_p$ for $p \in (0,1)$. Put $L_t(x) = \mathsf{P}^t(Y_t < x)$ where $Y_t = \psi_t^{-1}\Lambda_t$. Then for any $\varepsilon > 0$ and $y_t \in \mathbf{R}^1$

$$\begin{aligned} \mathsf{P}^t(Y_t = y_t) &\le L_t(y_t + \varepsilon) - L_t(y_t) \\ &= [L_t(y_t + \varepsilon) - L(y_t + \varepsilon)] \\ &\quad - [L_t(y_t) - L(y_t)] + [L(y_t + \varepsilon) - L(y_t)]. \end{aligned} \tag{2.4.8}$$

By the Pólya theorem (see, for example, [**16**]) we have

$$\lim_{t\to\infty} \sup_y |L_t(y) - L(y)| = 0. \tag{2.4.9}$$

Since the continuity of the function $L(x)$ implies its uniform continuity, we obtain from (2.4.8) and (2.4.9)

$$\lim_{t\to\infty} \mathsf{P}^t(Y_t = y_t) = 0. \tag{2.4.10}$$

Therefore it follows from (2.4.5) and (2.4.10) that

$$\lim_{t\to\infty} \alpha_t = \alpha \Longleftrightarrow \lim_{t\to\infty} L_t(y_t) = 1 - \alpha, \tag{2.4.11}$$

whence

$$\lim_{t\to\infty} \alpha_t = \alpha \Longleftrightarrow \lim_{t\to\infty} y_t = l_{1-\alpha}, \tag{2.4.12}$$

since $L(x)$ is continuous and strictly increasing in the interval $(\underline{l}, \bar{l})$.

By (2.1.9), we have for any $\varepsilon > 0$

$$\begin{aligned}\beta\left(\delta_t^{+,\alpha_t}\right) &\geq \mathsf{E}^t z_t\left(1-\delta_t^{+,\alpha_t}\right) \geq \mathsf{E}^t I(y_t-\varepsilon \leq Y_t) z_t\left(1-\delta_t^{+,\alpha_t}\right) \\ &\geq \mathsf{P}^t(y_t-\varepsilon \leq Y_t < y_t)\exp\bigl((y_t-\varepsilon)\psi_t\bigr).\end{aligned} \tag{2.4.13}$$

Using (2.4.13) we get

$$\lim_{t\to\infty} y_t = l_{1-\alpha} \Rightarrow \lim_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) = l_{1-\alpha} \tag{2.4.14}$$

in view of the second implication in (2.4.2) and since ε is arbitrary. The relation

$$\lim_{t\to\infty} y_t = l_{1-\alpha} \Leftarrow \lim_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) = l_{1-\alpha}$$

follows from

$$\lim_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) = l_{1-\alpha} \Rightarrow \limsup_{t\to\infty} y_t \leq l_{1-\alpha} \tag{2.4.15}$$

in view of the second implication in (2.4.3).

We prove (2.4.15) by contradiction. Let

$$\lim_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) = l_{1-\alpha}, \qquad \limsup_{t\to\infty} y_t > l_{1-\alpha}, \tag{2.4.16}$$

and let (t_n) be a sequence such that $t_n \to \infty$ and $y_{t_n} \to \limsup_{t\to\infty} y_t$ as $n \to \infty$. Then estimate (2.4.13) implies that

$$\limsup_{t\to\infty} \psi_t^{-1} \ln \beta\left(\delta_t^{+,\alpha_t}\right) \geq \limsup_{n\to\infty} \psi_{t_n}^{-1} \ln \beta\left(\delta_{t_n}^{+,\alpha_{t_n}}\right) > l_{1-\alpha},$$

contradicting the equality in (2.4.16). This proves implication (2.4.15). By combining (2.4.12), (2.4.14), (2.4.15), and (2.4.3), we obtain (2.4.7). □

REMARK 2.4.3. If the limit distribution L in condition $\Lambda 4$ is singular and concentrated at a point $x_0 \in (-\infty, 0)$, then the law of large numbers $\Lambda 1$ holds, and the statements of the previous section apply. If the function $L(x)$ in condition $\Lambda 4$ is discontinuous at a point $x_0 \in (-\infty, 0)$ such that $0 < L(x_0) < L(x_0+0) < 1$, then Theorem 2.4.1 implies the following result: *If $\alpha_t \to \alpha$ as $t \to \infty$, then we have for any $\alpha \in (1 - L(x_0+0), 1 - L(x_0))$*

$$\lim_{t\to\infty} \frac{d_t}{\psi_t} = l_{1-\alpha}, \qquad \lim_{t\to\infty} \frac{\ln \beta\left(\delta_t^{+,\alpha_t}\right)}{\psi_t} = l_{1-\alpha} \tag{2.4.17}$$

where $l_{1-\alpha} = x_0$. Moreover, if condition $\Lambda 4$ holds and $L^{-1}(\alpha) = \inf\{u: L(u) > \alpha\}$, $\alpha \in [0,1]$, is the inverse function of $L(x)$, then Theorems 2.4.1 and 2.4.2 imply the following result: *If $\alpha_t \to \alpha$ as $t \to \infty$, then for any $\alpha \in (0,1)$ that is a point of continuity of $L^{-1}(\alpha)$, relations* (2.4.17) *hold with $l_{1-\alpha} = L^{-1}(\alpha)$.*

REMARK 2.4.4. If the assumptions of Theorem 2.4.2 hold and δ_t^{+,α_t} is the Neyman–Pearson test whose level α_t has a limit value $\alpha \in (0,1)$, then Theorem 2.4.2 implies that

$$d_t = l_{1-\alpha}\psi_t + o(\psi_t), \qquad \ln \beta\left(\delta_t^{+,\alpha_t}\right) = l_{1-\alpha}\psi_t + o(\psi_t).$$

Here, $l_{1-\alpha}$ is strictly decreasing from 0 to $-\infty$ as α increases from 0 to 1. Therefore the rate of decay of $\beta(\delta_t^{+,\alpha_t})$ depends on α for each particular value of α. This differs

crucially from the case where the law of large numbers $\Lambda 1$ holds; in the latter case, the rate of decay of $\beta(\delta_t^{+,\alpha_t})$ is independent of α (see the preceding section).

Second-order behavior of the Neyman–Pearson test. Now we consider the case where the law of large numbers holds for Λ_t. In this case, the rate of decay of the type II error probability of the Neyman–Pearson test δ_t^{+,α_t} is independent of $\alpha = \lim_{t\to\infty} \alpha_t$ if $\alpha \in (0,1)$. Below is a condition enabling us to evaluate the second term in the asymptotic expansion of $\ln \beta(\delta_t^{+,\alpha_t})$ as $t \to \infty$.

$\Lambda 5$. $\mathcal{L}\left(\varphi_t^{-1}(\Lambda_t + \psi_t) | P^t\right) \xrightarrow{w} L$ as $t \to \infty$ where φ_t and ψ_t are positive functions such that $\varphi_t \to \infty$ and $\psi_t \to \infty$ as $t \to \infty$, $\varphi_t = o(\psi_t)$, and L is a probability distribution on $\mathbf{R}^1$ whose distribution function is $L(x)$.

It is easy to see that $\Lambda 5 \Rightarrow \Lambda 1$ for $\chi_t = \psi_t$.

THEOREM 2.4.3. *If condition* $\Lambda 5$ *holds, then for any* $\alpha \in (0,1)$

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \limsup_{t\to\infty} \frac{d_t + \psi_t}{\varphi_t} \le \bar{l}_{1-\alpha} \tag{2.4.18}$$
$$\Rightarrow \limsup_{t\to\infty} \frac{\ln \beta(\delta_t^{+,\alpha_t}) + \psi_t}{\varphi_t} \le \bar{l}_{1-\alpha},$$

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \liminf_{t\to\infty} \frac{\ln \beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t}{\varphi_t} \ge \underline{l}_{1-\alpha} \tag{2.4.19}$$
$$\Rightarrow \liminf_{t\to\infty} \frac{d_t + \psi_t}{\varphi_t} \ge \underline{l}_{1-\alpha}$$

where $\underline{l}_p$ *and* $\bar{l}_p$ *are defined by* (2.4.4).

PROOF. It is sufficient to follow the proof of Theorem 2.4.1 with appropriate modifications. First of all we note that equality (2.4.5) holds with

$$Y_t = \varphi_t^{-1}(\Lambda_t + \psi_t), \qquad y_t = \varphi_t^{-1}(d_t + \psi_t). \tag{2.4.20}$$

The proof of the first implication in (2.4.18) is the same as that of the first implication in (2.4.2). The second implications in (2.4.18) and (2.4.19) follow from the estimate (2.3.10).

Now let $\alpha_t \to \alpha$ as $t \to \infty$ and let $u \in \mathbf{R}^1$ be a point of continuity of the function $L(x)$ such that $L(u) < 1 - \alpha$. We have

$$\beta\left(\delta_t^{+,\alpha_t}\right) \ge \mathsf{E}^t I(Y_t \ge u)\left(1 - \delta_t^{+,\alpha_t}\right) z_t \ge \exp(u\varphi_t - \psi_t)\left[\mathsf{P}^t(Y_t \ge u) - \alpha_t\right]$$

by equality (2.1.8), whence the first implication in (2.4.19) follows. □

The following result improves Theorem 2.4.3 for the case where the function $L(x)$ in condition $\Lambda 5$ is continuous.

THEOREM 2.4.4. *Assume that condition* $\Lambda 5$ *holds and the function* $L(x)$ *is continuous and strictly increasing on the interval* $(\underline{l}, \bar{l})$ *where* $\underline{l}$ *and* $\bar{l}$ *are defined by* (2.4.6). *Then for any* $\alpha \in (0,1)$

$$\lim_{t\to\infty} \alpha_t = \alpha \Longleftrightarrow \lim_{t\to\infty} \frac{d_t + \psi_t}{\varphi_t} = l_{1-\alpha} \Longleftrightarrow \lim_{t\to\infty} \frac{\ln \beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t}{\varphi_t} = l_{1-\alpha} \tag{2.4.21}$$

where l_p *is a* p*-quantile of the distribution* L.

PROOF. It suffices to repeat the proof of Theorem 2.4.2 with appropriate modifications. First, we should use equality (2.4.5) where Y_t and y_t are given by (2.4.20). By literally repeating the proof of equivalence (2.4.12), the first equivalence in (2.4.21) is proved.

Next, we have for any $\varepsilon > 0$

$$\begin{aligned}\beta\left(\delta_t^{+,\alpha_t}\right) &\geq \mathsf{E}^t I(y_t - \varepsilon \leq Y_t) z_t \left(1 - \delta_t^{+,\alpha_t}\right) \\ &\geq \mathsf{P}^t(y_t - \varepsilon \leq Y_t < y_t) \exp((y_t - \varepsilon)\varphi_t - \psi_t).\end{aligned} \tag{2.4.22}$$

Then the second implication in (2.4.18) implies

$$\lim_{t\to\infty} y_t = l_{1-\alpha} \Rightarrow \lim_{t\to\infty} \varphi_t^{-1}\left(\ln\beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t\right) = l_{1-\alpha}, \tag{2.4.23}$$

since ε is arbitrary in (2.4.22) and $\bar{l}_p = l_p$ for $p \in (0,1)$ in view of the continuity and strict monotonicity of the function $L(x)$ in the interval $(\underline{l}, \bar{l})$. Further, the relation

$$\lim_{t\to\infty} y_t = l_{1-\alpha} \Rightarrow \lim_{t\to\infty} \varphi_t^{-1}\left(\ln\beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t\right) = l_{1-\alpha}$$

follows from the second implication in (2.4.19) if

$$\lim_{t\to\infty} \varphi_t^{-1}\left(\ln\beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t\right) = l_{1-\alpha} \Rightarrow \limsup_{t\to\infty} y_t \leq l_{1-\alpha}. \tag{2.4.24}$$

Relation (2.4.24) is readily proved by contradiction. Indeed, assume that

$$\lim_{t\to\infty} \varphi_t^{-1}\left(\ln\beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t\right) = l_{1-\alpha}, \qquad \limsup_{t\to\infty} y_t > l_{1-\alpha} \tag{2.4.25}$$

and let (t_n) be a sequence such that $t_n \to \infty$ and $y_{t_n} \to \limsup_{t\to\infty} y_t$ as $n \to \infty$. In view of estimate (2.4.22)

$$\limsup_{t\to\infty} \varphi_t^{-1}\left(\ln\beta\left(\delta_t^{+,\alpha_t}\right) + \psi_t\right) \geq \varlimsup_{n\to\infty} \varphi_{t_n}^{-1}\left(\ln\beta\left(\delta_{t_n}^{+,\alpha_{t_n}}\right) + \psi_{t_n}\right) > l_{1-\alpha},$$

contradicting the equality in (2.4.25). This proves implication (2.4.24), completing the proof of the second equivalence in (2.4.21). □

REMARK 2.4.5. If the assumptions of Theorem 2.4.4 hold and if $\alpha_t \to \alpha \in (0,1)$ as $t \to \infty$, then Theorem 2.4.4 gives the following asymptotic expansions for the Neyman–Pearson test δ_t^{+,α_t}:

$$d_t = -\psi_t + l_{1-\alpha}\varphi_t + o(\varphi_t), \tag{2.4.26}$$

$$\ln\beta\left(\delta_t^{+,\alpha_t}\right) = -\psi_t + l_{1-\alpha}\varphi_t + o(\varphi_t). \tag{2.4.27}$$

Expansions (2.4.26) and (2.4.27) show that, under these assumptions, the asymptotic behavior of the test δ_t^{+,α_t} depends, in the second term of the asymptotic expansion, on the limit value α of the level α_t (cf. Remark 2.4.4 and relation (2.3.12) in the case where the law of large numbers $\Lambda 1$ holds).

EXAMPLE 2.4.1. Let an observation ξ be the vector

$$\xi^n = (\xi_1, \xi_2, \dots, \xi_n), \qquad n = 1, 2, \dots,$$

where the random variables $\xi_1, \xi_2, \dots, \xi_n$ form a first order autoregressive process

$$\xi_i = \theta \xi_{i-1} + w_i, \qquad i = 1, 2, \dots, \ \xi_0 = 0, \tag{2.4.28}$$

where $\theta \in \mathbf{R}$ is an unknown parameter and $w_1, w_2, \dots$ are independent Gaussian random variables with the $\mathcal{N}(0,1)$ distribution (which are independent of θ). Denote by P_θ^n the probability measure generating the distribution of the observation ξ^n. Let the measures P_θ^n and $\mathsf{P}_{\tilde\theta}^n$ correspond to the hypotheses H^n and $\widetilde{H}^n$, respectively, where θ and $\tilde\theta$ are some points in $\mathbf{R}$ such that $|\theta| > 1$ and $\tilde\theta \neq \theta$. We assume that θ is independent of n, while $\tilde\theta$ depends, generally speaking, on n. We write $\tilde\theta = \theta_n$ if $\tilde\theta$ depends on n. It is clear that the measures P_θ^n and $\mathsf{P}_{\tilde\theta}^n$ generate Gaussian distributions. Moreover, $\mathsf{P}_{\tilde\theta}^n \sim \mathsf{P}_\theta^n$ and, in view of (2.4.28), the logarithm of the density of the measure $\mathsf{P}_{\tilde\theta}^n$ with respect to the measure P_θ^n can be represented as follows (P_θ^n-a.s.):

$$\Lambda_n = (\tilde\theta - \theta) \sum_{i=1}^n \xi_{i-1} w_i - \frac{1}{2} (\tilde\theta - \theta)^2 \sum_{i=1}^n \xi_{i-1}^2. \tag{2.4.29}$$

By (2.4.28), we have

$$\mathsf{E}_\theta^n \left(\theta^{-n-1} \xi_{n+1} - \theta^{-n} \xi_n \right)^2 = \mathsf{E}_\theta^n \left(\theta^{-n-1} w_{n+1} \right)^2 = \theta^{-2(n+1)}.$$

Using Proposition II.4.2 in [**42**] we obtain that $\theta^{-n}\xi_n$ is a Cauchy sequence with probability 1 and therefore the limit $\lim_{n\to\infty} \theta^{-n}\xi_n$ exists almost surely with respect to the probability P_θ^n. On the other hand (2.4.28) implies that the random variable ξ_i is normally distributed with mean 0 and variance

$$\mathsf{E}_\theta^n \xi_i^2 = \theta^{2i} \left(1 + \theta^{-2} + \theta^{-4} + \dots + \theta^{-2i} \right). \tag{2.4.30}$$

Thus

$$\sqrt{\theta^2 - 1}\, \theta^{-n} \xi_n \to \eta, \qquad n \to \infty, \tag{2.4.31}$$

almost surely with respect to the probability P_θ^n where the random variable η is normally $\mathcal{N}(0,1)$ distributed. Further, we obtain by (2.4.31)

$$\begin{aligned} \theta^{-2n} \left(\theta^2 - 1 \right)^2 \sum_{i=1}^n \xi_{i-1}^2 &= \left(\theta^2 - 1 \right)^2 \sum_{i=1}^n \left(\theta^{i-n} \xi_{n-i} \right)^2 \theta^{-2i} \\ &= \left(\theta^2 - 1 \right)^2 \sum_{i=1}^\infty \left(\theta^{i-n} \xi_{n-i} \right)^2 I_{[1,n]}(i) \theta^{-2i} \to \eta \end{aligned} \tag{2.4.32}$$

as $n \to \infty$ with probability 1. Here, we have used the dominated convergence, since for any ω where convergence (2.4.31) holds, there exists a constant $C(\omega)$ such that $|\theta^{-j} \xi_j(\omega)| \leq C(\omega)$ for all $j = 1, 2, \dots$. It is easy to show, in view of (2.4.28) and (2.4.30), that

$$\mathsf{E}_\theta^n \left(\sum_{i=1}^n \xi_{i-1} w_i \right)^2 = \sum_{i=1}^n \mathsf{E}_\theta^n \xi_{i-1}^2 \leq \sum_{i=1}^{2n} \theta^{2(i-1)} \frac{\theta^2}{\theta^2 - 1} \leq \frac{\theta^{2n}}{(\theta^2 - 1)^2}. \tag{2.4.33}$$

Now assume that $\tilde{\theta} = \theta_n$ depends, generally speaking, on n and $\theta^{2n}\Delta_n^2 \to \infty$ as $n \to \infty$ where $\Delta_n = \theta_n - \theta$. Then (2.4.29) implies the following representation (P_θ^n-a.s.):

$$\psi_n^{-1}\Lambda_n = 2\left(\theta^2-1\right)^2\Delta_n^{-1}\theta^{-2n}\sum_{i=1}^n \xi_{i-1}w_i - \left(\theta^2-1\right)^2\theta^{-2n}\sum_{i=1}^n \xi_{i-1}^2 \tag{2.4.34}$$

where $\psi_n = 2^{-1}(\theta^2-1)^{-2}\theta^{2n}\Delta_n^2$. But, in view of estimate (2.4.33), we have

$$\mathsf{E}_\theta^n\left(\theta^{-2n}\Delta_n^{-1}\sum_{i=1}^n \xi_{i-1}w_i\right)^2 \le \Delta_n^{-2}\theta^{-2n}\left(\theta^2-1\right)^{-2} \to 0, \qquad n \to \infty. \tag{2.4.35}$$

Relations (2.4.32), (2.4.34), and (2.4.35) imply that condition $\Lambda 4$ holds where $\psi_n = 2^{-1}(\theta^2-1)^{-2}\theta^{2n}\Delta_n^2$, L is the distribution of the random variable $-\eta^2$ whose distribution function is

$$L(x) = \mathsf{P}\left(-\eta^2 < x\right) = 2\left(1-\Phi(\sqrt{-x})\right), \qquad x \le 0,$$

$L(x) = 1$ for $x > 1$, and $\Phi(x)$ is the distribution function of the normal $\mathcal{N}(0,1)$ distribution.

Since the function $L(x)$ is continuous, Lemma 2.4.1 implies that the complete asymptotic distinguishability $(H^n) \triangle (\widetilde{H}^n)$ holds and, moreover, Theorem 2.4.2 applies. Observe that $\underline{l} = -\infty$, $\bar{l} = 0$, and for any $p \in (0,1)$ a p-quantile of the distribution L can be represented in the form $l_p = -z_{1-p/2}^2$ where z_p is a p-quantile of the distribution $\mathcal{N}(0,1)$. Therefore relations (2.4.7) become of the following form for any $\alpha \in (0,1)$:

$$\lim_{n\to\infty}\alpha_n = \alpha \iff \lim_{n\to\infty}\frac{d_n}{\psi_n} = -z_{\frac{1+\alpha}{2}}^2 \iff \lim_{n\to\infty}\frac{\ln\beta\left(\delta_n^{+,\alpha_n}\right)}{\psi_n} = -z_{\frac{1+\alpha}{2}}^2.$$

EXAMPLE 2.4.2. Let an observation be a sample $\xi^n = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables. Assume that the distribution of ξ_i under the hypothesis H^n is generated by a measure P whose density with respect to some σ-finite measure μ is $p(x)$, while under the hypothesis $\widetilde{H}^n$ the distribution is given by a measure $\widetilde{\mathsf{P}}$ having density $\widetilde{p}(x)$ with respect to μ. Consider the likelihood ratios

$$z_n(x) = \prod_{i=1}^n z(x_i), \qquad z(x_i) = \frac{p(x_i)}{\widetilde{p}(x_i)}, \qquad x = (x_1, x_2, \ldots, x_n).$$

Put

$$\Lambda_n(x) = \ln z_n(x), \qquad \lambda(x_i) = \ln z(x_i), \qquad x = (x_1, x_2, \ldots, x_n),$$
$$\Lambda_n = \Lambda_n(\xi^n), \qquad \lambda_i = \ln z(\xi_i), \qquad i = 1, 2, \ldots, n.$$

Then

$$\Lambda_n = \sum_{i=1}^n \lambda_i, \qquad n = 1, 2, \ldots,$$
$$I(\mathsf{P}^n|\widetilde{\mathsf{P}}^n) = nI(\mathsf{P}|\widetilde{\mathsf{P}}), \qquad I(\mathsf{P}|\widetilde{\mathsf{P}}) = -\mathsf{E}\lambda_1,$$

where P^n and $\widetilde{\mathsf{P}}^n$ are the measures generating the distribution of ξ^n under the hypothesis H^n and $\widetilde{H}^n$, respectively (see Section 2.3).

Assume that

$$0 < I(\mathsf{P}|\widetilde{\mathsf{P}}) < \infty, \qquad 0 < \sigma^2(\mathsf{P}|\widetilde{\mathsf{P}}) = \mathsf{D}\lambda_1 < \infty$$

where the symbol D stands for the variance under the hypothesis H^n. Then the central limit theorem [**47**] implies

$$\mathcal{L}\left(\sqrt{n}\sigma(\mathsf{P}|\widetilde{\mathsf{P}})\left(\Lambda_n + nI(\mathsf{P}|\widetilde{\mathsf{P}})\right)|\mathsf{P}^n\right) \xrightarrow{w} \mathcal{N}(0,1), \qquad n \to \infty.$$

Therefore condition $\Lambda 5$ holds with

$$\varphi_n = \sqrt{n}\sigma(\mathsf{P}|\widetilde{\mathsf{P}}), \qquad \psi_n = nI(\mathsf{P}|\widetilde{\mathsf{P}}), \qquad L = \mathcal{N}(0,1).$$

Thus the assumptions of Theorem 2.4.4 hold with $\underline{l} = -\infty$, $\bar{l} = \infty$, and $L(x) = \Phi(x)$. Hence relation (2.4.21) holds for any $\alpha \in (0,1)$ where $l_p = z_p$ is a p-quantile of the distribution $\mathcal{N}(0,1)$. Therefore the following asymptotic expansion holds for the Neyman–Pearson test δ_n^{+,α_n} if $\alpha_n \to \alpha \in (0,1)$, $n \to \infty$:

$$\ln \beta(\delta_n^{+,\alpha_n}) = -nI(\mathsf{P}|\widetilde{\mathsf{P}}) - \sqrt{n}\sigma(\mathsf{P}|\widetilde{\mathsf{P}})z_{1-\alpha} + o(\sqrt{n})$$

(cf. relation (2.3.22)).

EXAMPLE 2.4.3. Let an observation be a column-vector $\xi^n = (\xi_1, \xi_2, \ldots, \xi_n)'$ having Gaussian distributions P^n and $\widetilde{\mathsf{P}}^n$ under the hypotheses H^n and $\widetilde{H}^n$, respectively, namely

$$\mathcal{L}(\xi_n|H^n) = \mathcal{N}(\mu, \sigma^2 R_n), \qquad \mathcal{L}\left(\xi_n|\widetilde{H}^n\right) = \mathcal{N}(\widetilde{\mu}, \widetilde{\sigma}^2 R_n)$$

where $\mathcal{N}(a, B)$ is the Gaussian distribution with the vector of means a and matrix of the second order mixed moments B, $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$, $\widetilde{\mu} = (\widetilde{\mu}_1, \widetilde{\mu}_2, \ldots, \widetilde{\mu}_n)'$, and $R_n = (R_{ij})$ is an $n \times n$-matrix. This model can be written as

$$\xi^n = \theta + v\zeta^n$$

where $\zeta^n = (\zeta_1, \zeta_2, \ldots, \zeta_n)'$ is a Gaussian vector with distribution

$$\mathcal{L}(\zeta|H^n) = \mathcal{L}\left(\zeta|\widetilde{H}^n\right) = \mathcal{N}(0, R_n)$$

and where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ and $v \in (0, \infty)$ are unknown parameters such that $\theta = \mu$ and $v = \sigma$ under the hypothesis H^n, while $\theta = \widetilde{\mu}$ and $v = \widetilde{\sigma}$ under the hypothesis $\widetilde{H}^n$.

Assume that the matrix R_n is nondegenerate for all $n = 1, 2, \ldots$. Then the measures $\widetilde{\mathsf{P}}^n$ and P^n are mutually absolutely continuous. Therefore the likelihood ratio z_n has the form $z_n(x) = d\widetilde{\mathsf{P}}^n/d\mathsf{P}^n(x)$, $x = (x_1, x_2, \ldots, x_n)$, where

$$\begin{aligned}\Lambda_n(x) = \ln z_n(x) &= n \ln \frac{\sigma}{\widetilde{\sigma}} - \frac{1}{2\widetilde{\sigma}^2}(x - \widetilde{\mu})' R_n^{-1}(x - \widetilde{\mu}) + \frac{1}{2\sigma^2}(x - \mu)' R_n^{-1}(x - \mu) \\ &= n \ln \frac{\sigma}{\widetilde{\sigma}} + \frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\widetilde{\sigma}^2}\right) y' R_n^{-1} y - \frac{1}{\widetilde{\sigma}^2} m' R_n^{-1} y - \frac{1}{2\widetilde{\sigma}^2} m' R_n^{-1} m\end{aligned}$$

and $m = \widetilde{\mu} - \mu$ and $y = x - \mu$. This implies that

$$\begin{aligned} \Lambda_n &= \Lambda_n(\xi^n) \\ &= n \ln \frac{\sigma}{\widetilde{\sigma}} + \frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\widetilde{\sigma}^2}\right) \eta' R_n^{-1} \eta - \frac{1}{\widetilde{\sigma}^2} m' R_n^{-1} \eta - \frac{1}{2\widetilde{\sigma}^2} m' R_n^{-1} m \end{aligned} \tag{2.4.36}$$

where $\eta = \xi^n - \mu$. Note that Theorem 1.4.1 in [**38**] implies $\mathsf{E}^n \eta' R_n^{-1} \eta = n\sigma^2$, whence we obtain

$$I(\mathsf{P}^n | \widetilde{\mathsf{P}}^n) = -\mathsf{E}^n \Lambda_n = \frac{n}{2}\left(\frac{\sigma^2}{\widetilde{\sigma}^2} - 1 - \ln \frac{\sigma^2}{\widetilde{\sigma}^2}\right) + \frac{1}{2\widetilde{\sigma}^2} m' R_n^{-1} m \tag{2.4.37}$$

by equality (2.4.36).

Put $c_n = m' R_n^{-1} m$ and let $I(\mathsf{P}^n | \widetilde{\mathsf{P}}^n) \to \infty$ as $n \to \infty$. Considering (2.4.37) we distinguish the following three cases.

a) Let $\widetilde{\sigma} = \sigma$ and $c_n \to \infty$ as $n \to \infty$. Then (2.4.36) and (2.4.37) imply for any $n = 1, 2, \ldots$ that

$$\mathcal{L}\left(c_n^{-1/2} \widetilde{\sigma} \left[\Lambda_n + \frac{c_n}{2\widetilde{\sigma}^2}\right] \mid \mathsf{P}^n\right) = \mathcal{N}(0,1)$$

and thus condition $\Lambda 5$ holds with

$$\varphi_n = \frac{c_n^{1/2}}{\widetilde{\sigma}}, \qquad \psi_n = \frac{c_n}{2\widetilde{\sigma}^2}, \qquad L = \mathcal{N}(0,1), \qquad L(x) = \Phi(x).$$

Therefore the assumptions of Theorem 2.4.4 are satisfied and thus

$$\ln \beta\left(\delta_n^{+,\alpha_n}\right) = -\frac{c_n}{2\widetilde{\sigma}^2} - \frac{c_n^{1/2}}{\widetilde{\sigma}} z_{1-\alpha} + o\left(c_n^{1/2}\right) \tag{2.4.38}$$

if $\alpha_n \to \alpha \in (0,1)$ as $n \to \infty$.

b) Let $\widetilde{\sigma} \neq \sigma$ and $n/c_n \to \infty$ as $n \to \infty$. Equalities (2.4.36) and (2.4.37) imply that

$$\frac{1}{\sqrt{n}}\left[\Lambda_n + I(\mathsf{P}^n | \widetilde{\mathsf{P}}^n)\right] = \frac{1}{\sqrt{2}}\left(\frac{\sigma^2}{\widetilde{\sigma}^2} - 1\right) G_n - H_n \tag{2.4.39}$$

where

$$H_n = \frac{1}{\sqrt{n}\widetilde{\sigma}^2} m' R_n^{-1} \eta, \qquad G_n = \frac{1}{\sqrt{2n}}\left(\frac{1}{\widetilde{\sigma}^2} \eta' R_n^{-1} \eta - n\right). \tag{2.4.40}$$

It is clear that $H_n \to 0$ in probability as $n \to \infty$ under the hypothesis H^n. By the central limit theorem and in view of Theorem 1.4.1 in [**38**], we have

$$\mathcal{L}(G_n | \mathsf{P}^n) \xrightarrow{w} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

Therefore condition $\Lambda 5$ holds with

$$\varphi_n = \sqrt{\frac{n}{2}} \left|\frac{\sigma^2}{\widetilde{\sigma}^2} - 1\right|, \qquad \psi_n = I(\mathsf{P}^n | \widetilde{\mathsf{P}}^n), \qquad L = \mathcal{N}(0,1).$$

Then Theorem 2.4.4 implies that

$$\ln \beta\left(\delta_n^{+,\alpha_n}\right) = -I(\mathsf{P}^n | \widetilde{\mathsf{P}}^n) - \sqrt{\frac{n}{2}} \left|\frac{\sigma^2}{\widetilde{\sigma}^2} - 1\right| z_{1-\alpha} + o(\sqrt{n})$$

if $\alpha_n \to \alpha \in (0,1)$ as $n \to \infty$. Observe that this expansion takes the following form in the case of $c_n = o(\sqrt{n})$:

$$\ln \beta\left(\delta_n^{+,\alpha_n}\right) = -\frac{n}{2}\left(\frac{\sigma^2}{\widetilde{\sigma}^2} - 1 - \ln\frac{\sigma^2}{\widetilde{\sigma}^2}\right) - \sqrt{\frac{n}{2}}\left|\frac{\sigma^2}{\widetilde{\sigma}^2} - 1\right| z_{1-\alpha} + o(\sqrt{n}).$$

c) Let $\widetilde{\sigma} \neq \sigma$ and $n/c_n \to 0$ as $n \to \infty$. Then relations (2.4.39) and (2.4.40) imply that

$$\mathcal{L}\left(\frac{\widetilde{\sigma}^2}{\sigma\sqrt{c_n}}\left[\Lambda_n + I\left(\mathsf{P}^n|\widetilde{\mathsf{P}}^n\right)\right] \mid \mathsf{P}^n\right) \xrightarrow{w} \mathcal{N}(0,1)$$

as $n \to \infty$. Therefore condition $\Lambda 5$ holds with

$$\varphi_n = \frac{\sigma}{\widetilde{\sigma}^2} c_n^{1/2}, \qquad \psi_n = I\left(\mathsf{P}^n|\widetilde{\mathsf{P}}^n\right), \qquad L = \mathcal{N}(0,1).$$

Hence Theorem 2.4.4 implies that

$$\ln \beta\left(\delta_n^{+,\alpha_n}\right) = -I\left(\mathsf{P}^n|\widetilde{\mathsf{P}}^n\right) - \frac{\sigma}{\widetilde{\sigma}^2} c_n^{1/2} z_{1-\alpha} + o\left(c_n^{1/2}\right) \quad \text{if } \alpha_n \to \alpha \in (0,1).$$

The latter expansion becomes of the following form if $n = o(c_n^{1/2})$:

$$\ln \beta\left(\delta_n^{+,\alpha_n}\right) = -\frac{1}{2\widetilde{\sigma}^2} c_n - \frac{\sigma}{\widetilde{\sigma}^2} c_n^{1/2} z_{1-\alpha} + o\left(c_n^{1/2}\right)$$

(cf. (2.4.38)).

Examples where conditions $\Lambda 4$ and $\Lambda 5$ hold and the observation ξ^t is a stochastic process on the interval $[0,t]$ can be found in the monograph [**37**].

2.5. Contiguous families of hypotheses

Relative compactness and tightness of a family of probability measures. The concepts of relative compactness and tightness of families of probability measures play a fundamental role in studying contiguous families of statistical hypotheses. Assume that all underlying measures are defined on a metric space $(E, \mathscr{E}, \rho)$ equipped with a distance ρ where the σ-algebra $\mathscr{E}$ is generated by the metric ρ. In what follows, we often consider the case $(E, \mathscr{E}) = (\mathbf{R}^m, \mathscr{B}^m)$ with the Euclidean metric ρ.

DEFINITION 2.5.1. A family of probability measures $\mathscr{Q} = (\mathbf{Q}_u; u \in \mathfrak{U})$ is called *relatively compact* if any sequence of measures belonging to $\mathscr{Q}$ contains a subsequence that converges weakly to a probability measure.

Note that this definition does not assume that the limit probability measure belongs to the family $\mathscr{Q}$.

DEFINITION 2.5.2. A family of probability measures $\mathscr{Q} = (\mathbf{Q}_u; u \in \mathfrak{U})$ is called *tight* if for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset E$ such that

$$\sup\left\{\mathbf{Q}_u(E \setminus K_\varepsilon); u \in \mathfrak{U}\right\} \leq \varepsilon.$$

The following result is fundamental for the theory of weak convergence of probability measures.

THEOREM 2.5.1 (Prokhorov's theorem). *Assume that* $\mathcal{Q} = (\mathbf{Q}_u; u \in \mathfrak{U})$ *is a family of probability measures defined on a complete separable metric space* $(E, \mathscr{E}, \rho)$. *The family* $\mathcal{Q}$ *is relatively compact if and only if* $\mathcal{Q}$ *is tight.*

The proof of this theorem can be found in various textbooks on probability theory (see, for example, [**3, 47**]).

Consider a family of probability measures $\mathcal{Q} = (\mathbf{Q}_t; t \in \mathbf{R}_+)$ on the space $(\mathbf{R}, \mathscr{B})$ where $\mathbf{Q}_t$ is the probability measure defining the likelihood ratio z_t with respect to the measure $\widetilde{\mathsf{P}}^t$. If a family of hypotheses $(\widetilde{H}^t)$ is contiguous to a family (H^t), then Theorem 2.2.3 implies that the family (z_t) is tight with respect to $(\widetilde{\mathsf{P}}^t)$. Therefore the family of measures $\mathcal{Q}$ is tight and thus, by the Prokhorov theorem, relatively compact. In a similar way, if a family (H^t) is contiguous to the family $(\widetilde{H}^t)$, then the family of distributions of $\widetilde{z}_t$ with respect to the measure P^t is relatively compact. It is clear that the family (z_t) is tight with respect to (P^t). Therefore the family of distributions of $\Lambda_t = \ln z_t$ is relatively compact with respect to P^t, since $(H^t) \lhd (\widetilde{H}^t)$. Therefore if $(H^t) \lhd (\widetilde{H}^t)$, then every sequence of distributions of Λ_t with respect to P^t has a weakly convergent subsequence. For the sake of brevity we assume throughout this section that the distribution of Λ_t with respect to P^t is weakly convergent.

Weak convergence of the logarithm of the likelihood ratio. We introduce the following condition:

$\Lambda 6$. $\mathcal{L}\left(\Lambda_t | \mathsf{P}^t\right) \xrightarrow{w} L$ where L is a probability distribution on $\mathbf{R}$ whose distribution function is denoted by $L(x)$.

THEOREM 2.5.2. *If condition* $\Lambda 6$ *holds, then*

$$\mathcal{L}\left(\Lambda_t | \widetilde{\mathsf{P}}^t\right) \xrightarrow{w} \widetilde{L} \tag{2.5.1}$$

where $\widetilde{L}$ *is a probability distribution on the extended real line* $\overline{\mathbf{R}} = [-\infty, \infty]$ *whose distribution function is*

$$\widetilde{L}(x) = \int_{-\infty}^{x} e^y \, dL(y). \tag{2.5.2}$$

In this case, $\widetilde{L}(\infty) \leq 1$ *and*

$$\lim_{N \to \infty} \lim_{t \to \infty} \widetilde{\mathsf{P}}^t(\Lambda_t \geq N) = 1 - \widetilde{L}(\infty). \tag{2.5.3}$$

PROOF. By the Lebesgue decomposition we have for any $x \in \mathbf{R}$

$$\widetilde{\mathsf{P}}^t(\Lambda_t < x) = \mathsf{E}^t I(\ln z_t < x) z_t = \int_0^{\exp(x)} z \, dG_t(z) \tag{2.5.4}$$

where $G_t(z) = \mathsf{P}^t(z_t < z)$, $z \in \mathbf{R}_+$. By the Helly theorem [**47**], condition $\Lambda 6$ implies that

$$\lim_{t \to \infty} \int_0^x z \, dG_t(z) = \int_0^x z \, dG(z) \tag{2.5.5}$$

for any $x \in (0, \infty)$ such that $\ln x$ is a point of continuity of $L(x)$, where

$$G(z) = L(\ln z), \qquad z \in (0, \infty).$$

It is clear that

$$\int_0^{\exp(x)} z\,dG(z) = \int_{-\infty}^{y} e^x\,dL(x). \tag{2.5.6}$$

It follows from (2.5.4)–(2.5.6) that for any $y \in \mathbf{R}$ which is a point of continuity of $L(x)$,

$$\lim_{t\to\infty} \widetilde{\mathsf{P}}^t(\Lambda_t < y) = \int_{-\infty}^{y} e^x\,dL(x), \tag{2.5.7}$$

whence relations (2.5.1) and (2.5.2) follow. The inequality $\widetilde{L}(\infty) \le 1$ and relation (2.5.3) also follow from (2.5.7). □

COROLLARY 2.5.1. *We have*

$$\Lambda 6 \Rightarrow \lim_{t\to\infty} \overline{\alpha}_t = 1, \qquad \liminf_{t\to\infty} \overline{\beta}_t \ge \widetilde{L}(\infty). \tag{2.5.8}$$

PROOF. It is clear that for any $x \in \mathbf{R}$

$$1 - \overline{\alpha}_t = P^t(\Lambda_t = -\infty) \le \mathsf{P}^t(\Lambda_t < x), \qquad \overline{\beta}_t = \widetilde{\mathsf{P}}^t(\Lambda_t < \infty) \ge \widetilde{\mathsf{P}}^t(\Lambda_t < x).$$

Since the number $x \in \mathbf{R}$ is arbitrary, we obtain (2.5.8) from condition $\Lambda 6$ and relation (2.5.3). □

COROLLARY 2.5.2. *If condition $\Lambda 6$ holds for $L = \mathcal{N}(a, \sigma^2)$ where $a \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$, then $a \le -\sigma^2/2$ and (2.5.1) holds with the distribution function*

$$\widetilde{L}(x) = h\Phi\left(\frac{x - a - \sigma^2}{\sigma}\right) \tag{2.5.9}$$

where $h = \exp(a + \sigma^2/2)$.

PROOF. It is clear that

$$\int_{-\infty}^{x} e^y\,dL(y) = \int_{-\infty}^{x} e^y\,d\Phi\left(\frac{y-a}{\sigma}\right)dy = h\Phi\left(\frac{x-a-\sigma^2}{\sigma}\right).$$

By relation (2.5.7) and by equality (2.5.2), this implies (2.5.1) and equality (2.5.9). Since $\widetilde{\Lambda}(\infty) \le 1$ by Theorem 2.5.1, equality (2.5.9) implies that $a \le -\sigma^2/2$ in view of $\Phi(\infty) = 1$. □

REMARK 2.5.1. If condition $\Lambda 6$ holds, then it follows from Theorem 2.5.2 that the limit distribution $\widetilde{L}$ is, in general, a mixture of two probability distributions $\widetilde{L}^+$ and $\varepsilon_{\{\infty\}}$ weighted by $\widetilde{L}(\infty)$ and $1 - \widetilde{L}(\infty)$, respectively, that is,

$$\widetilde{L} = \widetilde{L}(\infty)\widetilde{L}^+ + (1 - \widetilde{L}(\infty))\varepsilon_{\{\infty\}}$$

where $\varepsilon_{\{\infty\}}$ is the Dirac measure concentrated at ∞ and $\widetilde{L}^+$ is a distribution on $\mathbf{R}$ determined by its distribution function $\widetilde{L}^+(x) = \widetilde{L}(x)/\widetilde{L}(\infty)$, $x \in \mathbf{R}$. In particular, by Corollary 2.5.2

$$\widetilde{L}(\infty) = h = \exp\left(a + \sigma^2/2\right), \qquad \widetilde{L}^+ = \mathcal{N}\left(a + \sigma^2, \sigma^2\right)$$

for $L = \mathcal{N}(a, \sigma^2)$.

REMARK 2.5.2. Assume that condition $\Lambda 6$ holds where the distribution L is a mixture of the normal distributions $\mathcal{N}\left(-\sigma^2/2, \sigma^2\right)$ with respect to the parameter σ with a probability distribution K on $(0, \infty)$. Then the distribution function $L(x)$ is continuous and increasing in $(-\infty, \infty)$. It is clear that the distribution $\widetilde{L}$ is a mixture of the normal distributions $\mathcal{N}\left(\sigma^2/2, \sigma^2\right)$ with respect to the parameter σ distributed according to the same distribution K. Therefore $\widetilde{L}(\infty) = 1$.

THEOREM 2.5.3. *We have*

$$\Lambda 6 \Rightarrow (H^t) \lhd (\widetilde{H}^t). \tag{2.5.10}$$

In particular, if condition $\Lambda 6$ *holds, then*

$$(\widetilde{H}^t) \lhd (H^t) \iff \widetilde{L}(\infty) = 1, \tag{2.5.11}$$

$$(\widetilde{H}^t) \overline{\lhd} (H^t) \iff \widetilde{L}(\infty) < 1. \tag{2.5.12}$$

PROOF. Condition $\Lambda 6$ implies that

$$\lim_{N \to -\infty} \lim_{t \to \infty} \mathsf{P}^t(\Lambda_t < N) = 0. \tag{2.5.13}$$

By Theorem 2.2.3, with the hypotheses H^t and $\widetilde{H}^t$ interchanged, we obtain (2.5.10) in view of the equality

$$\mathsf{P}^t(z_t < N) = \mathsf{P}^t\left(\widetilde{z}_t > N^{-1}\right)$$

and relation (2.5.13).

Now we assume that condition $\Lambda 6$ is satisfied. Then $\widetilde{L}(\infty) \leq 1$ by Theorem 2.2.3 and thus (2.5.3) holds. Hence Theorem 2.2.3 implies (2.5.11). Relation (2.5.12) follows from the inequality $\widetilde{L}(\infty) \leq 1$, equivalence (2.5.11), and the following property: either $(\widetilde{H}^t) \lhd (H^t)$ or $(\widetilde{H}^t) \overline{\lhd} (H^t)$. □

REMARK 2.5.3. The implication $\Leftarrow$ in (2.5.12) is known as the *first Le Cam theorem* (see [**22**]).

REMARK 2.5.4. The following property follows from Theorem 2.5.3 under condition $\Lambda 6$: either $(\widetilde{H}^t) \lhd\rhd (H^t)$ (type **a**) or $(\widetilde{H}^t) \overline{\lhd} (H^t)$ (type **b**). Moreover

$$(\widetilde{H}^t) \lhd\rhd (H^t) \iff \widetilde{L}(\infty) = 1,$$

$$(\widetilde{H}^t) \overline{\lhd} (H^t) \iff \widetilde{L}(\infty) < 1.$$

If $L = \mathcal{N}\left(a, \sigma^2\right)$ in condition $\Lambda 6$, then $a \leq -\sigma^2/2$ by Corollary 2.5.2. Moreover,

$$\widetilde{L}(\infty) = 1 \iff a = -\sigma^2/2,$$

$$\widetilde{L}(\infty) < 1 \iff a < -\sigma^2/2.$$

Behavior of the Neyman–Pearson tests. The following result establishes a relationship between the behavior of the level α_t and that of the type II error probability $\beta(\delta_t^{+,\alpha_t})$ for the Neyman–Pearson test under condition $\Lambda 6$.

THEOREM 2.5.4. *If condition* $\Lambda 6$ *holds, then for any* $\alpha \in (0,1)$:

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \limsup_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \le \widetilde{L}(\bar{l}_{1-\alpha} + 0), \tag{2.5.14}$$

$$\lim_{t\to\infty} \alpha_t = \alpha \Rightarrow \liminf_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \ge \widetilde{L}(\underline{l}_{1-\alpha}) \tag{2.5.15}$$

where $\underline{l}_p$ *and* $\bar{l}_p$ *are defined by equalities* (2.4.4).

PROOF. Assume that $\alpha_t \to \alpha \in (0,1)$ as $t \to \infty$. First suppose that u is a point of continuity of the function $L(x)$ satisfying $L(u) > 1 - \alpha$. As in the proof of Theorem 2.4.1, we obtain that $\limsup_{t\to\infty} d_t \le u$. Therefore

$$\limsup_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \le \limsup_{t\to\infty} \widetilde{\mathsf{P}}^t(\Lambda_t \le d_t) \le \lim_{t\to\infty} \widetilde{\mathsf{P}}^t(\Lambda_t \le u) = \widetilde{L}(u), \tag{2.5.16}$$

since, by (2.5.2), the point u is also a point of continuity of $\widetilde{L}(x)$.

Inequality (2.5.16) yields

$$\limsup_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \le \inf\left\{\widetilde{L}(u): L(u) > 1 - \alpha\right\} = \widetilde{L}(\bar{l}_{1-\alpha} + 0)$$

where the infimum is taken over all points u of continuity of $L(x)$ such that

$$L(u) > 1 - \alpha.$$

Therefore implication (2.5.14) is proved.

Now we suppose that u is a point of continuity of $L(x)$ such that $L(u) < 1 - \alpha$. Then $\liminf_{t\to\infty} d_t \ge u$, whence

$$\liminf_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \ge \liminf_{t\to\infty} \widetilde{\mathsf{P}}^t(\Lambda_t < d_t) \ge \lim_{t\to\infty} \widetilde{\mathsf{P}}^t(\Lambda_t < u) = \widetilde{L}(u).$$

Therefore

$$\liminf_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) \ge \sup\left\{\widetilde{L}(u) : L(u) < 1 - \alpha\right\} = \widetilde{L}(\underline{l}_{1-\alpha})$$

where the supremum is taken over all points u of continuity of $L(x)$ such that $L(u) < 1 - \alpha$. Implication (2.5.15) is also proved. □

If condition $\Lambda 6$ holds and the function $L(x)$ is continuous, then Theorem 2.5.4 can be sharpened. First we prove an auxiliary result which is also of interest on its own.

LEMMA 2.5.1. *Let* $(Z^t, \mathfrak{A}^t, \mathbf{S}^t)$, $t \in \mathbf{R}_+$, *be a family of probability measures and let* Y_t *be a measurable mapping of the space* $(Z^t, \mathfrak{A}^t)$ *into the space* $(\mathbf{R}, \mathscr{B})$ *such that*

$$\mathcal{L}\left(Y_t | \mathbf{S}^t\right) \xrightarrow{w} S, \qquad t \to \infty, \tag{2.5.17}$$

where S *is a probability distribution on* $\overline{\mathbf{R}}$ *whose distribution function* $S(x)$ *is continuous for* $x \in (-\infty, \infty)$ *and such that* $S(-\infty) = 0$ *and* $S(\infty) \le 1$. *Then*

$$\lim_{t\to\infty} \mathbf{S}^t(Y_t = y_t) = 0 \tag{2.5.18}$$

for any family (y_t) *of numbers such that* $y_t \in \mathbf{R}$ *and* $\limsup_{t\to\infty} y_t < \infty$ *if* $S(\infty) < 1$. *Further assume that the function* $S(x)$ *is strictly increasing in the interval* $(\underline{x}, \overline{x})$ *where*

$$\underline{x} = \sup\left\{x: S(x) = 0\right\}, \qquad \overline{x} = \inf\left\{x: S(x) = S(\infty)\right\}.$$

Let (y_t) *and* (ε_t) *be arbitrary families of numbers such that* $y_t \in \mathbf{R}$, $\varepsilon_t \in [0,1]$, *and the limit*

$$\lim_{t\to\infty}\left[\mathbf{S}^t(Y_t > y_t) + \varepsilon_t \mathbf{S}^t(Y_t = y_t)\right] = \beta \tag{2.5.19}$$

exists. Then the limit $\lim_{t\to\infty} y_t$ *exists for any* $\beta \in (1 - S(\infty), 1)$ *and*

$$\lim_{t\to\infty} y_t = s_{1-\beta} \tag{2.5.20}$$

where s_p *is a* p*-quantile of the distribution* S. *Moreover, if* $\beta = 1$, *then*

$$\limsup_{t\to\infty} y_t \le \underline{x}, \tag{2.5.21}$$

while if $\beta = 1 - S(\infty)$, *then*

$$\liminf_{t\to\infty} y_t \ge \overline{x}. \tag{2.5.22}$$

PROOF. First assume that $S(\infty) = 1$ and prove relations (2.5.18) and (2.5.20). Put $S_t(y) = \mathbf{S}^t(Y_t < y)$. Then

$$\begin{aligned}\mathbf{S}^t(Y_t = y_t) &\le S_t(y_t + \varepsilon) - S_t(y_t)\\ &= [S_t(y_t + \varepsilon) - S(y_t + \varepsilon)] - [S_t(y_t) - S(y_t)] + [S(y_t + \varepsilon) - S(y_t)]\end{aligned}$$

for any $\varepsilon > 0$. By the Pólya theorem (see [**16**]) and since the function $S(x)$ is uniformly continuous, the latter inequality and condition (2.5.17) imply (2.5.18).

It is clear that relation (2.5.18) implies that

$$(2.5.19) \iff \lim_{t\to\infty} S_t(y_t) = 1 - \beta.$$

Therefore, again by the Pólya theorem, we obtain

$$(2.5.19) \iff \lim_{t\to\infty} S(y_t) = 1 - \beta. \tag{2.5.23}$$

The properties of the function $S(x)$ and relation (2.5.23) imply that the limit $\lim_{t\to\infty} y_t$ exists and (2.5.20) holds.

Now we assume that $S(\infty) < 1$. It follows from condition (2.5.17) and the equality $S(-\infty) = 0$ that

$$\lim_{t\to\infty} S_t(x) = S(x) \tag{2.5.24}$$

uniformly in $x \le N$ for any $N < \infty$. Now the proof of the required relation (2.5.18) follows the lines of the proof of the same relation in the case where $S(\infty) = 1$. The only difference is that the reference to the Pólya theorem is replaced with the uniform convergence (2.5.24) in the interval $\left(-\infty, \limsup_{t\to\infty} y_t + \varepsilon\right)$ where $\varepsilon > 0$.

It is clear that $\beta \in [1 - S(\infty), 1]$. First assume that $\beta \in (1 - S(\infty), 1)$. In order to prove that the limit $\lim_{t\to\infty} y_t$ exists and relation (2.5.20) holds in this case, note that

$$\lim_{N\to-\infty} \lim_{t\to\infty} S_t(N) = 0$$

by (2.5.17) and in view of $S(-\infty) = 0$. Therefore conditions (2.5.17) and (2.5.19) yield $\limsup_{t\to\infty} y_t < \infty$. Let us prove this fact by contradiction. Assume that $\limsup_{t\to\infty} y_t = \infty$. Then there exists a sequence (t_n) such that $t_n \to \infty$ and $y_{t_n} \to \infty$ as $n \to \infty$. Let $\varepsilon > 0$ be arbitrary and let $N_\varepsilon \in \mathbf{R}$ be such that $S(N_\varepsilon) > S(\infty) - \varepsilon$. By (2.5.24), there exists an integer $n_0 = n_0(\varepsilon)$ such that

$|S_{t_n}(N_\varepsilon) - S(N_\varepsilon)| < \varepsilon$ for all $n > n_0$. Now suppose that $n_1 = n_1(\varepsilon)$ is an integer and $y_{t_n} > N_\varepsilon$ for all $n > n_1$. Therefore

$$S_{t_n}(y_{t_n}) \geq S_{t_n}(N_\varepsilon) \geq S(N_\varepsilon) - \varepsilon \geq S(\infty) - 2\varepsilon$$

for all $n > n_0 \vee n_1$. Hence

$$\liminf_{t\to\infty} \mathbf{S}^t(Y_t \geq y_t) \leq \liminf_{n\to\infty} \mathbf{S}^{t_n}(Y_{t_n} \geq y_{t_n}) \leq 1 - S(\infty),$$

since ε is arbitrary. On the other hand,

$$\liminf_{t\to\infty} \mathbf{S}^t(Y_t \geq y_t) \geq \lim_{t\to\infty} \left[\mathbf{S}^t(Y_t > y_t) + \varepsilon_t \mathbf{S}^t(Y_t = y_t)\right] = \beta > 1 - S(\infty)$$

by condition (2.5.19). This contradiction proves that $\limsup_{t\to\infty} y_t < \infty$. By (2.5.18), we obtain $\mathbf{S}^t(Y_t = y_t) \to 0$ as $t \to \infty$. Now the proof of equality (2.5.20) follows the lines of the same proof in the case $S(\infty) = 1$.

Now we prove (2.5.21) and (2.5.22) assuming that $S(\infty) \leq 1$. Both proofs are carried out by contradiction.

Suppose that $\beta = 1$ but $\limsup_{t\to\infty} y_t > \underline{x}$. Then for any $N \in (\underline{x}, \overline{y})$, where $\overline{y} = \limsup_{t\to\infty} y_t$, there exists $t_0 = t_0(N) \in \mathbf{R}_+$ such that $y_t > N$ for all $t > t_0$. Therefore the inequality

$$\mathbf{S}^t(Y_t < y_t) + (1 - \varepsilon_t)\mathbf{S}^t(Y_t = y_t) \geq \mathbf{S}^t(Y_t < y_t) \geq \mathbf{S}^t(Y_t < N)$$

holds for $t > t_0$. By inequalities (2.5.17) and (2.5.19) and by the equality $S(\underline{x}) = 0$ we obtain

$$\lim_{t\to\infty} \left[\mathbf{S}^t(Y_t < y_t) + (1 - \varepsilon_t)\mathbf{S}^t(Y_t = y_t)\right] > 0,$$

since $S(x)$ is strictly monotone in the interval $(\underline{x}, \overline{x})$. The latter inequality contradicts (2.5.19) for $\beta = 1$. Therefore inequality (2.5.21) is true.

Now assume that $\beta = 1 - S(\infty)$, but $\underline{y} = \liminf_{t\to\infty} y_t < \overline{x}$. Then there exists a sequence (t_n) such that $t_n \to \infty$ and $y_{t_n} \to \underline{y}$ as $n \to \infty$. By conditions (2.5.17) and (2.5.19) and by relation (2.5.18) we obtain

$$\lim_{n\to\infty} \left[\mathbf{S}^{t_n}(Y_{t_n} > y_{t_n}) + \varepsilon_{t_n} \mathbf{S}^{t_n}(Y_{t_n} = y_{t_n})\right] = 1 - S(\underline{y}) > 1 - S(\overline{x}) = 1 - S(\infty),$$

since $S(x)$ is strictly monotone in the interval $(\underline{x}, \overline{x})$. This contradicts (2.5.19) for $\beta = 1 - S(\infty)$. Therefore (2.5.22) is true. □

REMARK 2.5.5. If $\underline{x} = -\infty$ under the assumptions of Theorem 2.5.1, then the limit $\lim_{t\to\infty} y_t$ exists for $\beta = 1$ and equality (2.5.20) holds with $s_0 = \underline{x} = -\infty$. In a similar way, if $\overline{x} = \infty$, then the limit $\lim_{t\to\infty} y_t$ exists for $\beta = 1 - S(\infty)$ and equality (2.5.20) holds with $s_{1-\beta} = s_{S(\infty)} = \overline{x} = \infty$. In the case $S = \mathcal{N}(a, \sigma^2)$, see [45] for additional information if $a = -\sigma^2/2$, and [36] if $a < -\sigma^2/2$.

THEOREM 2.5.5. *Assume that condition* $\Lambda 6$ *holds and the function* $L(x)$ *is continuous and strictly increasing in the interval* $(\underline{l}, \overline{l})$ *where*

$$\underline{l} = \sup\{x: L(x) = 0\}, \qquad \overline{l} = \inf\{x: L(x) = 1\}.$$

Then for any $\alpha \in [0, 1]$

$$\lim_{t\to\infty} \alpha_t = \alpha \Longleftrightarrow \lim_{t\to\infty} \beta\left(\delta_t^{+,\alpha_t}\right) = \widetilde{L}(l_{1-\alpha}) \tag{2.5.25}$$

where l_p is a p-quantile of the distribution L and $\widetilde{L}(x)$ is the distribution function of the distribution $\widetilde{L}$ defined by equality (2.5.2).

PROOF. Condition $\Lambda 6$ and Theorem 2.5.1 imply weak convergence (2.5.1), where $\widetilde{L}$ is the probability distribution on $\overline{\mathbf{R}}$ whose distribution function $\widetilde{L}(x)$ is given by (2.5.2) and $\widetilde{L}(\infty) \leq 1$. Equality (2.5.2) shows that

$$\sup\{x: \widetilde{L}(x) = 0\} = \underline{l}, \qquad \inf\{x: \widetilde{L}(x) = \widetilde{L}(\infty)\} = \bar{l},$$

and that the function $\widetilde{L}(x)$ is continuous on $\mathbf{R}$ and strictly increasing on the interval $(\underline{l}, \bar{l})$.

Assume that $\alpha_t \to \alpha$ as $t \to \infty$. If $0 < \alpha < 1$, then the limit $\lim_{t\to\infty} d_t = l_{1-\alpha}$ exists by Theorem 2.5.1. Therefore $\limsup_{t\to\infty} d_t < \infty$. Then, again by Theorem 2.5.1, we have $\widetilde{\mathsf{P}}^t(\Lambda_t = d_t) \to 0$ as $t \to \infty$. By (2.5.1) and (2.5.2), by the inequality $\limsup_{t\to\infty} d_t < \infty$, and since $\widetilde{\mathsf{P}}^t(\Lambda_t < y)$ converges to $\widetilde{L}(y)$ uniformly in $y \leq N$ for any $N < \infty$, we obtain

$$\beta\left(\delta_t^{+,\alpha_t}\right) \to \widetilde{L}(l_{1-\alpha}), \qquad t \to \infty. \tag{2.5.26}$$

If $\alpha = 0$, then Lemma 2.5.1 implies that $\liminf_{t\to\infty} d_t \geq \bar{l}$. Therefore, for any $N \in (-\infty, \bar{l})$, there exists $t' = t'(N)$ such that $d_t > N$ for all $t > t'$. Hence

$$\beta\left(\delta_t^{+,\alpha_t}\right) \geq \widetilde{\mathsf{P}}^t(\Lambda_t < d_t) \geq \widetilde{\mathsf{P}}^t(\Lambda_t < N)$$

for all $t > t'$. Since $l_{1-\alpha} \geq \bar{l}$ for $\alpha = 0$ and N is arbitrary, convergence (2.5.1) implies convergence (2.5.26) for $\alpha = 0$.

If $\alpha = 1$, Lemma 2.5.1 implies $\limsup_{t\to\infty} d_t \leq \underline{l}$. Then for any $N \in (\bar{l}, \infty)$ there exists $t'' = t''(N)$ such that $d_t < N$ for any $t > t''$. Therefore

$$\beta\left(\delta_t^{+,\alpha_t}\right) \leq \widetilde{\mathsf{P}}^t(\Lambda_t < N) + \widetilde{\mathsf{P}}^t(\Lambda_t = d_t) \tag{2.5.27}$$

for all $t > t''$.

It is clear that $\limsup_{t\to\infty} d_t < \infty$ and $\widetilde{\mathsf{P}}^t(\Lambda_t = d_t) \to 0$ as $t \to \infty$ by Lemma 2.5.1. Therefore we obtain (2.5.26) for $\alpha = 1$ from estimate (2.5.27) and convergence (2.5.1). This completes the proof of the implication $\Rightarrow$ in (2.5.25).

Now assume that $\beta(\delta_t^{+,\alpha_t}) \to \widetilde{L}(l_{1-\alpha})$ as $t \to \infty$. Then $1 - \beta(\delta_t^{+,\alpha_t}) \to \beta$ where $\beta = 1 - \widetilde{L}(l_{1-\alpha}) \in [1 - \widetilde{L}(\infty), 1]$. For $\alpha \in (0,1)$, we have $\beta \in (1 - \widetilde{L}(\infty), 1)$ and, by Lemma 2.5.1, $d_t \to \tilde{l}_{1-\beta} = \tilde{l}_{\widetilde{L}(l_{1-\alpha})} = l_{1-\alpha}$ where $\tilde{l}_p$ is a p-quantile of the distribution $\widetilde{L}$. By condition $\Lambda 6$ and by Lemma 2.5.1, we obtain $\alpha_t \to 1 - L(l_{1-\alpha}) = \alpha$ as $t \to \infty$.

If $\alpha = 0$, then $\beta = 1 - \widetilde{L}(\infty)$. By Lemma 2.5.1, we have $\liminf_{t\to\infty} d_t \geq \bar{l}$. Then for any $N \in (-\infty, \bar{l})$ there exists $t_0 = t_0(N)$ such that $d_t > N$ for all $t > t_0$. Hence

$$\alpha_t \leq \mathsf{P}^t(\Lambda_t > N) + \varepsilon_t \mathsf{P}^t(\Lambda_t = d_t)$$

for all $t > t_0$. By condition $\Lambda 6$ and Lemma 2.5.1, we obtain $\alpha_t \to 0$ as $t \to \infty$, since N is arbitrary.

The proof of the relation $\alpha_t \to \alpha$ as $t \to \infty$ for $\alpha = 1$ is similar. Therefore the implication $\Leftarrow$ in (2.5.25) is also proved. □

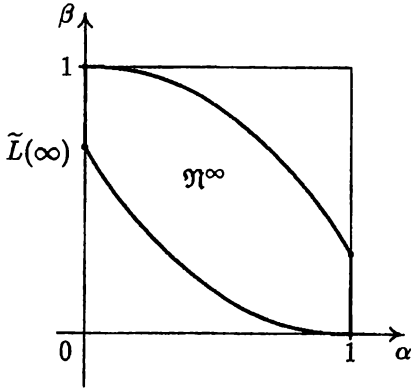

FIGURE 2.5.1

REMARK 2.5.6. Under the assumptions of Theorem 2.5.5, the function $\widetilde{L}(l_{1-\alpha})$ determines the equation for the lower bound of the limit $\mathfrak{N}^\infty$ of the set $\mathfrak{N}^t$ as $t \to \infty$. It is clear that this limit exists in our case. An example of the set $\mathfrak{N}^\infty$ is shown in Figure 2.5.1.

Behavior of Bayes tests and minimax tests. Theorem 2.5.2 combined with Lemma 2.5.1 enables us to obtain results on the asymptotic behavior of the probabilities of error for Bayes tests and minimax tests.

Let δ_t^π be the Bayes test with respect to the a priori distribution $(\pi, \widetilde{\pi}), \pi + \widetilde{\pi} = 1$, and the loss function $A_{ij} = 1 - \delta_{ij}$.

THEOREM 2.5.6. *Assume that condition* $\Lambda 6$ *holds and* $x = \ln(\pi/\widetilde{\pi})$ *is a point of continuity of the function* $L(x)$. *Then*

$$\lim_{t\to\infty} \alpha(\delta_t^\pi) = 1 - L\left(\ln \frac{\pi}{\widetilde{\pi}}\right), \tag{2.5.28}$$

$$\lim_{t\to\infty} \beta(\delta_t^\pi) = \widetilde{L}\left(\ln \frac{\pi}{\widetilde{\pi}}\right), \tag{2.5.29}$$

$$\lim_{t\to\infty} e_\pi(\delta_t^\pi) = \pi\left(1 - L\left(\ln \frac{\pi}{\widetilde{\pi}}\right)\right) + \widetilde{\pi}\widetilde{L}\left(\ln \frac{\pi}{\widetilde{\pi}}\right) \tag{2.5.30}$$

where $\widetilde{L}(x)$ *is defined by* (2.5.2).

PROOF. Since the Bayes test δ_t^π can be represented as $\delta_t^\pi = I(\Lambda_t \geq \ln(\pi/\widetilde{\pi}))$, condition $\Lambda 6$ implies (2.5.28), while relation (2.5.29) follows from Theorem 2.5.2 and definition (2.5.2) of the function $\widetilde{L}(x)$. Relation (2.5.30) follows from the equality

$$e_\pi(\delta_t^\pi) = \pi\alpha(\delta_t^\pi) + \widetilde{\pi}\beta(\delta_t^\pi)$$

in view of (2.5.28) and (2.5.29). □

Now assume that δ_t^* is the minimax test for testing the hypotheses H^t and $\widetilde{H}^t$.

THEOREM 2.5.7. *Let condition* $\Lambda 6$ *hold where the function* $L(x)$ *is continuous and strictly increasing in the interval* $(\underline{l}, \overline{l})$. *Then*

$$\lim_{t\to\infty} \alpha(\delta_t^*) = \lim_{t\to\infty} \beta(\delta_t^*) = \lim_{t\to\infty} e(\delta_t^*) = \alpha^* \tag{2.5.31}$$

where α^* is a unique solution of the equation $\widetilde{L}(l_{1-\alpha}) = \alpha$, $\widetilde{L}(x)$ is defined by (2.5.2), *and l_p is a p-quantile of the distribution L in condition* $\Lambda 6$.

PROOF. Observe that Theorem 1.2.4 yields

$$\alpha(\delta_t^*) = \beta(\delta_t^*) = e(\delta_t^*). \tag{2.5.32}$$

Now, by Remark 2.5.6 and by Theorem 1.2.4 we obtain (2.5.31) from (2.5.32). The existence and uniqueness of a solution of the equation $\widetilde{L}(l_{1-\alpha}) = \alpha$ follows, since $L(x)$ is continuous and strictly monotone in the interval $(\underline{l}, \bar{l})$. □

REMARK 2.5.7. It is clear that it is sufficient to find solution d^* of the equation $1 - L(d) = \widetilde{L}(d)$ in order to find solution α^* of the equation $\widetilde{L}(l_{1-\alpha}) = \alpha$. Then $\alpha^* = 1 - L(d^*) = \widetilde{L}(d^*)$, that is, $d^* = l_{1-\alpha^*} = \tilde{l}_{\alpha^*}$. Further, by Theorem 1.2.4, there exists a Bayes test $\delta_t^{\pi_t^*}$ with respect to the a prior distribution $(\pi_t^*, 1 - \pi_t^*)$ and a loss function A_{ij} such that $\alpha(\delta_t^{\pi_t^*}) = \beta(\delta_t^{\pi_t^*})$. This implies that $\alpha(\delta_t^{\pi_t^*}) \to \alpha^*$. Therefore, by Lemma 2.5.1, $\pi_t^* \to \pi^*$ as $t \to \infty$ where $\pi^* \in (0,1)$ is such that $\ln(\pi^*/(1-\pi^*)) = d^*$.

Independent observations. Let an observation be the vector

$$\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$$

where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables with the distribution P_θ having density $p(x;\theta)$ with respect to some σ-finite measure μ. Here, θ is an unknown parameter taking values in an open set $\Theta \subset \mathbf{R}$. Then the distribution P_θ^n of the vector $\xi^{(n)}$ has density with respect to the measure μ^n and this density is of the form $p_n(x;\theta) = \prod_{i=1}^n p(x_i;\theta)$, $x = (x_1, x_2, \ldots, x_n)$. For a fixed point $t \in \Theta$, introduce the following regularity conditions (R_t) on the family of probability distributions $\{\mathsf{P}_\theta, \theta \in \Theta\}$:

1) the function $p(x;\theta)$ is absolutely continuous with respect to θ in some neighborhood of the point $\theta = t$ for all $x \in \mathbf{R}$;
2) the derivative $p'_\theta(x;\theta) = \partial p(x;\theta)/\partial\theta$ exists for any θ belonging to a neighborhood of the point $\theta = t$ for μ-almost all $x \in \mathbf{R}$;
3) the function $I(\theta) = \mathsf{E}_\theta(\partial \ln p(\xi_1;\theta)/\partial\theta)^2$ is continuous and positive for $\theta = t$.

The function $I(\theta)$ is the *Fisher information* (see [**25, 38**]) and E_θ is expectation with respect to the measure P_θ.

Suppose that the hypothesis H^n is that the distribution of $\xi^{(n)}$ is determined by the measure P_t^n, while the hypothesis $\widetilde{H}^n$ is that the distribution of $\xi^{(n)}$ is determined by the measure $\mathsf{P}_{t+u/\sqrt{n}}^n$ where u is a fixed number such that

$$t + u/\sqrt{n} \in \Theta.$$

Then the logarithm of the likelihood ratio is given by

$$\Lambda_n = \sum_{i=1}^n \ln\left(p(\xi_i; t + u/\sqrt{n})/p(\xi_i;t)\right). \tag{2.5.33}$$

Consider the random variables

$$\eta_{in} = \left(p(\xi_i; t + u/\sqrt{n})/p(\xi_i;t)\right)^{1/2} - 1, \qquad i = 1, 2, \ldots, n, \tag{2.5.34}$$

and events $A_n = \{\max_{1\le i\le n}|\eta_{in}| < \varepsilon\}$ where $\varepsilon > 0$ is a fixed (small) number. If the event A_n occurs, then we use (2.5.33) and (2.5.34) to expand the logarithm into the Taylor series and obtain

$$\Lambda_n = 2\sum_{i=1}^{n}\ln(1+\eta_{in}) = 2\sum_{i=1}^{n}\eta_{in} - \sum_{i=1}^{n}\eta_{in}^2 + \sum_{i=1}^{n}\alpha_{in}|\eta_{in}|^3 \tag{2.5.35}$$

where α_{in} are some numbers such that $|\alpha_{in}| < 1$.

Before studying the asymptotic behavior of Λ_n as $n \to \infty$, consider three auxiliary results.

LEMMA 2.5.2. *Let a nonnegative function $g(y)$ be absolutely continuous in the interval $[a,b]$ and, moreover,*

$$\int_a^b \frac{|g'(y)|}{\sqrt{g(y)}}\,dy < \infty.$$

Then the function $\sqrt{g(y)}$ is also absolutely continuous in the interval $[a,b]$.

PROOF. Assume that $g(y) > 0$ in an interval $(\alpha,\beta) \subset [a,b]$. Then it is clear that the function $\sqrt{g(y)}$ is absolutely continuous in the interval $[\alpha,\beta]$ and

$$\sqrt{g(\beta)} - \sqrt{g(\alpha)} = \int_\alpha^\beta \frac{g'(y)}{\sqrt{g(y)}}dy.$$

Given $c \in [a,b]$, consider the open set $\{y \in (a,c): g(y) > 0\}$. As is well known [41], this set can be represented as the union of an at most countable number of disjoint intervals (α_i,β_i) such that $g(\alpha_i) = g(\beta_i) = 0$ if $\alpha_i \ne a$ and $\beta_i \ne c$. Therefore

$$\int_a^c \frac{g'(y)}{\sqrt{g(y)}}\,dy = \sum_{i=1}^{\infty}\int_{\alpha_i}^{\beta_i}\frac{g'(y)}{\sqrt{g(y)}}\,dy = \sqrt{g(c)} - \sqrt{g(a)},$$

that is, the function $\sqrt{g(y)}$ is also absolutely continuous. □

Consider the random variables

$$\zeta(u) = \sqrt{\frac{p(\xi_1,t+u)}{p(\xi_1,t)}} - 1, \qquad \varphi(\xi_1) = \frac{1}{2}\frac{\partial \ln(\xi_1;t)}{\partial t}$$

where u is a number such that $t + u \in \Theta$.

LEMMA 2.5.3. *If regularity conditions (R_t) hold, then*

$$\mathsf{E}_t\left(\frac{\zeta(u)}{u} - \varphi(\xi_1)\right)^2 \to 0 \tag{2.5.36}$$

as $u \to 0$.

PROOF. Since the function $I(\theta)$ is continuous in a neighborhood of the point $\theta = t$, we have by the Fubini theorem

$$\int_{-\infty}^{\infty} \left(\int_{t-\varepsilon}^{t+\varepsilon} \frac{(p'_\theta(x;\theta))^2}{p(x;\theta)}\, d\theta \right) d\mu = \int_{t-\varepsilon}^{t+\varepsilon} I(\theta)\, d\theta < \infty$$

where the internal integral on the left-hand side is μ-finite, whence it follows that the integral

$$\int_{t-\varepsilon}^{t+\varepsilon} \frac{|p'_\theta(x;\theta)|}{\sqrt{p(x;\theta)}}\, d\theta$$

also is μ-finite. Then we obtain in view of Lemma 2.5.2 that the function $\sqrt{p(x;\theta)}$ is absolutely continuous in a neighborhood of the point $\theta = t$ for μ-almost all x. Then we apply the Cauchy–Bunyakovskiĭ inequality to obtain

$$\begin{aligned} \mathsf{E}_t \left(\frac{\zeta(u)}{u} \right)^2 &\le \frac{1}{u^2} \int \left(\sqrt{p(x;t+u)} - \sqrt{p(x;t)} \right)^2 d\mu \\ &= \frac{1}{u^2} \int \left(\int_t^{t+u} \frac{p'_\theta(x;\theta)}{2\sqrt{p(x;\theta)}} d\theta \right)^2 d\mu \le \frac{1}{4u} \int_t^{t+u} I(\theta)\, d\theta. \end{aligned} \tag{2.5.37}$$

Since the function $I(\theta)$ is continuous at $\theta = t$, we deduce from (2.5.37) that

$$\limsup_{u \to 0} \mathsf{E}_t \left(\frac{\zeta |u|}{u} \right)^2 \le \frac{I(t)}{4}. \tag{2.5.38}$$

It is clear that

$$\mathsf{E}_t \varphi^2(\xi_1) = \frac{1}{4} I(t), \tag{2.5.39}$$

$$\frac{\zeta(u)}{u} \to \varphi(\xi_1), \qquad u \to 0, \qquad (\mathsf{P}_t\text{-a.s.}). \tag{2.5.40}$$

Therefore the required relation (2.5.36) follows from Theorem 1.A.4 in [**25**]. □

LEMMA 2.5.4. *If regularity conditions* (R_t) *hold, then*

$$\mathsf{E}_t \zeta^2(u) - \frac{1}{4} I(t) u^2 = o(u^2), \tag{2.5.41}$$

$$\mathsf{E}_t \left(\zeta^2(u) - \left(\frac{\partial \sqrt{p(\xi_1;t)}}{\partial t} \right)^2 u^2 \right) = o(u^2), \tag{2.5.42}$$

$$\mathsf{P}_t \{ |\zeta(u)| > \varepsilon \} = o(u^2), \tag{2.5.43}$$

$$\mathsf{E}_t \zeta(u) + \frac{1}{8} I(t) u^2 = o(u^2) \tag{2.5.44}$$

as $u \to 0$.

PROOF. By Lemma 2.5.3, relation (2.5.36) holds and moreover it can be rewritten as

$$\mathsf{E}_t (\zeta(u) - \varphi(\xi_1) u)^2 = o\left(u^2\right), \qquad u \to 0, \tag{2.5.45}$$

whence (2.5.42) follows. Further, equality (2.5.41) follows from (2.5.39) and (2.5.40) in view of Theorem 5.3 in [2] and inequality (2.5.38). Using (2.5.41), we obtain from (2.5.37)

$$\lim_{u\to 0}\frac{1}{u^2}\int\left(\sqrt{p(x;t+u)}-\sqrt{p(x;t)}\right)^2 d\mu=\frac{1}{4}I(t). \tag{2.5.46}$$

We split the domain of integration in (2.5.46) into the sets $\{x: p(x;t)=0\}$ and $\{x: p(x;t)\neq 0\}$ obtaining

$$\int_{\{x:p(x;t)=0\}} p(x;t+u)\,d\mu = o\left(u^2\right), \qquad u\to 0, \tag{2.5.47}$$

in view of (2.5.41). Further we use (2.5.47) and obtain

$$\begin{aligned}\mathsf{E}_t\zeta^2(u) &= \int_{\{x:p(x;t)\neq 0\}}\left(\sqrt{p(x;t+u)}-\sqrt{p(x;t)}\right)^2 d\mu\\ &= 2+o\left(u^2\right)-2\mathsf{E}_t\left(p(\xi_1;t+u)/p(\xi_1;t)\right)^{1/2} = -2\mathsf{E}_t\zeta(u)+o\left(u^2\right).\end{aligned}$$

This implies (2.5.44) in view of (2.5.41).

Now we prove (2.5.43). It is clear that

$$\begin{aligned}\mathsf{P}_t\{|\zeta(u)|>\varepsilon\} &\le \mathsf{P}_t\left\{\left|\zeta(u)-\frac{1}{2}\frac{\partial\ln p(\xi_1;t)}{\partial t}u\right|>\frac{\varepsilon}{2}\right\}+\mathsf{P}_t\left\{\left(\frac{\partial\ln p(\xi_1;t)}{\partial t}u\right)^2>\varepsilon^2\right\}\\ &\le \frac{4}{\varepsilon^2}\mathsf{E}_t\left(\zeta(u)-\frac{u}{2}\frac{\partial\ln p(\xi_1;t)}{\partial t}\right)^2\\ &\quad+\frac{u^2}{\varepsilon^2}\int I\left(x:\left|\frac{\partial\ln p(x;t)}{\partial t}\right|>\frac{\varepsilon}{|u|}\right)\left(\frac{\partial\ln p(x;t)}{\partial t}\right)^2 d\mu.\end{aligned}$$

The first term on the right-hand side is $o(u^2)$ by (2.5.45) and the second term is also $o(u^2)$, since $I(t)$ is finite. □

REMARK 2.5.8. It is clear that (2.5.45) implies

$$\mathsf{E}_t\zeta(u)-\frac{u}{2}\mathsf{E}_t\frac{\partial\ln p(\xi_1;t)}{\partial t}=o(u), \qquad u\to 0.$$

By (2.5.44), we obtain then

$$\mathsf{E}_t\frac{\partial}{\partial t}\ln p(\xi_1;t)=\int_{\{x:p(x;t)\neq 0\}}\frac{\partial p(x;t)}{\partial t}\,\mu(dx)=0. \tag{2.5.48}$$

The following result gives an asymptotic expansion of Λ_n as $n\to\infty$ under regularity conditions (R_t).

THEOREM 2.5.8. *If regularity conditions* (R_t) *hold, then*

$$\Lambda_n=\frac{u}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\ln p(\xi_i;t)}{\partial t}-\frac{u^2}{2}I(t)+\psi_n(u,t) \tag{2.5.49}$$

where

$$\mathsf{P}_t^n\{|\psi_n(u,t)|>\varepsilon\}\to 0, \qquad n\to\infty, \tag{2.5.50}$$

for any $\varepsilon > 0$ and all $u \in \mathbf{R}$. We also have

$$\mathcal{L}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln p(\xi_i;t)}{\partial t}\,\Big|\,\mathsf{P}_t^n\right) \xrightarrow{w} \mathcal{N}(0, I(t)), \qquad n \to \infty. \tag{2.5.51}$$

PROOF. The Taylor expansion (2.5.35) is valid if the event A_n occurs. In this case (2.5.35) implies representation (2.5.49)–(2.5.51) if for all $\varepsilon > 0$

$$\mathsf{P}_t^n\left(A_n^C\right) \to 0, \tag{2.5.52}$$

$$\mathsf{P}_t^n\left\{\left|2\sum_{i=1}^{n}\eta_{in} - \frac{u}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln p(\xi_i;t)}{\partial t} + \frac{u^2}{4}I(t)\right| > \varepsilon\right\} \to 0, \tag{2.5.53}$$

$$\mathsf{P}_t^n\left\{\left|\sum_{i=1}^{n}\eta_{in}^2 - \frac{u^2}{4}I(t)\right| > \varepsilon\right\} \to 0, \tag{2.5.54}$$

$$\mathsf{P}_t^n\left\{\sum_{i=1}^{n}|\eta_{in}|^3 > \varepsilon\right\} \to 0 \tag{2.5.55}$$

as $n \to \infty$.

First we prove (2.5.52). Since the random variables $\eta_{1n}, \ldots, \eta_{nn}$ are identically distributed and $\eta_{1n} = \zeta(u/\sqrt{n})$, relation (2.5.43) implies that

$$\mathsf{P}_t^n(A_n^C) \le \sum_{i=1}^{n}\mathsf{P}_t\{|\eta_{in}| > \varepsilon\} = n\mathsf{P}_t\left\{\left|\zeta\left(\frac{u}{\sqrt{n}}\right)\right| > \varepsilon\right\} = o(1)$$

as $n \to \infty$. Further,

$$\begin{aligned}\mathsf{P}_t^n\left\{\left|\sum_{i=1}^{n}\eta_{in}^2 - \frac{u^2}{n}\sum_{i=1}^{n}\left(\frac{\partial\sqrt{p(\xi_i;t)}}{\partial t}\right)^2\right| > \varepsilon\right\} &\le \frac{1}{\varepsilon}\mathsf{E}_t^n\sum_{i=1}^{n}\left|\eta_{in}^2 - \frac{u^2}{n}\left(\frac{\partial\sqrt{p(\xi_i;t)}}{\partial t}\right)^2\right| \\ &\le \frac{n}{\varepsilon}\mathsf{E}_t\left|\zeta^2\left(\frac{u}{\sqrt{n}}\right) - \frac{u^2}{n}\left(\frac{\partial\sqrt{p(\xi_1;t)}}{\partial t}\right)^2\right| \\ &\to 0\end{aligned}$$

by (2.5.42) where E_t^n is mathematical expectation with respect to the measure P_t^n. Moreover, since the sum

$$\sum_{i=1}^{n}\left(\partial\sqrt{p(\xi_i;t)}/\partial t\right)^2$$

converges in probability P_t^n to $I(t)/4$ by the law of large numbers, relation (2.5.54) is also proved.

Relation (2.5.55) follows from (2.5.52) and (2.5.54), since

$$\begin{aligned}\mathsf{P}_t^n\left\{\sum_{i=1}^{n}|\eta_{in}|^3 > \varepsilon\right\} \le{}& \mathsf{P}_t^n\left\{\max_{1\le i\le n}|\eta_{in}| > \frac{\varepsilon}{1+I(t)u^2}\right\} \\ &+ \mathsf{P}_t^n\left\{\sum_{i=1}^{n}\eta_{in}^2 > 1 + I(t)u^2\right\}.\end{aligned}$$

It remains to prove (2.5.53). Applying (2.5.44), we get

$$\mathsf{E}_t\eta_{ni} = \mathsf{E}_t\zeta\left(\frac{u}{\sqrt{n}}\right) = -\frac{I(t)u^2}{8n} + o\left(\frac{1}{n}\right)$$

as $n \to \infty$. Therefore (2.5.53) is equivalent to

$$J_n = \mathsf{P}_t^n\left\{2\left|\sum_{i=1}^n\left(\eta_{in} - \mathsf{E}_t\eta_{in} - \frac{u}{2\sqrt{n}}\frac{\partial \ln p(\xi_i;t)}{\partial t}\right)\right| > \varepsilon\right\} \to 0.$$

Since $\xi_1, \ldots, \xi_n$ are independent, we obtain from (2.5.48) that

$$\begin{aligned} J_n &\le \frac{4}{\varepsilon^2}\mathsf{E}_t^n\left[\sum_{i=1}^n\left(\eta_{in} - \mathsf{E}_t\eta_{in} - \frac{u}{2\sqrt{n}}\frac{\partial \ln p(\xi_i;t)}{\partial t}\right)\right]^2 \\ &= \frac{4n}{\varepsilon^2}\mathsf{E}_t\left(\zeta\left(\frac{u}{\sqrt{n}}\right) - \mathsf{E}_t\zeta\left(\frac{u}{\sqrt{n}}\right) - \frac{u}{2\sqrt{n}}\frac{\partial \ln p(\xi_1;t)}{\partial t}\right)^2 \\ &= \frac{4n}{\varepsilon^2}\left[\mathsf{E}_t\left(\zeta\left(\frac{u}{\sqrt{n}}\right) - \frac{u}{2\sqrt{n}}\frac{\partial \ln p(\xi_1;t)}{\partial t}\right)^2 - \left(\mathsf{E}_t\zeta\left(\frac{u}{\sqrt{n}}\right)\right)^2\right]. \end{aligned}$$

Since the right-hand side of the latter inequality tends to zero as $n \to \infty$ by (2.5.36) and (2.5.44), relation (2.5.53) as well as representation (2.5.49)–(2.5.50) is proved.

Relation (2.5.51) follows from equality (2.5.48), since the Fisher information $I(t)$ is finite in view of the central limit theorem for sums of independent identically distributed random variables. □

COROLLARY 2.5.3. *If regularity conditions* (R_t) *hold, then*

$$\mathcal{L}\left(\Lambda_n|\mathsf{P}_t^n\right) \xrightarrow{w} \mathcal{N}\left(-\frac{1}{2}I(t)u^2, I(t)u^2\right), \qquad n \to \infty, \tag{2.5.56}$$

that is, condition $\Lambda 6$ *is satisfied with* $L = \mathcal{N}\left(-\frac{1}{2}I(t)u^2, I(t)u^2\right)$.

REMARK 2.5.9. Asymptotic representations like (2.5.49)–(2.5.51) are known as the *local asymptotic normality* (LAN) property of a family of probability measures $\{\mathsf{P}_\theta^n, \theta \in \Theta\}$ at the point $\theta = t$ as $n \to \infty$. Representation (2.5.49)–(2.5.51) for the observation $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ can be found in the monographs [**25, 37**] for the case of independent but not necessarily identically distributed random variables $\xi_1, \xi_2, \ldots, \xi_n$.

REMARK 2.5.10. Relation (2.5.56) and Corollary 2.5.2 imply that convergence (2.5.1) holds where $\widetilde{L}(x)$ is the distribution function of the law $\mathcal{N}\left(\frac{1}{2}I(t)u^2, I(t)u^2\right)$. According to Remark 2.5.4, the mutual contiguity $(H^n) \lhd\rhd (\widetilde{H}^n)$ holds in this case.

CHAPTER 3

Goodness-of-Fit Tests

3.1. The setting of the problem. Kolmogorov test

Main definitions. Throughout this chapter we assume that the observation is a sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ where $\xi_1, \xi_2, \ldots, \xi_n$ are independent identically distributed random variables. Thus the random variables $\xi_1, \xi_2, \ldots, \xi_n$ are independent observations of a random variable ξ. Consider a hypothesis H about the distribution of the random variable ξ. We call H the *main* or *null hypothesis*. Our aim is to test the hypothesis H, that is, either accept H or reject it. The decision is to be made on the basis of the information contained in the sample $\xi^{(n)}$. The alternative hypothesis is that the null hypothesis H is false. The alternative hypothesis is denoted by K. The null hypothesis H can be either simple or composite. Therefore our aim is to decide whether results of the observation $\xi^{(n)}$ are in agreement with the hypothesis H. The tests described above are called *goodness-of-fit tests*.

We follow the general procedure to construct a goodness-of-fit test. Namely we introduce a statistic $T = T(\xi^{(n)})$ treated as the measure of disagreement between the data $\xi^{(n)}$ and the hypothesis H. We require that, if H is true, the distribution of this statistic is known exactly or at least approximately. In particular, if the hypothesis H is composite, then the distribution of the statistic $T(\xi^{(n)})$ should be the same for all simple hypotheses forming H. If we treat $T(x)$ as a mapping of $(\mathbf{R}^n, \mathscr{B}^n)$ into a measurable space $(Y, \mathscr{S})$, then the probability of the event $\{T(\xi^{(n)}) \in B\}$, $B \in \mathscr{S}$, is well defined if H is true. We write in this case $\mathsf{P}\{T(\xi^{(n)}) \in B/H\}$.

Consider a set $Y_\alpha \in \mathscr{S}$ of large deviations of the hypothesis H from the data $\xi^{(n)}$ such that $\mathsf{P}\{T(\xi^{(n)}) \in Y_\alpha/H\} \leq \alpha$ where $\alpha > 0$ is a sufficiently small number. Then the goodness-of-fit test can be described as follows. If $T(\xi^{(n)}) \in Y_\alpha$, then, under the assumption that the hypothesis H is true, an event occurs whose probability is small and thus the hypothesis H should be rejected, since it contradicts the observation $\xi^{(n)}$. Otherwise, that is, if $T(\xi^{(n)}) \notin Y_\alpha$, then there is no reason to reject the hypothesis H, since the observation does not contradict the hypothesis.

The goodness-of-fit test $\delta(x)$, $x \in \mathbf{R}^n$, for the hypothesis H is then such that $\delta(x) = 1$ for all $x \in \{x: T(x) \in Y_\alpha\}$ and $\delta(x) = 0$ for all $x \in \{x: T(x) \in Y \setminus Y_\alpha\}$, that is, $\delta(x)$ is a nonrandomized test for distinguishing two composite, generally speaking, hypotheses (see Section 1.3). The statistic T is called the *statistic of the test* δ, while the set Y_α (or the set $\{x: T(x) \in Y_\alpha\}$) is called the *critical set* for the hypothesis H. As usual, the number α is called the *significance level* or *type* I *error probability* for the test δ. Below we consider some examples of goodness-of-fit tests.

Kolmogorov goodness-of-fit test. Let the simple hypothesis H be that the distribution function of a random variable ξ equals $F(x)$. As the measure of

disagreement between the data $\xi^{(n)}$ and hypothesis H we consider the statistic

$$D_n = D_n(\xi^{(n)}) = \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \tag{3.1.1}$$

where $F_n(x)$ is the empirical distribution function constructed from the sample $\xi^{(n)}$, that is,

$$F_n(x) = \nu_n(x)/n, \qquad x \in \mathbf{R}, \tag{3.1.2}$$

$$\nu_n(x) = \sum_{i=1}^{n} I_{(-\infty,x)}(\xi_i) \tag{3.1.3}$$

where $I_A(x)$ is the indicator of the set $A \subset \mathbf{R}$. It follows from the Glivenko theorem that

$$\mathsf{P}\left\{\lim_{n\to\infty} D_n = 0/H\right\} = 1 \tag{3.1.4}$$

(see [**38**], Theorem 1.1.1).

The latter relation allows one to use the statistic D_n to construct a goodness-of-fit test for the hypothesis H by treating small values of the statistic D_n in favor of the hypothesis H. Large values of D_n suggest to a statistician that the hypothesis H is false and it should be rejected. The following result, known as the Kolmogorov theorem, allows one to construct the test for the hypothesis H if $F(x)$ is a continuous function.

THEOREM 3.1.1 (Kolmogorov). *If the function $F(x)$ is continuous, then for all $z > 0$*

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{n}D_n < z/H\right\} = K(z) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2z^2}. \tag{3.1.5}$$

Note that $K(z)$ is called the *Kolmogorov distribution function* (obviously, we have $K(z) = 0$ for $z \le 0$).

The proof of Theorem 3.1.1 is quite complicated and we omit it. We only mention that the method of proof of the Kolmogorov theorem and many other limit results for functionals of empirical distribution functions is to show first that

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{n}D_n < z/H\right\} = \mathsf{P}\left\{\sup_{0\le t\le 1} |w^0(t)| < z\right\} \tag{3.1.6}$$

for all $z \in \mathbf{R}$ where $w^0(t)$, $0 \le t \le 1$, is the Brownian bridge (see [**48**], §8). Using then the exact distribution of $\sup_{0\le t\le 1} |w^0(t)|$ (see [**48**]) we obtain (3.1.5) from (3.1.6). Also see [**16, 24**] for the proof of Theorem 3.1.1.

Now we construct the goodness-of-fit test for the hypothesis H based on the statistic D_n defined by (3.1.1). Let $\alpha > 0$ be the significance level of the test and let $z(\alpha)$ be a solution of the equation $K(z) = 1 - \alpha$ with respect to z. Then (3.1.5) implies for large n that

$$\mathsf{P}\{\sqrt{n}D_n \ge z(\alpha)/H\} \cong 1 - K(z(\alpha)) = \alpha. \tag{3.1.7}$$

Put $\delta_n = I(\sqrt{n}D_n \ge z(\alpha))$. It follows from (3.1.7) that the level of the test δ_n for large n is approximately equal to α. The test δ_n is called the *Kolmogorov goodness-of-fit test.* This test rejects the hypothesis H if $\sqrt{n}D_n \ge z(\alpha)$, that is,

if $D_n \geq z(\alpha)/\sqrt{n}$. Otherwise $\sqrt{n}D_n < z(\alpha)$ and there is no reason to reject the hypothesis H and it is accepted.

Let K_G be the simple hypothesis that the distribution function of a random variable ξ is G. If $\sup_x |G(x) - F(x)| \neq 0$, that is, if the distribution function G differs from F, then the hypothesis H is false given the alternative hypothesis K_G is true. The behavior of the statistic D_n if the hypothesis K_G is true is described in the following result.

THEOREM 3.1.2. *Let the distribution function $F(x)$ be continuous, and let $G(x)$ be another distribution function such that $\sup_x |G(x) - F(x)| \neq 0$. Then for all $z > 0$*

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{n}D_n < z/K_G\right\} = 0. \tag{3.1.8}$$

PROOF. According to the Glivenko theorem, for all $\varepsilon > 0$ and $\delta > 0$ there exists $n_0 = n_0(\varepsilon, \delta)$ such that for all $n > n_0$

$$\mathsf{P}\left\{\sup_x |F_n(x) - G(x)| > \varepsilon/K_G\right\} < \delta.$$

Let $\varepsilon < \sup_x |G(x) - F(x)|$. Then

$$\begin{aligned}
\mathsf{P}\left\{\sqrt{n}D_n < z/K_G\right\} &= \mathsf{P}\left\{\sqrt{n}D_n < z, \sup_x |F_n(x) - G(x)| \leq \varepsilon/K_G\right\} \\
&\quad + \mathsf{P}\left\{\sqrt{n}D_n < z, \sup_x |F_n(x) - G(x)| > \varepsilon/K_G\right\} \\
&\leq \mathsf{P}\left\{\sqrt{n}(\sup_x |G(x) - F(x)| - \varepsilon) < z/K_G\right\} \\
&\quad + \mathsf{P}\left\{\sup_x |F_n(x) - G(x)| > \varepsilon/K_G\right\}.
\end{aligned} \tag{3.1.9}$$

Since $\{\sqrt{n}(\sup_x |G(x) - F(x)| - \varepsilon) < z\}$ is a null event for $n > n_1 = n_1(\varepsilon, z)$ where

$$n_1(\varepsilon, z) = z^2 \left(\sup_x |G(x) - F(x)| - \varepsilon\right)^2,$$

the first term on the right-hand side of inequality (3.1.9) is zero for $n > n_1$. Taking into account (3.1.9) we obtain for $n > n_0 \vee n_1$ that $\mathsf{P}\{\sqrt{n}D_n < z/K_G\} \leq \delta$. Thus relation (3.1.8) is proved. □

If we put $z = z(\alpha)$ in relation (3.1.8), then

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{n}D_n < z(\alpha)/K_G\right\} = 0 \tag{3.1.10}$$

which means that the probability to accept the hypothesis H for the Kolmogorov test of level α given the hypothesis K_G is true and if $\sup_x |G(x) - F(x)| \neq 0$ tends to zero as $n \to \infty$. The tests satisfying condition (3.1.10) are called *consistent*. Thus the Kolmogorov test of level α is consistent. The behavior of the probability $\mathsf{P}\{\sqrt{n}D_n < z(\alpha)/K_G\}$ is studied in [7], §3.12, for distribution functions $G(x)$ that are close to $F(x)$ in a certain sense.

3.2. The Pearson test

The hypothesis. Measure of disagreement between a sample and the hypothesis. Let ξ be a random variable assuming values in a measurable space $(X, \mathscr{B})$. Consider a partition of the set X into r, $r \geq 2$, domains:

$$X = \bigcup_{i=1}^{r} S_i \tag{3.2.1}$$

where $S_i \cap S_j = \varnothing$, $i \neq j$, and $S_i \in \mathscr{B}$ for all i. Consider the following hypothesis H about the distribution of ξ:

$$\mathsf{P}\{\xi \in S_i / H\} = p_i, \qquad i = 1, 2, \ldots, r,$$

where $p_1, p_2, \ldots, p_r$ are given positive numbers such that $p_1 + p_2 + \cdots + p_r = 1$. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample from the distribution of ξ and denote by ν_i the total number of members of the sample $\xi^{(n)}$ belonging to the domain S_i, that is,

$$\nu_i = \sum_{j=1}^{n} I_{S_i}(\xi_j), \qquad i = 1, 2, \ldots, r. \tag{3.2.2}$$

It is clear that $\nu_1 + \nu_2 + \cdots + \nu_r = n$. We consider the following measure of disagreement between the sample $\xi^{(n)}$ and hypothesis H:

$$\zeta_n = \sum_{i=1}^{r} c_i \left(\frac{\nu_i}{n} - p_i\right)^2 \tag{3.2.3}$$

where c_i, $i = 1, 2, \ldots, r$, are some positive constants. By the Borel strong law of large numbers

$$\mathsf{P}\left\{\lim_{n\to\infty} \frac{\nu_i}{n} = p_i / H\right\} = 1, \qquad i = 1, 2, \ldots, r.$$

Thus $\mathsf{P}\{\lim_{n\to\infty} \zeta_n = 0 / H\} = 1$. Moreover if $K_{\widetilde{p}}$ is a hypothesis of the form

$$\mathsf{P}\{\xi \in S_i / K_{\widetilde{p}}\} = \widetilde{p}_i, \qquad i = 1, 2, \ldots, r,$$

where $\widetilde{p} = (\widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_r) \neq (p_1, p_2, \ldots, p_r)$, then again by the Borel strong law of large numbers

$$\mathsf{P}\left\{\lim_{n\to\infty} \zeta_n = \sum_{i=1}^{r} c_i (\widetilde{p}_i - p_i)^2 > 0 / K_{\widetilde{p}}\right\} = 1.$$

Thus the random variable ζ_n defined by (3.2.3) can be used as a measure of disagreement between the data and hypothesis H.

Pearson studied the behavior of ζ_n in the case of $c_i = n/p_i$, $i = 1, 2, \ldots, r$. The random variable ζ_n can be rewritten in this case as

$$\zeta_n = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \sum_{i=1}^{r} \frac{\nu_i^2}{np_i} - n. \tag{3.2.4}$$

Consider the distribution of the random variable ζ_n defined by (3.2.4) under the condition that the hypothesis H is true. Then

$$\mathsf{E}\{\zeta_n / H\} = r - 1. \tag{3.2.5}$$

Indeed $\mathsf{E}\{\nu_i / H\} = np_i$ and $\mathsf{D}\{\nu_i / H\} = np_iq_i$ where $q_i = 1 - p_i$. This implies that

$$\mathsf{E}\{\nu_i^2 / H\} = np_iq_i + (np_i)^2 = np_i(1 + (n-1)p_i).$$

Thus

$$\mathsf{E}\{\zeta_n/H\} = \sum_{i=1}^{r}(1+(n-1)p_i) - n = r-1$$

and equality (3.2.5) is proved. Similarly we obtain that

$$\mathsf{D}\{\zeta_n/H\} = 2(r-1) + \frac{1}{n}\left(\sum_{i=1}^{r}\frac{1}{p_i} - r^2 - 2r + 2\right). \tag{3.2.6}$$

Relations (3.2.5) and (3.2.6) imply that the first two moments of the random variable ζ_n under the condition that the hypothesis H is true converge to the corresponding moments of the $\chi^2(r-1)$ distribution as $n \to \infty$. Here $\chi^2(r-1)$ stands for the chi-square distribution with $r-1$ degrees of freedom.

Pearson theorem. The following result, known as the *Pearson theorem*, provides the limit distribution of random variables ζ_n as $n \to \infty$ under the condition that the hypothesis H is true.

THEOREM 3.2.1. *For all* $x > 0$

$$\lim_{n\to\infty} \mathsf{P}\{\zeta_n < x/H\} = K_{r-1}(x) \tag{3.2.7}$$

where $K_{r-1}(x)$ *is the chi-square distribution function with* $r-1$ *degrees of freedom.*

PROOF. Consider the random variables

$$\mu_k^{(i)} = I_{S_i}(\xi_k), \qquad i = 1,2,\ldots,r,\ k = 1,2,\ldots,n.$$

Then

$$\nu_i = \mu_1^{(i)} + \mu_2^{(i)} + \cdots + \mu_n^{(i)}, \qquad i = 1,2,\ldots,r.$$

The random variables $\mu_1^{(i)}, \mu_2^{(i)}, \ldots, \mu_n^{(i)}$ are independent, identically distributed, and such that

$$\mathsf{P}\left\{\mu_k^{(i)} = 1/H\right\} = \mathsf{P}\{\xi_k \in S_i/H\} = p_i, \qquad i = 1,2,\ldots,r.$$

Put $\nu = (\nu_1, \nu_2, \ldots, \nu_r)'$. Then $\nu = \sum_{j=1}^{n}\mu_j$ where $\mu_j = \left(\mu_j^{(1)}, \mu_j^{(2)}, \ldots, \mu_j^{(r)}\right)'$, $j = 1,2,\ldots,n$, are independent identically distributed random variables such that

$$\mathsf{P}\{\mu_j = e_k/H\} = \mathsf{P}\left\{\mu_j^{(k)} = 1/H\right\} = p_k, \qquad k = 1,2,\ldots,r.$$

Here $e_k = (0,\ldots,0,1,0,\ldots,0)'$ is the r-dimensional vector whose k-th coordinate equals 1 and all other coordinates are zero. If $\varphi_{\mu_j}(t)$ denotes the characteristic function of the vector μ_j, then

$$\varphi_{\mu_j}(t) = \mathsf{E}\left\{e^{it'\mu_j}/H\right\} = \sum_{k=1}^{r} p_k e^{it'e_k} = \sum_{k=1}^{r} p_k e^{it_k}$$

where $t = (t_1,\ldots,t_r)'$. This determines the characteristic function of the vector ν:

$$\varphi_\nu(t) = \prod_{j=1}^{n}\varphi_{\mu_j}(t) = \left(\sum_{k=1}^{r} p_k e^{it_k}\right)^n. \tag{3.2.8}$$

Now we introduce the random variables

$$\eta_i = \frac{\nu_i - np_i}{\sqrt{np_i}}, \qquad i = 1,2,\ldots,r. \tag{3.2.9}$$

It is clear that

$$\sum_{i=1}^{r} \eta_i \sqrt{p_i} = 0. \tag{3.2.10}$$

It follows from (3.2.4) and (3.2.9) that

$$\zeta_n = \sum_{i=1}^{r} \eta_i^2. \tag{3.2.11}$$

According to (3.2.8), the characteristic function of the vector $\eta = (\eta_1, \eta_2, \dots, \eta_r)'$ can be rewritten as

$$\begin{aligned}
\varphi_\eta(t) &= \mathsf{E}\{e^{it'\eta}/H\} \\
&= \mathsf{E}\left\{\exp\left(i\left[\nu_1 \frac{t_1}{\sqrt{np_1}} + \dots + \nu_r \frac{t_r}{\sqrt{np_r}}\right]\right.\right. \\
&\qquad \left.\left. - i[t_1\sqrt{np_1} + \dots + t_r\sqrt{np_r}]\right)/H\right\} \\
&= \exp\left(-i\sum_{j=1}^{r} t_j \sqrt{np_j}\right) \varphi_\nu\left(\frac{t_1}{\sqrt{np_1}}, \dots, \frac{t_r}{\sqrt{np_r}}\right) \\
&= \exp\left(-i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j}\right)\left(\sum_{k=1}^{r} p_r \exp\left(i\frac{t_k}{\sqrt{np_k}}\right)\right)^n.
\end{aligned} \tag{3.2.12}$$

Taking the logarithm of both sides of equality (3.2.12) and then expanding the exponent $\exp(it_k/\sqrt{np_k})$ and logarithm $\ln(1+x)$ into the Taylor series we obtain

$$\begin{aligned}
\ln \varphi_\eta(t) &= n \ln \sum_{k=1}^{r} p_k e^{it_k/\sqrt{np_k}} - i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j} \\
&= n \ln\left[1 + \frac{i}{\sqrt{n}}\sum_{k=1}^{r} t_k\sqrt{p_k} - \frac{1}{2n}\sum_{k=1}^{r} t_k^2 + O\left(n^{-3/2}\right)\right] \\
&\quad - i\sqrt{n}\sum_{j=1}^{r} t_j\sqrt{p_j} \\
&= -\frac{1}{2}\sum_{k=1}^{r} t_k^2 + \frac{1}{2}\left(\sum_{k=1}^{r} t_k\sqrt{p_k}\right)^2 + O\left(n^{-1/2}\right).
\end{aligned} \tag{3.2.13}$$

Passing to the limit in (3.2.13) as $n \to \infty$ we prove that

$$\lim_{n\to\infty} \varphi_\eta(t) = \exp\left(-\frac{1}{2}Q(t)\right) \tag{3.2.14}$$

uniformly with respect to t in any bounded domain where

$$Q(t) = \sum_{k=1}^{r} t_k^2 - \left(\sum_{k=1}^{r} t_k\sqrt{p_k}\right)^2, \qquad t = (t_1, t_2, \dots, t_r)'. \tag{3.2.15}$$

It is clear that the quadratic form $Q(t)$ defined by (3.2.15) can be represented as $Q(t) = t'\Lambda t$ where Λ is a matrix such that $\Lambda = I - pp'$. The symbol I stands for the $r \times r$ unit matrix, while $p = (\sqrt{p_1}, \dots, \sqrt{p_r})'$. Thus the right-hand side of (3.2.14)

is $\varphi_{\eta^0}(t) = \exp\{-\frac{1}{2}t'\Lambda t\}$ which is the characteristic function of the $\mathcal{N}(0,\Lambda)$ normal vector $\eta^0 = (\eta_1^0, \eta_2^0, \ldots, \eta_r^0)'$. Therefore convergence (3.2.14) implies that

$$\lim_{n\to\infty} \varphi_\eta(t) = \varphi_{\eta^0}(t) = \exp\left(-\frac{1}{2}t'\Lambda t\right)$$

uniformly with respect to t in any bounded domain. The continuity theorem for characteristic functions implies then that as $n \to \infty$

$$\mathcal{L}(\eta|H) \xrightarrow{w} \mathcal{L}(\eta^0|H) = \mathcal{N}(0,\Lambda),$$

whence

$$\mathcal{L}(\zeta_n|H) \xrightarrow{w} \mathcal{L}(\zeta^0|H), \qquad n \to \infty, \tag{3.2.16}$$

in view of (3.2.11) where

$$\zeta^0 = \sum_{i=1}^{r} (\eta_i^0)^2. \tag{3.2.17}$$

Equality (3.2.10) for the limit vector η^0 implies that

$$\sum_{i=1}^{r} \eta_i^0 \sqrt{p_i} = 0. \tag{3.2.18}$$

Let A be an orthogonal $r \times r$ matrix whose bottom row is $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_r})$. Then by (3.2.18)

$$\varkappa_r = \sum_{i=1}^{r} \eta_i^0 \sqrt{p_i} = 0 \tag{3.2.19}$$

where $\varkappa = (\varkappa_1, \ldots, \varkappa_r)' = A\eta^0$. On the other hand, it is known that orthogonal transformations do not change the canonical representation of quadratic forms. Thus the quadratic form (3.2.17) is given by

$$\zeta^0 = \sum_{i=1}^{r} (\eta_i^0)^2 = \sum_{i=1}^{r} \varkappa_i^2 = \sum_{i=1}^{r-1} \varkappa_i^2 \tag{3.2.20}$$

in view of (3.2.19).

The quadratic form $Q(t)$ can be rewritten for new coordinates $u = At$ as

$$Q(t) = \sum_{k=1}^{r} t_k^2 - \left(\sum_{k=1}^{r} t_k \sqrt{p_k}\right)^2 = \sum_{k=1}^{r} u_k^2 - u_r^2 = \sum_{k=1}^{r-1} u_k^2 = Q(A^{-1}u).$$

Therefore the characteristic function of the vector $\varkappa$ can be represented for new coordinates as

$$\varphi_\varkappa(u) = \mathsf{E}\left(e^{iu'\varkappa}/H\right) = \mathsf{E}\left(e^{iu'A\eta^0}/H\right) = \mathsf{E}\left(e^{i(A'u)'\eta^0}/H\right)$$
$$= \exp\left(-\frac{1}{2}Q(A^{-1}u)\right) = \exp\left(-\frac{1}{2}\sum_{k=1}^{r-1} u_k^2\right),$$

that is, the coordinates $\varkappa_1, \ldots, \varkappa_{r-1}$ of the vector $\varkappa = (\varkappa_1, \ldots, \varkappa_r)$ are independent $\mathcal{N}(0,1)$ identically distributed random variables and $\varkappa_r = 0$. Now it follows from (3.2.20) that $\mathcal{L}(\zeta^0|H) = \chi^2(r-1)$. Therefore (3.2.7) follows from (3.2.16). □

The Pearson goodness-of-fit test. Applying Theorem 3.2.1 one can construct a goodness-of-fit test for the hypothesis H in a way similar to that used to construct the Kolmogorov goodness-of-fit test based on the Kolmogorov limit theorem.

Let $\alpha > 0$ be a significance level and let $z_{r-1}(\alpha)$ be a solution of the equation $K_{r-1}(z) = 1 - \alpha$ with respect to z where $K_{r-1}(z)$ is the chi-square distribution function with $r - 1$ degrees of freedom. Then relation (3.2.7) implies for large n that

$$\mathsf{P}\{\zeta_n \geq z_{r-1}(\alpha)/H\} \approx 1 - K_{r-1}(z_{r-1}(\alpha)) = \alpha. \tag{3.2.21}$$

Now the test for the hypothesis H is $\delta_n = I(\zeta_n \geq z_{r-1}(\alpha))$. Relation (3.2.21) implies that the level of this test is approximately equal to α. The test is called the *Pearson goodness-of-fit test.* Sometimes it is called the *chi-square test.* The Pearson test rejects the hypothesis H if $\zeta_n \geq z_{r-1}(\alpha)$, and it accepts the hypothesis H if $\zeta_n < z_{r-1}(\alpha)$.

Let $K_{\widetilde{p}}$ be the simple hypothesis of the form $\mathsf{P}\{\xi \in S_i/K_{\widetilde{p}}\} = \widetilde{p}_i$, $i = 1, 2, \ldots, r$, where $\widetilde{p} = (\widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_r) \neq p = (p_1, p_2, \ldots, p_r)$. It turns out that the Pearson test of level $\alpha \in (0, 1)$ is consistent.

THEOREM 3.2.2. *For all vectors* $\widetilde{p} \neq p$

$$\lim_{n\to\infty} \mathsf{P}\{\zeta_n < z_{r-1}(\alpha)/K_{\widetilde{p}}\} = 0.$$

The proof of Theorem 3.2.2 can be found in [**37**], Theorem 3.2.

Examples of the Pearson goodness-of-fit tests for special models of observations can be found in [**7, 9, 14, 26, 34**].

REMARK 3.2.1. The hypothesis H tested with the help of the Pearson test is, generally speaking, composite. This hypothesis is simple only in the case where the vector $\mu = (\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(r)})$ assumes values e_k, $k = 1, 2, \ldots, r$, and the hypothesis H is that $\mathsf{P}\{\mu = e_k/H\} = p_k$, $k = 1, 2, \ldots, r$. One can show in this case that the Pearson test coincides with the likelihood ratio test (see [**7, 9**].

REMARK 3.2.2. Let H_0 be the hypothesis that the distribution function of ξ is $F(x)$, $x \in \mathbf{R} = (-\infty, \infty)$, and $\mathbf{R} = \bigcup_{i=1}^{r} X_i$ where $X_i \cap X_j = \varnothing$, $i \neq j$. Moreover let $\mathsf{P}\{\xi \in X_i/H_0\} = p_i$, $i = 1, 2, \ldots, r$, and $p_1 + p_2 + \cdots + p_r = 1$. As above let H be the hypothesis that $\mathsf{P}\{\xi \in X_i/H\} = p_i$, $i = 1, 2, \ldots, r$. The Pearson test constructed above for the hypothesis H is sometimes used to test the hypothesis H_0. However there are distributions $G(x)$ such that the hypothesis H is true but $\sup_x |G(x) - F(x)| > 0$, that is, the null hypothesis H_0 is false. The Pearson test does not detect the difference between such functions $G(x)$ and $F(x)$ and therefore is not consistent for testing the hypothesis H_0.

Quantile test. Sign test. Let a random variable ξ be real-valued, that is, $X = \mathbf{R} = (-\infty, \infty)$. Assume that the hypothesis H is that

$$F(y_i) = p_i, \qquad i = 1, 2, \ldots, r - 1,$$

where $F(x)$ is the distribution function of the random variable ξ,

$$0 < p_1 < \cdots < p_{r-1} < 1$$

are given numbers,

$$-\infty < y_1 < \cdots < y_{r-1} < \infty$$

are quantiles of levels $p_1, p_2, \ldots, p_{r-1}$, respectively, and $r \geq 2$. Thus the hypothesis H is composite and deals with all distributions with fixed quantiles and their levels.

Put $S_i = [y_{i-1}, y_i)$, $i = 1, 2, \ldots, r$, $y_0 = -\infty$, and $y_r = \infty$. Let

$$\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$$

be a sample of size n and let ν_i be the number of its members ξ_j belonging to the interval S_i. One can apply the Pearson test for the hypothesis H based on the statistic ζ_n defined by (3.2.4). The test in this case is called the *quantile test*. If $r = 2$ and $p_1 = 0,5$, then the corresponding test is called the *sign test*. The null hypothesis H in the latter case is that the median of the distribution of a random variable ξ is y_1. The statistic (3.2.4) in this case is $(4/n)(\nu_1 - n/2)^2$ where ν_1 is the total number of members of the sample $\xi^{(n)}$ belonging to the interval $(-\infty, y_1)$. In other words, ν_1 is the number of negative signs in the sequence $\xi_i - y_1$, $i = 1, 2, \ldots, n$.

The sign test is used under the following assumptions. Let $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ be a sample of size n whose members are independent observations of the vector (ξ, η). One needs to test the hypothesis H_0 that the coordinates ξ and η are independent and identically distributed, that is, $F(x, y) = F(x)F(y)$ where $F(x, y)$ is the distribution function of the vector (ξ, η) and $F(x)$ is the distribution function of the random variable ξ (and, of course, of η if the hypothesis H_0 is true). Let $\zeta_i = \xi_i - \eta_i$, $i = 1, 2, \ldots, n$. Then $\mathsf{P}\{\zeta_i < 0/H_0\} = \mathsf{P}\{\zeta_i > 0/H_0\} = 1/2$ and the null hypothesis H_0 is that the data $(\zeta_1, \zeta_2, \ldots, \zeta_n)$ is sampled from a distribution whose median is 0. The statistic ν_1 in this case is the total number of negative members in the sequence $\zeta_1, \zeta_2, \ldots, \zeta_n$. According to Theorem 3.2.1

$$\mathcal{L}\left(4(\nu_1 - n/2)^2/n | H_0\right) \xrightarrow{w} \chi^2(1), \qquad n \to \infty.$$

This relation allows one to construct the sign test for the null hypothesis H_0 that random variables ξ and η are independent and identically distributed for a given level α. The procedure is the same as in the case of the Pearson test. More details on the sign tests are given in [5].

The Pearson goodness-of-fit test for distributions with unknown parameters. Let a random variable ξ assume values in a measurable space $(X, \mathscr{B})$ and let its distribution depend on an unknown s-dimensional parameter

$$\theta = (\theta_1, \theta_2, \ldots, \theta_s)'$$

where $\theta_1, \theta_2, \ldots, \theta_s$ are real numbers. As above we introduce a partition (3.2.1) of the set X consisting of r domains.

Let the hypothesis H about the distribution of ξ be that

$$\mathsf{P}\{\xi \in S_i / H\} = p_i(\theta), \qquad i = 1, 2, \ldots, r, \tag{3.2.22}$$

where $p_1(\theta), p_2(\theta), \ldots, p_r(\theta)$ are known functions of the parameter θ such that $p_1(\theta) + p_2(\theta) + \cdots + p_r(\theta) = 1$. The parameter θ is unknown and our aim is to estimate it by the observation ξ. Let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample of size n and let $\nu_1, \nu_2, \ldots, \nu_r$ be the numbers of members of the sample $\xi_1, \xi_2, \ldots, \xi_n$ belonging to the domains $S_1, S_2, \ldots, S_r$, respectively (see relation (3.2.2)). The measure

of disagreement (3.2.4) between the sample and hypothesis (3.2.22) is in this case given by

$$\zeta_n = \zeta_n(\theta) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}. \tag{3.2.23}$$

Note that ζ_n depends on the unknown parameter θ and this does not allow one to use ζ_n to construct a test for the hypothesis H. Thus the first step is to exclude the unknown parameter from (3.2.23). To do so we substitute into (3.2.23) an estimator $\hat{\theta}_n = \hat{\theta}_n(\xi^{(n)})$ for θ and obtain the statistic

$$\widehat{\zeta}_n = \zeta_n(\hat{\theta}_n) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)}. \tag{3.2.24}$$

The statistic $\widehat{\zeta}_n$ defined by (3.2.24) depends on the sample $\xi^{(n)}$ and does not depend on θ and thus it can be used to test the hypothesis H. When constructing a test for the hypothesis H based on the statistic $\xi^{(n)}$ one needs to know its distribution or, at least, its limit distribution (the latter depends on the estimator $\hat{\theta}_n$). Below we consider the most famous method of estimation of the parameter θ which leads to a simple limit distribution of the statistic $\widehat{\zeta}_n$. This method was successfully used by R. Fisher, J. Neyman, and K. Pearson in the early twentieth century.

As an estimator of θ it is natural to take a value of the parameter for which $\zeta_n(\theta)$ defined by (3.2.23) attains its minimum. This is the so-called *minimum χ^2 method*. If the derivatives exist, then the problem of finding such a value is reduced to solving the following system of equations with respect to θ:

$$\frac{\partial \zeta_n(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^{r} \left(\frac{\nu_i - np_i(\theta)}{p_i(\theta)} + \frac{(\nu_i - np_i(\theta))^2}{2np_i^2(\theta)} \right) \frac{\partial p_i(\theta)}{\partial \theta_j} = 0, \tag{3.2.25}$$

$$j = 1, 2, \ldots, s.$$

Note however that this system is not easy to solve even in the simplest cases. On the other hand, one can show that the influence of the second term in parentheses is negligible for large n. Omitting this term, system (3.2.25) becomes of the form

$$\sum_{i=1}^{r} \frac{\nu_i - np_i(\theta)}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, s. \tag{3.2.26}$$

The method of estimation based on solving the system (3.2.26) is called the *modified minimum χ^2 method*. Under rather general assumptions both methods have the same limit distribution $\zeta_n(\hat{\theta}_n)$ as $n \to \infty$. Below we consider a simpler method based on solving the system (3.2.26).

Since $p_1(\theta) + p_2(\theta) + \cdots + p_r(\theta) = 1$ for all θ, system (3.2.26) becomes of the form

$$\sum_{i=1}^{r} \frac{\nu_i}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, s. \tag{3.2.27}$$

The system (3.2.27) can be rewritten as

$$\frac{\partial \ln L_n(\theta)}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, s, \tag{3.2.28}$$

where $L_n(\theta) = p_1^{\nu_1}(\theta)\dots p_r^{\nu_r}(\theta)$. The method of estimation based on solving the system (3.2.28) is nothing else but the maximum likelihood method for the polynomial distribution. Thus the estimator $\hat{\theta}_n$ obtained as a solution of system (3.2.27) (as well as that of (3.2.26)) is called the *maximum likelihood polynomial estimator.*

The limit distribution of the random variable $\widehat{\zeta}_n = \zeta_n(\hat{\theta}_n)$ as $n \to \infty$ is described in the following result where $\hat{\theta}_n$ is the maximum likelihood polynomial estimator.

THEOREM 3.2.3. *Let the functions* $p_j(\theta)$, $j = 1, 2, \dots, r$, $\theta = (\theta_1, \theta_2 \dots, \theta_s)'$, $s < r$, *be such that*:

1) $p_1(\theta) + p_2(\theta) + \dots + p_r(\theta) = 1$ *for all* θ;
2) $p_i(\theta) \geq c > 0$ *for all* $i = 1, 2, \dots, r$ *and the derivatives* $\frac{\partial p_i(\theta)}{\partial \theta_j}$ *and* $\frac{\partial^2 p_i(\theta)}{\partial \theta_j \partial \theta_k}$, $j, k = 1, 2, \dots, s$, $i = 1, 2, \dots, r$, *are continuous*;
3) *the rank of the* $r \times s$ *matrix* $\left\| \frac{\partial p_i(\theta)}{\partial \theta_j} \right\|$ *is equal to* s *for all* θ.

Then

$$\lim_{n\to\infty} \mathsf{P}\left\{\widehat{\zeta}_n < z/H\right\} = K_{r-s-1}(z) \quad \textit{for all } z > 0 \tag{3.2.29}$$

where $\widehat{\zeta}_n = \zeta_n(\hat{\theta}_n)$ *and* $\hat{\theta}_n$ *is the maximum likelihood polynomial estimator.*

The proof of Theorem 3.2.3 can be found in [**14**].

Based on Theorem 3.2.3 the goodness-of-fit test for the hypothesis (3.2.22) is constructed in the same way as in the case of the Pearson test for distributions whose parameters are known. The constructed test is also called the *Pearson test.*

REMARK 3.2.3. An estimator of θ can be evaluated without ranking the data. This can be done, for example, by maximizing the likelihood function

$$f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta)$$

with respect to θ where $f(x; \theta)$ is the density of the distribution of ξ with respect to some measure. The estimator of the parameter θ is not based in this case on frequencies $\nu_1, \nu_2, \dots, \nu_r$ for domains $S_1, S_2, \dots, S_r$ but uses the observations $\xi_1, \xi_2, \dots, \xi_n$ instead. Chernoff and Lehmann (see [**12**]) showed however that the limit relation (3.2.29) does not hold for this method of estimation.

EXAMPLE 3.2.1. Let a random variable ξ assume values $0, 1, 2, \dots$. Set

$$S_j = \{j - 1\}, \qquad j = 1, 2, \dots, r - 1,$$

and $S_r = \{r - 1, r, \dots\}$. Let the hypothesis H be defined by (3.2.22) where

$$p_j(\theta) = \frac{\theta^{j-1}}{(j-1)!} e^{-\theta}, \qquad j = 1, 2, \dots, r - 1,$$

$$p_r(\theta) = \sum_{i=r-1}^{\infty} \frac{\theta^i}{i!} e^{-\theta}, \qquad \theta > 0.$$

In this case $s = 1$ and thus the system (3.2.27) is reduced to the equation

$$\sum_{j=\nu}^{r-2} \left(\frac{j}{\theta} - 1\right) \nu_{j+1} + \nu_r \sum_{i=r-1}^{\infty} \left(\frac{i}{\theta} - 1\right) \frac{\theta^i}{i!} \left(\sum_{i=r-1}^{\infty} \frac{\theta^i}{i!}\right)^{-1} = 0,$$

whence

$$\theta = \frac{1}{n}\left[\sum_{j=\nu}^{r-2} j\nu_{j+1} + \nu_r \sum_{i=r-1}^{\infty} i\frac{\theta^i}{i!}\left(\sum_{i=r-1}^{\infty}\frac{\theta^i}{i!}\right)^{-1}\right]$$

where $\nu_1, \nu_2, \ldots, \nu_r$ are the numbers of members of the sample $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ belonging to $S_1, S_2, \ldots, S_r$, respectively. The first term in the square brackets is equal to the sum of all members of the sample $\xi_1, \xi_2, \ldots, \xi_n$ such that $\xi_i \leq r-2$, while the second term is approximately equal to the sum of all members of the sample $\xi_1, \xi_2, \ldots, \xi_n$ that are greater than or equal to $r-1$. Thus as an estimator $\hat{\theta}_n$ of the parameter θ one can take the sampling mean $\hat{\theta}_n = \overline{\xi}$. Note that the maximum likelihood estimator of the parameter θ is equal to $\overline{\xi}$ in the case of the Poisson distribution. Note also that the limit distribution as $n \to \infty$ of the statistic $\widehat{\zeta}_n = \zeta_n(\hat{\theta}_n)$ is the chi-square distribution with $r-2$ degrees of freedom (this result follows from Theorem 3.2.3).

EXAMPLE 3.2.2. Let

$$\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$$

be a sample of size n. Let $\nu_1, \nu_2, \ldots, \nu_r$ be the number of members of the sample $\xi^{(n)}$ belonging to $S_1, S_2, \ldots, S_r$, respectively (see relation (3.2.2)), where $S_1 = (-\infty, x_1 + \frac{1}{2}h)$, $S_i = (x_i - \frac{1}{2}h,\ x_i + \frac{1}{2}h)$ for $i = 2, 3, \ldots, r-1$, $S_r = (x_r - \frac{1}{2}h, \infty)$, $x_i = x_1 + (i-1)h$ for $i = 2, 3, \ldots, r$, and x_1 is some number of $(-\infty, \infty)$. Let the hypothesis H be defined by (3.2.22) where

$$p_i(\theta) = \int_{S_i} \varphi(x;\theta)\,dx, \qquad i = 1, 2, \ldots, r,$$

$\varphi(x;\theta)$ is the density of the normal $\mathcal{N}(m, \sigma^2)$ law, and $\theta = (m, \sigma)$. Then the system of equations (3.2.27) becomes of the form

$$m = \frac{1}{n}\sum_{i=1}^{r} \nu_i \int_{S_i} x\varphi(x;\theta)\,dx \left(\int_{S_i} \varphi(x;\theta)\,dx\right)^{-1},$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{r} \nu_i \int_{S_i} (x-m)^2\varphi(x;\theta)\,dx \left(\int_{S_i} \varphi(x;\theta)dx\right)^{-1}.$$

First we assume that x_1 and r are such that $\nu_1 = \nu_r = 0$. If h is small, then an approximate solution can be obtained by substituting the values of integrands at the middle points x_i of the corresponding intervals S_i instead of the integrands. Then we get the estimators $\widehat{m}_n$ and $\widehat{\sigma}_n$ defined by

$$\widehat{m}_n = \frac{1}{n}\sum_i \nu_i x_i, \qquad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_i \nu_i (x_i - \widehat{m}_n)^2.$$

To improve an approximation of the solution one can expand the integrands into the Taylor series in the neighborhoods of points x_i. Then the expressions for $\widehat{m}_n$ and $\widehat{\sigma}_n$ for small h become of the form

$$\widehat{m}_n = \frac{1}{n}\sum_i \nu_i x_i + O\left(h^4\right), \qquad \widehat{\sigma}_n^2 = \frac{1}{n}\sum_i \nu_i (x_i - \widehat{m}_n)^2 - \frac{h^2}{12} + O\left(h^4\right).$$

Thus omitting the terms of order h^4 we obtain estimators for m and σ^2. These estimators are the mean value and variance of the ranked sample $\overline{x}$ corrected with

the help of the *Sheppard correction* $\frac{h^2}{12}$ [**14**]. The *ranked sample* is obtained from the original sample if one substitutes the middle point of an underlying interval instead of any member of the sample belonging to this interval.

This procedure gives a better approximation even if h is not small but the end intervals are not empty and contain a small proportion of sampling values. Often it is convenient to merge the end intervals such that the union contains at least 10 sampling values. As estimators for m and σ^2 one can take $\overline{x}$ and s^2 evaluated by the original ranked sample with the Sheppard correction applied to s^2. If r' is the number of groups in the merged sample used for the evaluation of $\widehat{\zeta}_n$, then the limit distribution of $\widehat{\zeta}_n$ has $r'-3$ degrees of freedom, since two parameters are already estimated from the sample.

More details on the Sheppard correction are given in [**14**] (our Examples 3.2.1 and 3.2.2 are considered there for particular numerical values).

3.3. Smirnov test

The hypothesis on the homogeneity. Measure of disagreement between a sample and the hypothesis. It is an important applied problem to check whether the data is homogeneous. More precisely let $\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$ be a sample of size n consisting of n independent observations of a random variable ξ and let $\eta^{(m)} = (\eta_1, \eta_2, \ldots, \eta_m)$ be a sample of size m, independent of $\xi^{(n)}$ and consisting of m independent observations of a random variable η. The *hypothesis H on the homogeneity* is that the distributions of the random variables ξ and η coincide, that is, $\mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$ for all $x \in (-\infty, \infty)$. In other words, the hypothesis H means that the samples $\xi^{(n)}$ and $\eta^{(m)}$ are, in fact, the observations of the same random variable.

Let $S_n(x)$ be the empirical distribution function constructed from the sample $\xi^{(n)}$ and let $T_m(x)$ be the empirical distribution function constructed from the sample $\eta^{(m)}$. Consider the following measures of disagreement between the sample and hypothesis H:

$$D^+_{n,m} = \sup_x [S_n(x) - T_m(x)], \tag{3.3.1}$$

$$D_{n,m} = \sup_x |S_n(x) - T_m(x)|. \tag{3.3.2}$$

Since

$$S_n(x) - T_m(x) = (S_n(x) - F(x)) - (T_m(x) - F(x)),$$

where $F(x) = \mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$ is the common distribution function of ξ and η if the hypothesis H is true, the Glivenko theorem implies

$$\mathsf{P}\left\{\lim_{n,m\to\infty} D^+_{n,m} = 0/H\right\} = \mathsf{P}\left\{\lim_{n,m\to\infty} D_{n,m} = 0/H\right\} = 1.$$

On the other hand, if $H_{F,G}$ is the hypothesis that $\mathsf{P}\{\xi < x\} = F(x)$ and $\mathsf{P}\{\eta < x\} = G(x)$ where $F(x)$ and $G(x)$ are distribution functions such that $\sup_x |F(x) - G(x)| \neq$

0, then again by the Glivenko theorem

$$\mathsf{P}\left\{\lim_{n,m\to\infty} D^+_{n,m} = \sup_x [F(x) - G(x)]/H_{F,G}\right\} = 1,$$

$$\mathsf{P}\left\{\lim_{n,m\to\infty} D_{n,m} = \sup_x |F(x) - G(x)|/H_{F,G}\right\} = 1.$$

The above argument shows that the statistics defined by (3.3.1) and (3.3.2) can be used as measures of disagreement between the data the and hypothesis on the homogeneity.

Similarly to the Kolmogorov and Pearson goodness-of-fit tests we apply limit results on the behavior of measures of disagreement between the sample and the hypothesis. In the case of the hypothesis H we deal with $D^+_{n,m}$ and $D_{n,m}$ as $n, m \to \infty$. The method described below is due to Gnedenko. It allows one to find the distributions of statistics $D^+_{n,m}$ and $D_{n,m}$. We restrict the discussion to the case $m = n$. Other methods for studying the limit behavior of $D^+_{n,m}$ and $D_{n,m}$ can be found in [**24**].

The distributions of statistics $D^+_{n,m}$ and $D_{n,m}$. The following result contains the explicit expressions for the distribution functions of $D^+_{n,m}$ and $D_{n,m}$ if the hypothesis H is true.

THEOREM 3.3.1. *If the distribution function* $\mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$ *is continuous, then*

$$\mathsf{P}\left\{\sqrt{\frac{n}{2}} D^+_{n,m} < z/H\right\} = \begin{cases} 0, & z \le 0, \\ 1 - \binom{2n}{n-c}/\binom{2n}{n}, & 0 < z \le \sqrt{n/2}, \\ 1, & z > \sqrt{n/2}, \end{cases} \tag{3.3.3}$$

(3.3.4)

$$\mathsf{P}\left\{\sqrt{\frac{n}{2}} D_{n,m} < z/H\right\} = \begin{cases} 0, & z \le 1/\sqrt{2n}, \\ \sum_{|k|\le[n/c]} (-1)^k \binom{2n}{n-kc}/\binom{2n}{n}, & 1/\sqrt{2n} < z \le \sqrt{n/2}, \\ 1, & z > \sqrt{n/2}, \end{cases}$$

where $c = \left]z\sqrt{2n}\right[$ *is the minimal integer number which is greater than or equal to* $z\sqrt{2n}$.

PROOF. We assume below that the hypothesis H is true. Then the random variables $\xi_1, \dots, \xi_n$ and $\eta_1, \dots, \eta_n$ are independent and identically distributed and their common distribution function is $\mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$. We rearrange the random variables $\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_n$ in ascending order:

$$\zeta_1 < \zeta_2 < \cdots < \zeta_{2n}.$$

The equalities in this sequence may occur with probability 0, since the distribution function is continuous. Define the random variables $\chi_1, \chi_2, \dots, \chi_{2n}$ as follows: $\chi_k = 1$ if ζ_k is a member of the sample $\xi_1, \xi_2, \dots, \xi_n$, while $\chi_k = -1$ if ζ_k is a member of the sample $\eta_1, \eta_2, \dots, \eta_n$.

Put

$$S_0 = 0, \qquad S_k = \sum_{i=1}^{k} \chi_i, \quad k = 1, 2, \dots, 2n.$$

It is clear that

$$nD^+_{n,m} = \sup_{0\le k\le 2n} S_k, \qquad nD_{n,m} = \sup_{0\le k\le 2n} |S_k|. \tag{3.3.5}$$

Consider the points (k, S_k) on the plane (t, x) for $k = 0, 1, \dots, 2n$ and join them by segments of straight lines. As a result we get a polygonal line for which there are n subintervals $[k-1, k]$ of $[0, 2n]$ where the line goes up and n subintervals $[l-1, l]$ of $[0, 2n]$ where it goes down. This line starts at the point $(0, 0)$ and ends at the point $(2n, 0)$. Polygonal lines with these properties are called *trajectories*. Since the number of intervals of any trajectory where it rises is equal to n and the number of intervals where it descends is equal to n, the total number of trajectories is $\binom{2n}{n}$. All these trajectories are equiprobable. Indeed, any trajectory corresponds to the event $\{\xi_{i_1} < \xi_{i_2} < \dots < \xi_{i_{2n}}\}$ where $i_1, i_2, \dots, i_{2n}$ is a permutation of the numbers $1, 2, \dots, 2n$ and $\xi_{i_1} < \xi_{i_2} < \dots < \xi_{i_{2n}}$ are random variables $\xi_1, \xi_2, \dots, \xi_n, \xi_{n+1}, \dots, \xi_{2n}$ arranged in ascending order, $\xi_{n+k} = \eta_k$, $k = 1, 2, \dots, n$. Since the random variables $\xi_1, \xi_2, \dots, \xi_{2n}$ are identically distributed, we have

$$\mathsf{P}\{\xi_{i_1} < \xi_{i_2} < \dots < \xi_{i_{2n}}/H\} = \mathsf{P}\{\xi_{j_1} < \xi_{j_2} < \dots < \xi_{j_{2n}}/H\}$$

for all permutations $i_1, i_2, \dots, i_{2n}$ and $j_1, j_2, \dots, j_{2n}$ of the numbers $1, 2, \dots, 2n$. Thus the probability of any of the trajectories is $1/\binom{2n}{n}$.

According to (3.3.5) the random variables $nD^+_{n,n}$ and $nD_{n,n}$ assume integer values and thus

$$\mathsf{P}\left\{\sqrt{\frac{n}{2}} D^+_{n,n} < z/H\right\} = \mathsf{P}\left\{nD^+_{n,n} < c/H\right\},$$

$$\mathsf{P}\left\{\sqrt{\frac{n}{2}} D_{n,n} < z/H\right\} = \mathsf{P}\left\{nD_{n,n} < c/H\right\}$$

where $c =]z\sqrt{2n}[$. It remains to evaluate the probabilities

$$\mathsf{P}\left\{\sup_{0\le k\le 2n} S_k < c/H\right\} \quad \text{and} \quad \mathsf{P}\left\{\sup_{0\le k\le 2n} |S_k| < c/H\right\}$$

to complete the proof of (3.3.3) and (3.3.4). Since all the trajectories are equiprobable, we need to determine the total numbers of trajectories favorable to the events

$$\left\{\sup_{0\le k\le 2n} S_k < c\right\} \quad \text{and} \quad \left\{\sup_{0\le k\le 2n} |S_k| < c\right\}.$$

First we prove equality (3.3.3). We determine the total number of trajectories favorable to the event

$$\left\{\sup_{0\le k\le 2n} S_k < c\right\} = \left\{nD^+_{n,n} < c\right\},$$

that is, the total number of trajectories below the straight line $x = c$ (line α). We obtain this number by evaluating the total number of trajectories favorable to the converse event

$$\left\{\sup_{0\le k\le 2n} S_k \ge c\right\} = \left\{nD^+_{n,n} \ge c\right\},$$

that is, the total number of trajectories that have common points with the line α (we say in this case that a trajectory meets α). Every trajectory meeting the line α

(and called an old trajectory in this case) corresponds to another (new) trajectory defined as follows: the new trajectory coincides with the old one from the point $(0,0)$ until it meets the line α for the first time; after this point the new trajectory is the mirror reflection of the old one. Thus the new trajectory starts at the point $(0,0)$ and ends at the point $(2n,2c)$. The total number of different new trajectories (hence the trajectories meeting the line α) is equal to $\binom{2n}{n-c}$, since the number of intervals where a new trajectory rises is equal to $n+c$, while the number of intervals where it descents is equal to $n-c$. Thus the total number of trajectories that do not meet the line α is equal to $\binom{2n}{n} - \binom{2n}{n-c}$ and equality (3.3.3) is proved.

Now we turn to equality (3.3.4). We split the set $\mathfrak{M}$ of all trajectories into disjoint subsets $\mathfrak{A}_i$, $i \geq 0$, and $\mathfrak{B}_i$, $i \geq 1$: the set $\mathfrak{A}_0$ consists of trajectories that do not meet both lines $x = c$ (line α) and $x = -c$ (line β); the set $\mathfrak{A}_1$ consists of trajectories meeting the line α but not the line β; the set $\mathfrak{B}_1$ consists of trajectories meeting β but not α; the set $\mathfrak{A}_2$ consists of trajectories meeting first α, then β, and then not meeting α anymore; the set $\mathfrak{B}_2$ consists of trajectories meeting first β, then α, and then not meeting β anymore; the set $\mathfrak{A}_3$ consists of trajectories meeting first α, then β, then again α, and then not meeting β anymore, and so on. Obviously these sets are eventually empty. Moreover

$$\mathfrak{M} = \mathfrak{A}_0 \cup \left(\bigcup_{i \geq 1} (\mathfrak{A}_i \cup \mathfrak{B}_i) \right) \tag{3.3.6}$$

and the sets $\mathfrak{A}_1, \mathfrak{A}_2, \ldots$ and $\mathfrak{B}_1, \mathfrak{B}_2, \ldots$ are disjoint.

Along with the sets defined above we introduce the following sequence: the set A_1 consists of trajectories meeting α at least one time; the set B_1 consists of trajectories meeting β at least one time; the set A_2 consists of trajectories meeting α at least one time and then meeting β; the set B_2 consists of trajectories meeting β at least one time and then meeting α; the set A_3 consists of trajectories meeting α at least two times and β at least one time each in the following order: first α, then β, then α, and so on. It is clear that

$$A_1 = \mathfrak{A}_1 \cup \left(\bigcup_{i \geq 2} (\mathfrak{A}_i \cup \mathfrak{B}_i) \right), \qquad B_1 = \mathfrak{B}_1 \cup \left(\bigcup_{i \geq 2} (\mathfrak{A}_i \cup \mathfrak{B}_i) \right),$$

$$A_2 = \mathfrak{A}_2 \cup \left(\bigcup_{i \geq 3} (\mathfrak{A}_i \cup \mathfrak{B}_i) \right), \qquad B_2 = \mathfrak{B}_2 \cup \left(\bigcup_{i \geq 3} (\mathfrak{A}_i \cup \mathfrak{B}_i) \right),$$

and so on. This implies for all $i \geq 1$ that

$$(A_{2i-1} \setminus A_{2i}) \cup (B_{2i-1} \setminus B_{2i}) = \mathfrak{A}_{2i-1} \cup \mathfrak{A}_{2i} \cup \mathfrak{B}_{2i-1} \cup \mathfrak{B}_{2i}.$$

The latter equality together with (3.3.6) implies

$$\mathfrak{A}_0 = \mathfrak{M} \setminus \left(\bigcup_{i \geq 1} \left[(A_{2i-1} \setminus A_{2i}) \cup (B_{2i-1} \setminus B_{2i}) \right] \right). \tag{3.3.7}$$

To complete the proof we determine the total number of trajectories in the sets A_{2i-1}, A_{2i}, B_{2i-1}, and B_{2i} for $i = 1, 2, \ldots$. We demonstrate the method for the case of sets A_1 and A_2. Every trajectory starting from the point $(0,0)$

and meeting the line α corresponds to a new trajectory starting from the point $(0,0)$ and coinciding with the original trajectory until the point where it meets the line α; then the new trajectory is the mirror reflection of the old trajectory about the line α. The new trajectory ends at the point $(2n, 2c)$. The number of such trajectories is already determined above and it is equal to $\binom{2n}{n-c}$. Note that the cardinality of the set A_1 also is $\binom{2n}{n-c}$. If the original trajectory meets the line α first and then meets the line β, then the new trajectory meets the line $x = 3c$ (this is the mirror reflection of the line β).

To determine the cardinality of the set A_2 we introduce new trajectories as follows: the new trajectory coincides with the original one from the point $(0,0)$ until it meets α, then it coincides with the first mirror reflection about the line α until meeting the line $x = 3c$, and, finally, the last part of the new trajectory is the second reflection about the line $x = 3c$ of the trajectory reflected first. Such new trajectories end at the point $(2n, 4c)$. The total number of such trajectories (thus, the cardinality of the set A_2) is equal to $\binom{2n}{n-2c}$.

A similar reasoning (using an appropriate number of reflections) proves that the cardinality of the set A_i is $\binom{2n}{n-ic}$. In the same way we find that the number of trajectories in the set B_i also is $\binom{2n}{n-ic}$. Since the sets $A_{2i-1} \setminus A_{2i}$ and $B_{2i-1} \setminus B_{2i}$ are disjoint and the terms in (3.3.7) also are disjoint, we obtain from (3.3.7) that the total number of trajectories in the set $\mathfrak{A}_0$ is

$$\binom{2n}{n} - 2\sum_{i\geq 1}\left(\binom{2n}{n-(2i-1)c} - \binom{2n}{n-2ic}\right).$$

Thus equality (3.3.4) is proved. □

REMARK 3.3.1. The process S_k, $k = 0, 1, \ldots, 2n$, used in the proof of Theorem 3.3.1 is called a *random walk* on the axis. The method of evaluation of the number of trajectories in the sets A_i and B_i applied in the proof of Theorem 3.3.1 is well known in the theory of random walks as the *reflection method* [**24, 47**].

The Smirnov limit theorem. The following result describes the asymptotic behavior of the statistics $D^+_{n,n}$ and $D_{n,n}$ as $n \to \infty$.

THEOREM 3.3.2 (Smirnov). *If the distribution function*

$$\mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$$

is continuous, then

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{\frac{n}{2}} D^+_{n,n} < z/H\right\} = \begin{cases} 0, & z \leq 0, \\ 1 - e^{-2z^2}, & z > 0, \end{cases} \tag{3.3.8}$$

$$\lim_{n\to\infty} \mathsf{P}\left\{\sqrt{\frac{n}{2}} D_{n,n} < z/H\right\} = K(z) \tag{3.3.9}$$

where $K(z)$ is the Kolmogorov distribution function defined by (3.1.5).

PROOF. Let $z > 0$ be a fixed number. Consider the ratio

$$I_k = \frac{\binom{2n}{n-kc}}{\binom{2n}{n}} = \frac{(n!)^2}{(n-kc)!\,(n+kc)!}$$

where k is a fixed constant independent of n. Using the Stirling formula

$$m! = \sqrt{2\pi m}m^m e^{-m}(1+o(1)), \qquad m \to \infty,$$

we obtain

$$I_k = \left(1-\frac{kc}{n}\right)^{-n+kc}\left(1+\frac{kc}{n}\right)^{-n-kc}(1+o(1)). \tag{3.3.10}$$

Since $c = \left]z\sqrt{2n}\right[= z\sqrt{2n}(1+o(1))$ as $n \to \infty$, we have $kc/n = az\sqrt{2/n}(1+o(1))$. Then taking the logarithm of both sides of (3.3.10) and expanding the result into the Taylor series we get that

$$\ln I_k = -\frac{k^2c^2}{n} + o(1) = -2k^2z^2 + o(1), \qquad n \to \infty.$$

Thus

$$I_k = e^{-2k^2z^2}(1+o(1)). \tag{3.3.11}$$

This equality for $k = 1$ proves relation (3.3.8).

Now we turn to relation (3.3.9). For given $z > 0$ and $\varepsilon > 0$ there is a number $N(\varepsilon, z) > 0$ such that

$$e^{-2N^2z^2} < \frac{\varepsilon}{16}, \qquad \left|\sum_{|k|>N}(-1)^k e^{-2k^2z^2}\right| < \frac{\varepsilon}{4} \tag{3.3.12}$$

for $N > N(\varepsilon, z)$. Since

$$\binom{2n}{n-kc} > \binom{2n}{n-(k+1)c}, \qquad k > 0,$$

we have

$$\left|\sum_{N<|k|\le[n/c]}(-1)^k I_k\right| < 2I_N.$$

Taking into account (3.3.11) and the first inequality in (3.3.12) we obtain for sufficiently large n that

$$\left|\sum_{N<|k|\le[n/c]}(-1)^k I_k\right| \le 4e^{-2N^2z^2} < \frac{\varepsilon}{4}. \tag{3.3.13}$$

Now we apply the second inequality in (3.3.12) and inequality (3.3.13) to prove that for large n

$$\left|\sum_{N<|k|\le[n/c]}(-1)^k I_k - \sum_{|k|>N}(-1)^k e^{-2k^2z^2}\right| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}. \tag{3.3.14}$$

Further we deduce from (3.3.11) that for a fixed N and sufficiently large n

$$\left|\sum_{|k|\le N}(-1)^k I_k - \sum_{|k|\le N}(-1)^k e^{-2k^2z^2}\right| < \frac{\varepsilon}{2}. \tag{3.3.15}$$

Thus (3.3.14) and (3.3.15) imply for fixed $z > 0$ and $\varepsilon > 0$ and sufficiently large n that

$$\left| \mathsf{P}\left\{ \sqrt{\frac{n}{2}} D_{n,n}^{+} < z/H \right\} - K(z) \right| < \varepsilon.$$

Thus relation (3.3.9) is proved. □

The Smirnov homogeneity test. The hypothesis H on the homogeneity, that is, $\mathsf{P}\{\xi < x\} = \mathsf{P}\{\eta < x\}$, can be tested in the case of continuous distribution functions by using either the statistic $D_{n,m}^{+}$ or the statistic $D_{n,m}$. As above we consider the case $m = n$. To construct a test one can use either Theorem 3.3.1 or Theorem 3.3.2. For the sake of simplicity we construct a test by using Theorem 3.3.2.

We consider a goodness-of-fit test for the hypothesis H on the homogeneity based on the statistic $D_{n,n}$. Let $\alpha > 0$ be a significance level and let $z(\alpha)$ be a solution of the equation $K(z) = 1 - \alpha$ with respect to z. Then relation (3.3.9) implies that for sufficiently large n

$$\mathsf{P}\left\{ \sqrt{\frac{n}{2}} D_{n,n} \geq z(\alpha)/H \right\} \approx 1 - K(z(\alpha)) = \alpha. \tag{3.3.16}$$

The test is defined by

$$\delta_n = I\left\{ \sqrt{\frac{n}{2}} D_{n,n} \geq z(\alpha) \right\}.$$

It follows from relation (3.3.16) that the level of the test δ_n is approximately equal to α for large n. The test δ_n is called the *Smirnov test* or *Smirnov homogeneity test* or *Smirnov goodness-of-fit test.* The Smirnov test rejects the hypothesis H if $\sqrt{n/2}\, D_{n,n} \geq z(\alpha)$, that is, if $D_{n,n} \geq z(\alpha)\sqrt{2/n}$. Otherwise there is no reason to reject the hypothesis H and it is accepted.

It is not hard to show that the test δ_n is consistent. Indeed, let $K_{F,G}$ be the hypothesis that $\mathsf{P}\{\xi < x/K_{F,G}\} = F(x)$ and $\mathsf{P}\{\eta < x/K_{F,G}\} = G(x)$ where $\sup_x |F(x) - G(x)| \neq 0$. Similarly to the proof of Theorem 3.1.2 we obtain from the Glivenko theorem that

$$\lim_{n\to\infty} \mathsf{P}\left\{ \sqrt{\frac{n}{2}} D_{n,n} < z(\alpha)/K_{F,G} \right\} = 0,$$

that is, the test δ_n is consistent.

In the same way one can construct a goodness-of-fit test for the hypothesis H based on the statistic $D_{n,n}^{+}$. Namely, we introduce a solution $z^{+}(\alpha)$ of the equation $e^{-2z^2} = \alpha$ with respect to z where $\alpha > 0$ is a given significance level. It follows from relation (3.3.8) that

$$\mathsf{P}\left\{ \sqrt{\frac{n}{2}} D_{n,n}^{+} \geq z^{+}(\alpha)/H \right\} \approx e^{-2(z^{+}(\alpha))^2} = \alpha$$

for sufficiently large n. Thus the level of the test

$$\delta_n^{+} = I\left(\sqrt{\frac{n}{2}} D_{n,n}^{+} \geq z^{+}(\alpha) \right)$$

is approximately equal to α for sufficiently large n. The test δ_n^{+} can also be used for the hypothesis H on the homogeneity and it is called the *Smirnov test*, too.

3.4. Other goodness-of-fit tests

We studied Kolmogorov, Smirnov, and Pearson tests in the preceding sections. Many other goodness-of-fit tests are well known in mathematical statistics. Some of them are considered in this section.

Symmetric tests. Let a random variable ξ be real valued and let

$$\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$$

be a sample of size n. We split the real axis $\mathbf{R} = (-\infty, \infty)$ into r intervals

$$S_i = [y_{i-1}, y_i), \qquad i = 1, 2, \dots, r, \quad y_0 = -\infty, \ y_r = \infty. \tag{3.4.1}$$

Let the hypothesis H be such that $\mathsf{P}\{\xi \in S_i / H\} = p_i = 1/r$ for all $i = 1, 2, \dots, r$.

Let $\nu = (\nu_1, \nu_2, \dots, \nu_r)$ be the vector whose coordinates equal the number of members of the sample $\xi^{(n)}$ belonging to intervals (3.4.1), that is, ν_i, $i = 1, 2, \dots, r$, are evaluated by (3.2.2). Consider the class of statistics

$$\zeta_{n,r}(g) = \sum_{i=1}^{n} g(\nu_i) \tag{3.4.2}$$

where $g(x)$ is some real function defined for all nonnegative integer arguments $x = 0, 1, 2, \dots$. Since the random variables $\nu_1, \nu_2, \dots, \nu_n$ form a symmetric expression in (3.4.2), the statistic $\zeta_{n,r}(g)$ is called *symmetric* and tests based on statistic (3.4.2) are called *symmetric*.

If $g(x) = I_{\{k\}}(x)$ in (3.4.2) for some $k \in \{0, 1, 2, \dots, n\}$, then the statistic $\zeta_{n,r}(g)$ is equal to the number of intervals among $S_1, S_2, \dots, S_r$ that contain exactly k members of the sample $\xi_1, \xi_2, \dots, \xi_n$. We denote this statistic by

$$\mu_k = \mu_k(n, r) = \sum_{i=1}^{n} I_{\{k\}}(\nu_i). \tag{3.4.3}$$

Thus we have a collection of symmetric statistics $\mu_0, \mu_1, \dots, \mu_n$ defined by (3.4.3). It is obvious that these statistics are such that

$$\sum_{k=0}^{n} \mu_k(n, r) = r, \qquad \sum_{k=0}^{n} k\mu_k(n, r) = n. \tag{3.4.4}$$

Relation (3.4.2) can be rewritten in terms of the statistics $\mu_0, \mu_1, \dots, \mu_n$:

$$\zeta_{n,r}(g) = \sum_{k=0}^{n} g(k)\mu_k, \tag{3.4.5}$$

whence it follows that an arbitrary symmetric statistic is a linear combination of $\mu_0, \mu_1, \dots, \mu_n$. The converse is also true, namely any linear combination of $\mu_0, \mu_1, \dots, \mu_n$ is a symmetric statistic, since

$$\sum_{k=0}^{n} c_k \mu_k = \sum_{i=1}^{r} g(\nu_i)$$

where $g(k) = c_k$, $k = 0, 1, \dots, n$. Thus the *class of symmetric statistics coincides with the class of all linear combinations of* $\mu_0, \mu_1, \dots, \mu_n$.

Note that the statistic ζ_n defined by (3.2.4) and used in Section 3.2 to construct the Pearson test coincides with the symmetric statistic (3.4.2) for $g(x) = rx^2/n - x$ and if intervals (3.4.1) are equiprobable.

Empty boxes test. The symmetric test based on the statistic μ_0 is called the *empty boxes test.* Note that μ_0 is the number of intervals $S_1, S_2, \dots, S_r$ that do not contain any member of the sample $\xi_1, \xi_2, \dots, \xi_n$.

Below we evaluate the first two moments of the statistic μ_0 if the hypothesis H_p is true where H_p is such that $\mathsf{P}\{\xi \in S_i/H_p\} = p_i$, $i = 1, 2, \dots, r$. Here the numbers p_i are arbitrary and $p = (p_1, p_2, \dots, p_r)$. Consider the random variables $\eta_1, \eta_2, \dots, \eta_r$ such that $\eta_i = 1$ if the interval S_i does not contain any member of the sample $\xi_1, \dots, \xi_n$, and $\eta_i = 0$ otherwise. It is clear that $\mu_0 = \eta_1 + \cdots + \eta_n$, whence

$$\mathsf{E}\{\mu_0/H_p\} = \sum_{i=1}^{r} \mathsf{E}\{\eta_i/H_p\} = \sum_{i=1}^{r} \mathsf{P}\{\eta_i = 1/H_p\},$$

$$\begin{aligned}\mathsf{D}\{\mu_0/H_p\} &= \sum_{i=1}^{r} \mathsf{D}\{\eta_i/H_p\} + 2\sum_{i<j} \mathsf{E}\big\{(\eta_i - \mathsf{E}\{\eta_i/H_p\})(\eta_j - \mathsf{E}\{\eta_j/H_p\})/H_p\big\} \\ &= \sum_{i=1}^{r} \mathsf{P}\{\eta_i = 1/H_p\}[1 - \mathsf{P}\{\eta_i = 1/H_p\}] \\ &\quad + 2\sum_{i<j}\big[\mathsf{P}\{\eta_i = 1, \eta_j = 1/H_p\} - \mathsf{P}\{\eta_i = 1/H_p\}\mathsf{P}\{\eta_j = 1/H_p\}\big].\end{aligned}$$

Since the random variables $\xi_1, \xi_2, \dots, \xi_n$ are independent and identically distributed and

$$\{\eta_i = 1\} = \bigcap_{j=1}^{n}\{\xi_j \notin S_i\}, \qquad \{\eta_i = 1, \eta_j = 1\} = \bigcap_{k=1}^{n}\{\xi_k \notin S_i, \xi_k \notin S_j\},$$

we have

$$\mathsf{P}\{\eta_i = 1/H_p\} = (1 - p_i)^n, \qquad \mathsf{P}\{\eta_i = 1, \eta_j = 1/H_p\} = (1 - p_i - p_j)^n.$$

Thus

$$\mathsf{E}\{\mu_0/H_p\} = \sum_{i=1}^{r}(1 - p_i)^n, \tag{3.4.6}$$

$$\mathsf{D}\{\mu_0/H_p\} = 2\sum_{i<j}(1 - p_i - p_j)^n + \mathsf{E}\{\mu_0/H_p\} - (\mathsf{E}\{\mu_0/H_p\})^2. \tag{3.4.7}$$

It is easy to show that $\mathsf{E}\{\mu_0/H_p\}$ as a function of $p = (p_1, p_2, \dots, p_r)$ attains its minimum at $p = p^0 = (p_1^0, p_2^0, \dots, p_r^0)$ where $p_1^0 = p_2^0 = \cdots = p_r^0 = 1/r$. Equalities (3.4.6) and (3.4.7) for $p = p_0$ become of the form

$$\mathsf{E}\{\mu_0/H\} = r(1 - 1/r)^n,$$

$$\mathsf{D}\{\mu_0/H\} = r(r-1)(1 - 2/r)^n + r(1 - 1/r)^n - r^2(1 - 1/r)^{2n}$$

(note that $H = H_{p^0}$ in this case).

This shows that if the hypothesis H is not true, that is, if not all probabilities for intervals $S_1, S_2, \dots, S_r$ are equal to $1/r$, the statistic μ_0 tends to increase, since $\mathsf{E}\{\mu_0/H_p\} > \mathsf{E}\{\mu_0/H_{p^0}\}$ for $p \neq p_0$. Thus large values of μ_0 lead to the rejection

of the hypothesis H. The empty boxes test for testing the hypothesis H is then as follows: the hypothesis H is rejected if $\mu_0 \geq t_\alpha(n,r)$, while it is accepted otherwise. The number $t_\alpha(n,r)$ can be evaluated by using the distribution of the statistic μ_0 given the hypothesis H is true. However this distribution is complicated and the limit results for μ_0 are often used instead (see [**26, 27**]).

The empty boxes test has the same disadvantage as the Pearson test (see Remark 3.2.2). To avoid this disadvantage one should assume that r is large enough, that is, $r \to \infty$ as $n \to \infty$. This assumption allows one to apply the empty boxes test for testing the simple null hypothesis H_0 that $\mathsf{P}\{\xi < x/H_0\} = F(x)$ and $\mathsf{P}\{\xi \in S_i/H_0\} = p_i = 1/2$, $i = 1,2,\ldots,r$, where $F(x)$ is a given distribution function. More details are given in [**26, 27**].

General symmetric tests. The empty boxes test is based on the statistic μ_0 that does not contain all the information available from the sample $\xi^{(n)}$. It is clear that the statistic

$$\sum_{k=0}^{n} c_k \mu_k$$

contains more statistical information than μ_0 where $c_1 > 0, \ldots, c_n > 0$ are some weights. It is natural to choose the weights $c_1, c_2, \ldots, c_n$ such that the test is the most powerful in the class of all symmetric tests. The theory of symmetric tests is considered in [**27**]. Note also that the Pearson test has the maximal asymptotic power among all symmetric tests under appropriate conditions (see [**26, 27**]).

Tests of the homogeneity. We considered the Smirnov test in Section 3.3, however there exist many other tests of the homogeneity. We briefly discuss some of them below.

Chi-square test of the homogeneity. Consider s independent samples of independent observations. Assume that samples $1,2,\ldots,s$ contain $n_1, n_2, \ldots, n_s$ members, respectively. We assume that an attribute ξ_i is checked in the sample i and denote by $\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i}$ the observations of the attribute, so that we deal with independent random variables ξ_{ij}, $j = 1,2,\ldots,n_i$, $i = 1,2,\ldots,s$. Further let the results of every observation fall into r groups according to their values. Namely the domain X of the attribute ξ_i fall into parts $S_1, S_2, \ldots, S_r$ such that $\bigcup_{k=1}^{r} S_k = X$ and $S_k \cap S_l = \varnothing$, $k \neq l$. Let ν_{ij} be the number of members of the sample $\xi_{i1}, \xi_{i2}, \ldots, \xi_{in_i}$ belonging to S_j, $j = 1,2,\ldots,r$. Put $p_{ij} = \mathsf{P}\{\xi_i \in S_j\}$, $i = 1,2,\ldots,s$, $j = 1,2,\ldots,r$. Consider the hypothesis $H\colon (p_{i1},\ldots,p_{ir}) = (p_{11},\ldots,p_{1r})$ for all $i = 1,2,\ldots,s$. We form the vector $p = (p_1, p_2, \ldots, p_r)$ and note that $p_1 + \cdots + p_r = 1$.

The vector p is, generally speaking, unknown. However if the vector p is known, then $\mathsf{E}\{\nu_{ij}/H\} = n_i p_j$ and

$$\zeta_n(p) = \sum_{j=1}^{r} \sum_{i=1}^{s} \frac{(\nu_{ij} - n_i p_j)^2}{n_i p_j} \tag{3.4.8}$$

can be viewed as the measure of disagreement between the data and the null hypothesis (cf. relation (3.2.4)). Since $p_1, p_2, \ldots, p_r$ are unknown in general, we modify the χ^2 method and substitute the estimators

$$\widehat{p}_j = \frac{\nu_{\cdot j}}{n}, \qquad j = 1,2,\ldots,r, \tag{3.4.9}$$

for $p_1, \dots, p_r$ where

$$\nu_{\cdot j} = \sum_{i=1}^{s} \nu_{ij}, \qquad n = \sum_{i=1}^{s} n_i = \sum_{i=1}^{s} \sum_{j=1}^{r} \nu_{ij}. \tag{3.4.10}$$

Substituting (3.4.9) into (3.4.8) and taking into account (3.4.10) we get the statistic

$$\widehat{\zeta}_n = \zeta_n(\widehat{p}) = n \sum_{j=1}^{r} \sum_{i=1}^{s} \frac{(\nu_{ij} - n_i \nu_{\cdot j}/n)^2}{n_i \nu_{\cdot j}} = n\left(\sum_{j=1}^{r} \sum_{i=1}^{s} \frac{\nu_{ij}^2}{n_i \nu_{\cdot j}} - 1 \right) \tag{3.4.11}$$

where $\widehat{p} = (\widehat{p}_1, \dots, \widehat{p}_r)$. Note that

$$\mathcal{L}\left(\widehat{\zeta}_n | H\right) \xrightarrow{w} \chi^2((r-1)(s-1)), \qquad n \to \infty \tag{3.4.12}$$

(see [**14**], §30.6).

Based on the statistic $\widehat{\zeta}_n$ and using the limit relation (3.4.12) we construct the goodness-of-fit test for the hypothesis H. Some examples of applications of this test can be found in [**14, 26**].

Empty blocks test. Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ and $\eta^{(m)} = (\eta_1, \dots, \eta_m)$ be two independent samples of independent observations of random variables ξ and η, respectively. Assume that both ξ and η have continuous distribution functions. The hypothesis H is that $\mathsf{P}\{\xi < x/H\} = \mathsf{P}\{\eta < x/H\}$ for all $x \in (-\infty, \infty)$. Consider the order statistics $\zeta_{n,1} \le \zeta_{n,2} \le \dots \le \zeta_{n,n}$ related to the sample $\xi^{(n)}$. These statistics split the set $(-\infty, \infty)$ into the intervals

$$S_i = [\zeta_{n,i-1}, \zeta_{n,i}), \qquad i = 1, 2, \dots, n+1,$$

where we put $\zeta_{n,0} = -\infty$ and $\zeta_{n,n+1} = \infty$. These intervals are called *sampling blocks*. Consider those sampling blocks that contain exactly r random variables of $\eta_1, \eta_2, \dots, \eta_m$ and let $s_r = s_r(n, m)$ be the number of such blocks, $r = 0, 1, 2, \dots, m$. Every linear combination

$$S_l(n, m) = \sum_{r=0}^{l} c_r s_r(n, m), \qquad l = 0, 1, 2, \dots, m,$$

can be viewed as a test for the goodness-of-fit test of the hypothesis H where $c_0, c_1, \dots, c_l$ are some positive numbers. The test corresponding to the case $l = 0$ is called the *empty blocks test.* The number of blocks that do not contain any observation of the second sample is the test statistic (denoted by s_0) in this case. The following result on the asymptotic distribution of the statistic s_0 is crucial for constructing this test: if $n, m \to \infty$ such that $m/n \to \rho \in (0, \infty)$, then

$$\mathcal{L}\left(\left(\frac{(1+\rho)^3}{n\rho^2} \right)^{1/2} \left(s_0(n, m) - \frac{n}{1+\rho} \right) \Big| H \right) \xrightarrow{w} \mathcal{N}(0, 1)$$

(see [**53**]). It is proved in [**53**] that the empty blocks test against the alternative $\mathsf{P}\{\xi < x\} = F_1(x) \ne F_2(x) = \mathsf{P}\{\eta < x\}$ is consistent if the derivative $g(u)$ of the function $F_2(F_1^{-1}(u))$, $u \in [0, 1]$, differs from 1 on a nonzero Lebesgue set. If the hypothesis H is true, then $g(u) \equiv 1$, $u \in [0, 1]$. The class of such alternative hypotheses is denoted by H^*.

Test of series. As in the case of empty blocks test we deal with two independent samples $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ and $\eta^{(m)} = (\eta_1, \dots, \eta_m)$ that are independent observations of random variables ξ and η, respectively. We also assume that ξ and η have continuous distribution functions. Consider the hypothesis H that

$$\mathsf{P}\{\xi < x\} = \mathsf{P}\{\eta < x\} \quad \text{for all } x \in (-\infty, \infty).$$

It is a quite interesting case when the alternative hypothesis H_{F_1,F_2} is such that $\mathsf{P}\{\xi < x/H_{F_1,F_2}\} = F_1$, $\mathsf{P}\{\eta < x/H_{F_1,F_2}\} = F_2$, and $F_1(x) > F_2(x)$ for all x. The random variable η in this case is stochastically bigger than the random variable ξ, since for all x the random variable η exceeds x with a larger probability than ξ does. The test detecting a disagreement between the data and the hypothesis H can be constructed as follows.

First we merge the samples $\xi^{(n)}$ and $\eta^{(m)}$ and obtain the sample

$$\zeta^{n+m} = (\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_m)$$

of size $n+m$. Then we construct the order statistics for the sample ζ^{n+m}. Finally, in the sequence of order statistics, we substitute the symbol C for all members of the sample $\xi^{(n)}$ and the symbol $\overline{C}$ for all members of the sample $\eta^{(m)}$. As a result we get a sequence of n symbols C and m symbols $\overline{C}$. The total number of such sequences is $\binom{n+m}{n}$. It is clear that if H is true, then all such sequences are equiprobable (the proof is the same as that in the case of Theorem 3.3.1; also see the proof of Theorem 14.3.1 in [**53**]). If the alternative hypothesis is H_{F_1,F_2} where $F_1 > F_2$, then it is more likely that symbols $\overline{C}$ appear far away from the origin of the sequence. The measure of the displacement of the symbols $\overline{C}$ to the right can be characterized by the statistic $W(n,m)$ which is the number of series of symbols C and $\overline{C}$. Any sequence of symbols C or $\overline{C}$ is called a *series*. The number of series is small if the symbols C (or $\overline{C}$) are grouped in a specified place of the sequence. Thus the critical set for testing the hypothesis H can be taken in the form $\{W(n,m) \le t_\alpha(n,m)\}$ where $t_\alpha(n,m)$ is a certain number defined by the level α. The test related to such a critical set was proposed by Wald and Wolfowitz in 1940 and is called the *test of series.*

The following result on the limit behavior of the statistic $W(n,m)$ is useful for the evaluation of $t_\alpha(n,m)$: if $n, m \to \infty$ such that $m/n \to \rho \in (0,\infty)$, then

$$\mathcal{L}\left(\left(\frac{(1+\rho)^3}{4n\rho^2}\right)^{1/2}\left(W(n,m) - \frac{2n\rho}{1+\rho}\right)|H\right) \xrightarrow{w} \mathcal{N}(0,1)$$

(see [**53**]).

Wald and Wolfowitz proved that the test of series is consistent if the alternative hypothesis belongs to the class H^*. The test of series is discussed in detail in [**53**].

Rank tests. Most nonparametric methods use observations ranked in order of their magnitude. Statistics constructed from ranks of observations are called *rank statistics*. Tests based on rank statistics are called *rank tests*.

Let $\xi^{(n)} = (\xi_1, \dots, \xi_n)$ and $\eta^{(m)} = (\eta_1, \dots, \eta_m)$ be two independent samples of independent observations of random variables ξ and η whose distribution functions are continuous. Using the samples $\xi^{(n)}$ and $\eta^{(m)}$ we want to test the hypothesis H that $\mathsf{P}\{\xi < x\} = \mathsf{P}\{\eta < x\}$ for all $x \in (-\infty, \infty)$. We merge the samples $\xi^{(n)}$ and $\eta^{(m)}$ and obtain the sample $\zeta^{n+m} = (\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_m)$. Then we construct the sequence of order statistics $\zeta_{N,1} \le \zeta_{N,2} \le \dots \le \zeta_{N,N}$, $N = n+m$,

from the sample ζ^{n+m}. Let R_i be the index in this sequence corresponding to the member ξ_i of the sample $\xi^{(n)}$, that is, $\xi_i = \zeta_{N,R_i}$, $i = 1, 2, \ldots, n$. Thus we deal with ranks $R_1, R_2, \ldots, R_n$ of the observations $\xi_1, \xi_2, \ldots, \xi_n$. Consider the statistic $T = R_1 + \cdots + R_n$, that is, T is the sum of indices of members of the first sample $\xi_1, \ldots, \xi_n$ in the sequence of order statistics constructed from the merged sample. The test based on the rank statistic T was proposed by Wilcoxon in 1945 for the case of identical sizes of samples ($n = m$) and is called the *Wilcoxon rank sum test.*

Consider the random variable Z_{ij} that equals 1 if $\xi_i < \eta_j$ and 0 otherwise. Put

$$U = U(n, m) = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}. \tag{3.4.13}$$

It is clear that U is the total number of the cases in the merged sample where members of the sample $\xi^{(n)}$ precede members of the sample $\eta^{(m)}$. The test for testing the hypothesis H on the homogeneity based on the statistic U was studied by Mann and Whitney in 1947 and is called the *Mann–Whitney U-test.*

One can show that $T + U = nm + n(n+1)/2$. Thus the Wilcoxon rank sum test and Mann–Whitney U-test are equivalent.

Let H_{F_1,F_2} be the hypothesis that $\mathsf{P}\{\xi < x\} = F_1(x)$ and $\mathsf{P}\{\eta < x\} = F_2(x)$ where $F_1(x)$ and $F_2(x)$ are continuous distribution functions. It follows from (3.4.13) that

$$\mathsf{E}\{U/H_{F_1,F_2}\} = nm\mathsf{E}\{Z_{11}/H_{F_1,F_2}\} = nma$$

where

$$a = \mathsf{E}\{\xi_1 < \eta_1/H_{F_1,F_2}\} = \int_{-\infty}^{\infty} F_1(x)\, dF_2(x).$$

If the hypothesis H is true, that is, if $F_1(x) \equiv F_2(x)$, then $a = 1/2$. Similarly

$$\mathsf{D}\{U/H_{F_1,F_2}\} = nm\big[a + (n-1)b + (m-1)c - (n+m-1)a^2\big]$$

where

$$b = \int_{-\infty}^{\infty} F_1^2(x)\, dF_2(x), \qquad c = \int_{-\infty}^{\infty} (1 - F_2(x))^2\, dF_1(x).$$

If the hypothesis H is true, then $b = c = 1/3$ and thus $\mathsf{D}\{U/H\} = nm(n+m+1)/12$. It is known that

$$\mathcal{L}\left(\left(\frac{nm(n+m+1)}{12}\right)^{-1/2}\left(U(n,m) - \frac{nm}{2}\right)\Big|H\right) \xrightarrow{w} \mathcal{N}(0,1) \tag{3.4.14}$$

as $n, m \to \infty$

Relation (3.4.14) is useful for the Wilcoxon test of homogeneity (and for the Mann–Whitney test, too). The critical set for this test depends on the alternative hypothesis H_{F_1,F_2} and especially on the value of a: either $a < 1/2$ or $a > 1/2$ or $a = 1/2$. More details on the Wilcoxon and Mann–Whitney tests are given in [**53**]. The general theory of rank tests as well as examples of various rank tests is presented in [**22**].

Tests of independence. Below we consider a couple of tests for testing the hypothesis that two random variables ξ and η are independent. The statistical inference is based on independent observations $(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots, (\xi_n, \eta_n)$ of the vector (ξ, η). If $F_{(\xi,\eta)}(x, y)$ is the distribution function of the vector (ξ, η), then the hypothesis H is that $F_{(\xi,\eta)}(x, y) = F_\xi(x)F_\eta(y)$ where $F_\xi(x)$ and $F_\eta(y)$ are

distribution functions of the random variables ξ and η, respectively. Note that the distribution functions $F_{(\xi,\eta)}(x,y)$, $F_\xi(x)$, and $F_\eta(y)$ are unknown.

Chi-square test of independence. Let X and Y be the sets of values of random variables ξ and η, respectively. Consider the following partitions of the sets X and Y:

$$X = \bigcup_{i=1}^{r} S_i^{(1)} \quad \text{and} \quad Y = \bigcup_{i=1}^{s} S_i^{(2)}$$

where

$$S_i^{(1)} \cap S_j^{(1)} = \varnothing, \ i \neq j, \quad \text{and} \quad S_k^{(2)} \cap S_l^{(2)} = \varnothing, \ k \neq l.$$

These partitions generate the partition of the set $X \times Y$:

$$X \times Y = \bigcup_{i=1}^{r} \bigcup_{j=1}^{s} \left(S_i^{(1)} \times S_j^{(2)} \right)$$

where

$$\left(S_i^{(1)} \times S_j^{(2)} \right) \cap \left(S_k^{(1)} \times S_l^{(2)} \right) = \varnothing, \qquad (i,j) \neq (k,l).$$

Here $S_i^{(1)} \times S_j^{(2)} = \left\{ (x,y) \colon x \in S_i^{(1)}, \, y \in S_j^{(2)} \right\}$ are rectangles in the set $X \times Y$.

The hypothesis of the independence H means in this case that $p_{ij} = p_i q_j$ for all $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, s$ where

$$p_{ij} = \mathsf{P}\left\{ \xi \in S_i^{(1)}, \eta \in S_j^{(2)} / H \right\}, \qquad p_i = \mathsf{P}\left\{ \xi \in S_i^{(1)} / H \right\},$$

and

$$q_j = \mathsf{P}\left\{ \eta \in S_j^{(2)} / H \right\}.$$

Denote by ν_{ij} the total number of observations (ξ_k, η_k), $k = 1, \ldots, n$, belonging to the set $\left(S_i^{(1)} \times S_j^{(2)} \right)$, so that $\sum_{i=1}^{r} \sum_{j=1}^{s} \nu_{ij} = n$. If the probabilities p_i and q_j are known, then one can use the statistic

$$\zeta_n(p,q) = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(\nu_{ij} - n p_i q_j)^2}{n p_i q_j} \tag{3.4.15}$$

to test the hypothesis H. However these probabilities are usually unknown and thus we use their estimators

$$\widehat{p}_i = \frac{\nu_{i\cdot}}{n}, \ i = 1, \ldots, n, \quad \text{and} \quad \widehat{q}_j = \frac{\nu_{\cdot j}}{n}, \ j = 1, \ldots, n, \tag{3.4.16}$$

where

$$\nu_{i\cdot} = \sum_{j=1}^{s} \nu_{ij}, \qquad \nu_{\cdot j} = \sum_{i=1}^{r} \nu_{ij}.$$

By substituting (3.4.16) into (3.4.15) instead of the probabilities p_i and q_j we obtain the statistic

$$\widehat{\zeta}_n = \zeta_n(\widehat{p}, \widehat{q}) = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(\nu_{ij} - \nu_{i\cdot}\nu_{\cdot j}/n)^2}{\nu_{i\cdot}\nu_{\cdot j}} = n \left(\sum_{i,j} \frac{\nu_{ij}^2}{\nu_{i\cdot}\nu_{\cdot j}} - 1 \right). \tag{3.4.17}$$

The Pearson theorem implies that

$$\mathcal{L}\left(\widehat{\zeta}_n | H \right) \xrightarrow{w} \chi^2\big((r-1)(s-1)\big), \qquad n \to \infty. \tag{3.4.18}$$

Now we use statistic (3.4.17) and construct the goodness-of-fit test for testing the hypothesis H of the independence of two random variables. This is a standard procedure using the limit relation (3.4.18) and leading to the so-called χ^2 *test of independence.* More details on the χ^2 test of independence are given in [**14, 26**].

Spearman test. Let R_i be the rank of the member ξ_i in the sequence of order statistics $\xi_{n1} \le \xi_{n2} \le \cdots \le \xi_{nn}$ constructed from the sample $(\xi_1, \xi_2, \ldots, \xi_n)$. Note that the sample $(\xi_1, \ldots, \xi_n)$ consists of the first components of the members of the sample $(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots, (\xi_n, \eta_n)$. Similarly, let S_i be the rank of the member η_i in the sequence of order statistics $\eta_{n1} \le \eta_{n2} \le \cdots \le \eta_{nn}$ constructed from the sample $(\eta_1, \ldots, \eta_n)$. The sample $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ thus generates the set of pairs of ranks $(R_1, S_1), \ldots, (R_n, S_n)$. We rearrange these pairs in ascending order with respect to their first component and denote the rearranged set of pairs by $(1, T_1), (2, T_2), \ldots, (n, T_n)$.

Consider the rank statistic

$$\rho = \sum_{i=1}^{n} (R_i - \overline{R})(S_i - \overline{S}) \Big(\sum_{i=1}^{n} (R_i - \overline{R})^2 \sum_{i=1}^{n} (S_i - \overline{S})^2 \Big)^{-1/2} \tag{3.4.19}$$

which is the coefficient of correlation between two sets of ranks $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ where

$$\overline{R} = \frac{1}{n} \sum_{i=1}^{n} R_i, \qquad \overline{S} = \frac{1}{n} \sum_{i=1}^{n} S_i.$$

The statistic ρ defined by (3.4.19) is called the *Spearman rank correlation coefficient.* The test based on the statistic ρ is called the *Spearman test.*

Since $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ are certain permutations of the numbers $(1, 2, \ldots, n)$, we have

$$\overline{R} = \overline{S} = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2}, \tag{3.4.20}$$

$$\sum_{i=1}^{n} (R_i - \overline{R})^2 = \sum_{i=1}^{n} (S_i - \overline{S})^2 = \sum_{i=1}^{n} i^2 - n \Big(\frac{n+1}{2} \Big)^2 = \frac{n(n^2-1)}{12}. \tag{3.4.21}$$

Combining (3.4.19)–(3.4.21) we get

$$\begin{aligned} \rho &= \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \Big(R_i - \frac{n+1}{2} \Big) \Big(S_i - \frac{n+1}{2} \Big) \\ &= \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \Big(i - \frac{n+1}{2} \Big) \Big(T_i - \frac{n+1}{2} \Big). \end{aligned} \tag{3.4.22}$$

There is another useful formula for the Spearman coefficient, namely

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - S_i)^2 = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (i - T_i)^2. \tag{3.4.23}$$

It is straightforward to check that (3.4.22) and (3.4.23) are equivalent.

Let two distribution functions $\mathsf{P}\{\xi < x/H\}$ and $\mathsf{P}\{\eta < x/H\}$ be continuous. Since all $n!$ permutations $(T_1, T_2, \ldots, T_n)$ of the numbers $(1, 2, \ldots, n)$ are equiprobable, we have

$$\mathsf{E}\{T_i/H\} = \sum_{j=1}^{n} j \frac{(n-1)!}{n!} = \frac{n+1}{2},$$

whence we obtain by (3.4.23) that

$$\mathsf{E}\{\rho/H\} = 1 - \frac{6}{n(n^2-1)}\left[2\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n} i\mathsf{E}\{T_i/H\}\right] = 0.$$

Similarly

$$\mathsf{D}\{\rho/H\} = \frac{1}{n-1}.$$

If the ranks coincide, that is, if $R_i = S_i$, $i = 1, 2, \ldots, n$, then $\rho = 1$, while if the ranks are opposite, that is, if $T_i = n - i + 1$, $i = 1, 2, \ldots, n$, then $\rho = -1$. In general, $-1 \le \rho \le 1$. If ρ is close to either -1 or 1, then the hypothesis H is false. Thus the critical set of the Spearman test is $\{|\rho| \ge t_\alpha(n)\}$ where $t_\alpha(n)$ is defined for the level α by using the distribution of the statistic ρ. One approach to evaluate $t_\alpha(n)$ is to use for $n = 2, 3, \ldots, 30$ the tables of the distribution of the statistic ρ (see references in [**22**]). Another approach is based on the limit relation

$$\mathcal{L}\left(\sqrt{n}\rho|H\right) \xrightarrow{w} \mathcal{N}(0,1), \qquad n \to \infty,$$

whence one can also find $t_\alpha(n)$ (see [**29**], §37.28).

Kendall test. Another rank test of independence, called the *Kendall test*, is based on the statistic

$$\tau = \frac{1}{c_n^2} \sum_{i<j} \operatorname{sign}(T_j - T_i)$$

where $\operatorname{sign}(a) = 1$ for $a > 0$ and $\operatorname{sign}(a) = -1$ for $a < 0$; c_n is a certain constant. The statistic τ is called the *Kendall statistic.* It is known that

$$\mathsf{E}\{\tau/H\} = 0, \qquad \mathsf{D}\{\tau/H\} = \frac{2(2n+5)}{9n(n-1)}, \tag{3.4.24}$$
$$\mathcal{L}\left(\frac{3}{2}\sqrt{n}\tau|H\right) \xrightarrow{w} \mathcal{N}(0,1), \qquad n \to \infty$$

(see [**29**]). The critical set for the Kendall test is $\{|\tau| \ge t_\alpha(n)\}$ where the constant $t_\alpha(n)$ can be found for the level α from relation (3.4.24). It is shown in [**29**] that the Spearman and Kendall tests are asymptotically equivalent as $n \to \infty$, since the coefficient of the correlation between the statistics ρ and τ approaches 1 as $n \to \infty$.

Other rank tests of independence can be found in [**22**].

The von Mises–Smirnov test. Let $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ be a sample of size n. We treat $\xi_1, \ldots, \xi_n$ as independent observations of a random variable ξ. Let the hypothesis H be such that $\mathsf{P}\{\xi < x/H\} = F(x)$ for all $x \in (-\infty, \infty)$. One possible approach to test the hypothesis H is to use the Kolmogorov test. Consider another goodness-of-fit test for testing the hypothesis H based on the statistic

$$\omega_n^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, dF(x) \tag{3.4.25}$$

where $F_n(x)$ is the empirical distribution function constructed from the sample $\xi^{(n)}$. It is proved in [9] that if $F(x)$ is a continuous function, then

$$\lim_{n\to\infty} \mathsf{P}\{n\omega_n^2 < x/H\} = \Omega(x) = \mathsf{P}\left\{\int_0^1 \left(w^0(t)\right)^2 dt < x\right\} \tag{3.4.26}$$

where $w^0(t)$, $0 \le t \le 1$, is the Brownian bridge. We follow standard procedure to construct the goodness-of-fit test for testing the hypothesis H based on statistic (3.4.25) and on the limit relation (3.4.26). This procedure leads to the so-called *von Mises–Smirnov test.* Sometimes it also is called the ω^2 *test.*

The distribution function of $\Omega(x)$ is complicated. However one can use the tables of values of the function $\Omega(x)$ (see [6]). Note that the distribution of the statistic ω_n^2 does not depend on the function $F(x)$ and moreover

$$\mathsf{E}\{\omega_n^2/H\} = \frac{1}{6n}, \qquad \mathsf{D}\{\omega_n^2/H\} = \frac{4n-3}{180n^3} \tag{3.4.27}$$

(see [14]). More details about the ω^2 test can be found in [40].

Moran test. As in the preceding section let ξ^n be a sample and let the hypothesis H be such that $\mathsf{P}\{\xi < x/H\} = F(x)$ for all x where $F(x)$ is a continuous distribution function. Consider the statistic

$$M_n = \sum_{k=0}^{n} [F(\zeta_{n,k+1}) - F(\zeta_{n,k})]^2 \tag{3.4.28}$$

where $\zeta_{n,k}$, $k = 1, 2, \ldots, n$, are order statistics constructed from the sample $\xi^{(n)}$,

$$\zeta_{n,1} \le \zeta_{n,2} \le \cdots \le \zeta_{n,n},$$

and $F(\zeta_{n,0}) = 0$ and $F(\zeta_{n,n+1}) = 1$. The test based on the statistic (3.4.28) is called the *Moran test.* It rejects the hypothesis H if $M_n > c_n(\alpha)$ where $c_n(\alpha)$ is a constant determined by a level α and the distribution of the statistic M_n.

Since the random variable $F(\zeta_{n,k})$ is uniformly distributed on the interval $[0, 1]$, the distribution of the statistic M_n does not depend on the function $F(x)$. Thus one can consider the test based on the statistic

$$M_n = \sum_{k=0}^{n} (\zeta_{n,k+1} - \zeta_{n,k})^2. \tag{3.4.29}$$

The aim of the Moran test is to test whether the distribution of the random variable ξ is uniform on the interval $[0, 1]$. The number $c_n(\alpha)$ for the Moran test can be evaluated by applying the following assertion: *if the distribution function $F(x)$ is continuous, then*

$$\mathcal{L}\left(\sqrt{n}\left(\frac{nM_n}{2} - 1\right) | H\right) \xrightarrow{w} \mathcal{N}(0, 1)$$

as $n \to \infty$ (see [9]).

It is also proved in [9] that the Moran test is consistent. On the other hand, the Moran test does not distinguish close hypotheses (see [9]).

To conclude this section we note that many other goodness-of-fit tests are known and an extensive literature is devoted to them. Some references are given above. We also mention the book [50] on the nonparametric statistics where special attention is given to the goodness-of-fit tests.

CHAPTER 4

Sequential Tests

4.1. Bayes sequential tests of hypotheses

Setting of the problem. Let $(\Theta, \mathscr{C})$ be a measurable space and let $\xi = (\xi_1, \xi_2, \dots)$ be a sequence of independent identically distributed random variables whose distribution P_θ depends on a parameter $\theta \in \Theta$. Let $\{\Theta_1, \Theta_2, \dots, \Theta_m\}$ be a partition of the space Θ, that is, $\Theta = \bigcup_{i=1}^m \Theta_i$ and $\Theta_i \cap \Theta_j = \varnothing$, $i \neq j$. Assume that m loss functions $A_i(\theta)$, $i = 1, 2, \dots, m$, are defined on Θ. The parameter θ determining the distribution P_θ is chosen in Θ according to the a priori distribution $\mathbf{Q}$ on $(\Theta, \mathscr{C})$.

Consider the problem of testing m hypotheses $H_i: \theta \in \Theta_i$, $i = 1, 2, \dots, m$, by the sequence of random variables $\xi_1, \xi_2, \dots$. The difference between *sequential tests* and tests with a fixed size of the sample is as follows. In the case of a sequential test, a statistician is free to decide at any time whether it is necessary to terminate the sampling. More precisely, a statistician may terminate the sampling at any time n and decide to accept a certain hypothesis H_i on the basis of the sample $\xi_1, \xi_2, \dots, \xi_n$. When deciding to terminate the sampling, the statistician takes into account the cost per observation on the one hand and the amount of information about the parameter θ available in the next observation on the other hand. Let $K_n(\xi^{(n)})$ be the *expenses caused by the observation* $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$. Assume that the expenses are such that

1) $K_{n+1}(\xi^{(n+1)}) \geq K_n(\xi^{(n)})$ a.s. for all $n \geq 0$;
2) $\lim_{n\to\infty} K_n(\xi^{(n)}) = \infty$ a.s.

Every sequential test is determined by two components: a stopping rule and a decision rule. Denote by $s(\mathbf{Q}, \xi)$ a *stopping rule* and let $\nu(s)$ be a random variable denoting the size of the sample if the stopping rule $s(\mathbf{Q}, \xi)$ is applied. A *decision rule* is denoted by $\delta(\mathbf{Q}, \xi)$. Our assumption is that the decision depends only on $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ if $\nu(s) = n$. Consider the following set of sequential tests.

1. After a statistician terminates the sampling he applies a decision rule $\delta(\mathbf{Q}, \xi)$ that is assumed to be Bayes under the a priori distribution $\mathbf{Q}$. This means that if $\mathbf{Q}$ and $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ are given and $\nu(s) = n$, then

$$\delta(\mathbf{Q}, \xi) = \left(q_1^\delta(\xi^{(n)}), q_2^\delta(\xi^{(n)}), \dots, q_n^\delta(\xi^{(n)})\right)$$

where the functions $q_i^\delta(\xi^{(n)})$ are defined by (1.3.42) with $\xi^{(n)}$ instead of x. We also assume that the measure P_θ is absolutely continuous with respect to some σ-finite measure μ and that its density is $p(x; \theta)$. Then the distribution of the sample $\xi^{(n)}$ is absolutely continuous with respect to the measure $\mu^n = \mu \times \mu \times \cdots \times \mu$ (n times) and its density is $p_n(x^{(n)}; \theta) = \prod_{i=1}^n p(x_i; \theta)$, $x^{(n)} = (x_1, x_2, \dots, x_n)$. Thus $p_n(x^{(n)}, t)$ should be used in (1.3.40) and (1.3.41) instead of $p(x; t)$.

2. The class of stopping rules $s(\mathbf{Q}, \xi)$ is the collection of all randomized rules that can be described as follows. For any $n \geq 1$, denote by $\mathscr{F}_n$ the σ-algebra generated by the vector $(\theta, \xi_1, \xi_2, \dots, \xi_n)$ and let $\mathscr{F}_0 = \mathscr{C}$. It is clear that

$$\mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots.$$

Let $\mathscr{S}$ denote the class of all sequences $s = (s_1, s_2, \dots)$ such that $s_1 = s_1(\mathbf{Q})$ and $s_{n+1} = s_{n+1}(\mathbf{Q}, \xi_1, \xi_2, \dots, \xi_n)$ is an $\mathscr{F}_n$-measurable random variable,

$$0 \leq s_{n+1}(\mathbf{Q}, \xi_1, \xi_2, \dots, \xi_n) \leq 1, \qquad n \geq 1.$$

We treat s_{n+1} as the probability that the random variable ξ_{n+1} occurs in the sampling. Set

$$J_{n+1} = J_{n+1}(\mathbf{Q}, \xi^{(n)}) = \begin{cases} 1, & \text{with probability } s_{n+1}(\mathbf{Q}, \xi^{(n)}), \\ 0, & \text{with probability } 1 - s_{n+1}(\mathbf{Q}, \xi^{(n)}). \end{cases}$$

Following the stopping rule generated by the sequence $s \in \mathscr{S}$ a statistician decides to terminate the sampling at the minimal $n \geq 0$ such that $J_{n+1} = 0$. The corresponding size of the sample is a random variable given by

$$\nu(s) = \min\left\{n \geq 0 \colon J_{n+1}(\mathbf{Q}, \xi^{(n)}) = 0\right\}.$$

It is clear that $\{\nu(s) = n\} \in \mathscr{F}_n$, $n \geq 0$, that is, $\nu(s)$ is a *stopping rule* with respect to the family of σ-algebras $\mathscr{F}_0, \mathscr{F}_1, \mathscr{F}_2, \dots$.

Given $(\mathbf{Q}, \xi^{(n)})$ denote by $\mathbf{Q}(\cdot/\xi^{(n)})$ the a posteriori distribution of the parameter θ. Let P_Q^n be the joint unconditional (weighted) distribution of the vector $\xi^{(n)}$ given $\mathbf{Q}$. Denote by $\rho_0(\mathbf{Q}, \xi^{(n)})$ the Bayes risk corresponding to the Bayes decision for given $(\mathbf{Q}, \xi^{(n)})$ and $\nu(s) = n$, that is,

$$\rho_0(\mathbf{Q}, \xi^{(n)}) = \min_{1 \leq i \leq m} \int_\Theta A_i(\theta)\, \mathbf{Q}(d\theta/\xi^{(n)}).$$

Then *a priori risk* corresponding to the stopping rule s and the Bayes decision rule is defined by

$$\begin{aligned} R(\mathbf{Q}; s) = \sum_{n=1}^{\infty} \int \Big[K_n(x^{(n)}) + \rho_0(\mathbf{Q}, x^{(n)})\Big] \\ \times \prod_{k=0}^{n-1} s_{k+1}(\mathbf{Q}, x^{(k)})\big(1 - s_{n+1}(\mathbf{Q}, x^{(n)})\big)\, \mathsf{P}_Q^n(dx^{(n)}) \\ + (1 - s_1(\mathbf{Q}))\rho_0(\mathbf{Q}) \end{aligned}$$

where $s_1(\mathbf{Q}, x^{(0)}) = s_1(\mathbf{Q})$ and $\rho_0(\mathbf{Q})$ is the Bayes risk to make a decision without sampling, that is,

$$\rho_0(\mathbf{Q}) = \min_{1 \leq i \leq m} \int_\Theta A_i(\theta)\, \mathbf{Q}(d\theta).$$

A stopping rule s^* is called *Bayes* (or *optimal*) for the a priori distribution $\mathbf{Q}$ if

$$R(\mathbf{Q}, s^*) = \inf_{s \in \mathscr{S}} R(\mathbf{Q}, s).$$

In what follows we assume that $R(\mathbf{Q}, s^*) < \infty$.

The problem of evaluating a Bayes stopping rule s^* for a given a priori distribution is not easy. Below we consider some general properties of the rule s^*.

Properties of a Bayes stopping rule. Let $\mathscr{S}^{(N)}$ be a *subclass of stopping rules* $s \in \mathscr{S}$ *truncated at the moment* $\nu = N$, that is, $s_{j+1}(\mathbf{Q}, \xi^{(j)}) = 0$ for all $j \geq N$ and all $s \in \mathscr{S}^{(N)}$. Below we obtain a Bayes truncated rule $s^{(N)} \in \mathscr{S}^{(N)}$ such that

$$R(\mathbf{Q}, s^{(N)}) = \inf_{s \in \mathscr{S}^{(N)}} R(\mathbf{Q}, s).$$

According to the dynamic programming method (see [**15**]), we define the rule $s^{(N)}$ by constructing the $N+1$ functions

$$\rho_0^{(N)}(\mathbf{Q}, \xi^{(N)}) = \rho_0(\mathbf{Q}, \xi^{(N)}),$$

$$\rho_j^{(N)}(\mathbf{Q}, \xi^{(N-j)}) = \min\left\{\rho_0(\mathbf{Q}, \xi^{(N-j)}), \mathsf{E}\{\Delta(\eta; \xi^{(N-j)}) + \rho_{j-1}^{(N)}(\mathbf{Q}, (\xi^{(N-j)}, \eta))/\mathscr{F}_{N-j}\}\right\},$$

$j = 1, 2, \ldots, N$, where $\mathsf{E}\{\cdot/\mathscr{F}_{N-j}\}$ is the conditional expectation with respect to the distribution

$$\mathsf{P}(\cdot/\mathscr{F}_{N-j}) = \int_\Theta \mathsf{P}_\theta(\cdot)\, \mathbf{Q}(d\theta/\xi^{(N-j)}), \qquad j = 0, 1, \ldots, N,$$

and

$$\Delta(\eta; \xi^{(N-j)}) = K_{N-j+1}(\xi^{(N-j)}, \eta) - K_{N-j}(\xi^{(n-j)}).$$

Let

$$\nu^{(N)} = \min\left\{n: 0 \leq n \leq N, \rho_{N-n}^{(N)}(\mathbf{Q}, \xi^{(n)}) = \rho_0(\mathbf{Q}, \xi^{(n)})\right\}.$$

The number $\nu^{(N)}$ is the size of the sample corresponding to the Bayes truncated stopping rule $s^{(N)}$. Since $\{\nu^{(N)} = n\} \in \mathscr{F}_n$ for all $n \geq 0$, $\nu^{(N)}$ is a stopping rule. The sequence $s^{(N)} = (s_1^{(N)}, s_2^{(N)}, \ldots)$ is a Bayes stopping rule, and moreover

$$s_j^{(N)} = s_j^{(N)}(\mathbf{Q}, \xi^{(j-1)}) = \begin{cases} 1, & \text{if } j \leq \nu(N), \\ 0, & \text{if } j > \nu(N). \end{cases}$$

For any $N \geq 1$, $\rho_N^{(N)}(\mathbf{Q})$ is the a priori risk corresponding to $\mathbf{Q}$ and $\mathscr{S}^{(N)}$. Denote by $\mathscr{S}_1^{(N)}$, $N \geq 1$, the subclass of stopping rules of $\mathscr{S}^{(N)}$ for which at least one observation is taken. Then

$$\rho_N^{(N)}(\mathbf{Q}) = \min\left\{\rho_0(\mathbf{Q}), \inf_{s \in \mathscr{S}_1^{(N)}} R(\mathbf{Q}, s)\right\}. \tag{4.1.1}$$

Since $\mathscr{S}_1^{(N)} \subset \mathscr{S}_1^{(N+1)}$ for all $N \geq 1$, we have

$$\inf_{s \in \mathscr{S}_1^{(N+1)}} R(\mathbf{Q}, s) \leq \inf_{s \in \mathscr{S}_1^{(N)}} R(\mathbf{Q}, s). \tag{4.1.2}$$

It follows from (4.1.1) and (4.1.2) that $\rho_N^{(N)}(\mathbf{Q}) \geq \rho_{N+1}^{(N+1)}(\mathbf{Q})$ for all $N \geq 1$. Since $\rho_n^{(N)}(\mathbf{Q}) \geq 0$ for all $N \geq 1$, the limit

$$\rho^{(\infty)}(\mathbf{Q}) = \lim_{N \to \infty} \rho_N^{(N)}(\mathbf{Q})$$

exists.

By $\rho(\mathbf{Q})$ we denote the a priori Bayes risk related to the class $\mathscr{S}$ of all stopping rules, that is, $\rho(\mathbf{Q}) = R(\mathbf{Q}, s^*)$. The risk $\rho(\mathbf{Q})$ satisfies the equation

$$\rho(\mathbf{Q}) = \min\{\rho_0(\mathbf{Q}),\ \mathsf{E}_Q\{K_1(\eta) + \rho(\mathbf{Q}_\eta)\}\} \tag{4.1.3}$$

where $\mathbf{Q}_\eta(\cdot) = \mathbf{Q}(\cdot/\eta)$ is the a posteriori distribution of the parameter θ given $(\mathbf{Q}, \eta)$.

Indeed, if $\mathscr{S}_1$ is the class of all stopping rules for which at least one observation is taken, then

$$\rho(\mathbf{Q}) = \min\left\{\rho_0(\mathbf{Q}),\ \inf_{s\in\mathscr{S}_1} R(\mathbf{Q}, s)\right\}. \tag{4.1.4}$$

For all $s \in \mathscr{S}_1$, $s = (s_1(\mathbf{Q}), s_2(\mathbf{Q}, \xi_1), s_3(\mathbf{Q}, \xi_1, \xi_2), \dots)$, we define the *reduced* stopping rule $s^+ = (s_2(\mathbf{Q}, \xi_1), s_3(\mathbf{Q}, \xi_1, \xi_2), \dots)$. Thus we have for all $s \in \mathscr{S}_1$ that

$$R(\mathbf{Q}, s) = \mathsf{E}_Q\left\{K_1(\eta) + R(\mathbf{Q}_\eta, s^+)\right\}. \tag{4.1.5}$$

Equality (4.1.5) and the Fatou lemma imply that

$$\begin{aligned}\inf_{s\in\mathscr{S}_1} R(\mathbf{Q}, s) &\geq \mathsf{E}_Q\left\{K_1(\eta) + \inf_{s\in\mathscr{S}_1} R(\mathbf{Q}_\eta, s^+)\right\} \\ &= \mathsf{E}_Q\left\{K_1(\eta) + \inf_{s\in\mathscr{S}_1} R(\mathbf{Q}_\eta, s)\right\} = \mathsf{E}_Q\left\{K_1(\eta) + \rho(\mathbf{Q}_\eta)\right\}.\end{aligned} \tag{4.1.6}$$

We obtain from (4.1.4) and (4.1.6) that

$$\rho(\mathbf{Q}) \geq \min\{\rho_0(\mathbf{Q}), \mathsf{E}_Q\{K_1(\eta) + \rho(\mathbf{Q}_\eta)\}\} \geq \inf_{s\in\mathscr{S}} R(\mathbf{Q}, s) = \rho(\mathbf{Q}),$$

that is, $\rho(\mathbf{Q})$ satisfies equation (4.1.3).

Similarly we show that for all $N \geq 1$

$$\rho_{N+1}^{(N+1)}(\mathbf{Q}) = \min\left\{\rho_0(\mathbf{Q}), \mathsf{E}_Q\left\{K_1(\eta) + \rho_N^{(N)}(\mathbf{Q}_\eta)\right\}\right\}. \tag{4.1.7}$$

If $\mathsf{E}_Q\rho_1^{(1)}(\mathbf{Q}_\eta) < \infty$, then the Lebesgue dominated convergence theorem implies that

$$\begin{aligned}\rho^{(\infty)} &= \lim_{N\to\infty} \rho_{N+1}^{(N+1)}(\mathbf{Q}) = \min\left\{\rho_0(\mathbf{Q}), \lim_{N\to\infty}\left\{K_1(\eta) + \rho_N^{(N)}(\mathbf{Q}_\eta)\right\}\right\} \\ &= \min\left\{\rho_0(\mathbf{Q}), \mathsf{E}_Q\left\{K_1(\eta) + \rho^{(\infty)}(\mathbf{Q}_\eta)\right\}\right\},\end{aligned}$$

since the risk $\rho_N^{(N)}$, $N \geq 1$, is monotone. Thus the limit $\rho^{(\infty)}(\mathbf{Q})$ also satisfies equation (4.1.3). Therefore we have proved the following result.

THEOREM 4.1.1. *If* $\mathsf{E}_Q K_1(\eta) < \infty$ *and* $\mathsf{E}_Q\rho_1^{(1)}(\eta_\eta) < \infty$ *for a given a priori distribution* $\mathbf{Q}$, *then*

$$\rho(\mathbf{Q}) = \lim_{N\to\infty} \rho_N^{(N)}(\mathbf{Q}).$$

Theorem 4.1.1 implies for all $\varepsilon > 0$ that there exists an integer number $N(\varepsilon)$ such that the Bayes truncated stopping rule $s^{(N)}$ for all $N \geq N(\varepsilon)$ is an ε-Bayes stopping rule for the nontruncated problem, that is,

$$0 \leq \rho_N^{(N)}(\mathbf{Q}) - \rho(\mathbf{Q}) < \varepsilon.$$

Below are two results about the existence of a Bayes (optimal) stopping rule.

THEOREM 4.1.2. *If* $\rho(\mathbf{Q}) = \lim_{N\to\infty} \rho_N^{(N)}(\mathbf{Q})$ *and for all* $n \geq N_0$

$$\rho_0(\mathbf{Q}, \xi^{(n)}) - \mathsf{E}\left\{\rho_0(\mathbf{Q}, (\xi^{(n)}, \eta))/\mathscr{F}_n\right\} \leq \mathsf{E}\left\{\Delta(\eta; \xi^{(n)})/\mathscr{F}_n\right\}$$

almost surely, then $\rho_{N_0}^{(N_0)} = \rho(\mathbf{Q})$ *and the truncated Bayes sequential rule* $s^{(N_0)}$ *is optimal for the nontruncated problem.*

THEOREM 4.1.3. *If for all* $n \geq 1$ *the a posteriori risk* $\rho_0(\mathbf{Q}, \xi)$ *does not depend on* $\xi^{(n)}$, *that is,* $\rho_0(\mathbf{Q}, \xi^{(n)}) = \rho_0(\mathbf{Q}, n)$ *almost surely for all* $n \geq 1$, *then the Bayes sequential stopping rule is the one with a fixed size* n_0 *of the sample where* n_0 *is the minimal positive integer number* n *that minimizes* $\rho_0(\mathbf{Q}, n) + \mathsf{E}_Q K_n(\xi^{(n)})$.

The proofs of Theorems 4.1.2 and 4.1.3 can be found in [**54**].

More details on the general theory of Bayes sequential rules are given in [**4, 15, 54**].

The evaluation of Bayes stopping rules is a complicated problem, especially in the case of composite hypotheses. The following example shows that this problem is complicated even in the case of two simple hypotheses.

EXAMPLE 4.1.1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables assuming two values 1 and 0 with probabilities θ and $1 - \theta$, respectively, where θ is an unknown parameter. Let θ assume only two values $1/3$ and $2/3$, that is, $\Theta = \{1/3, 2/3\}$. Let the a priori measure $\mathbf{Q}$ be determined by the number $q = \mathsf{P}\{\theta = 1/3\} = 1 - \mathsf{P}\{\theta = 2/3\}$. Thus we deal with two simple hypotheses $H_1: \theta = 1/3$ and $H_2: \theta = 2/3$. Let the loss functions $A_i(\theta)$ be defined by the numbers $A_1(1/3) = A_2(2/3) = 0$ and $A_1(2/3) = A_2(1/3) = 20$, while the cost per observation is determined by the equality $K_n(\xi^{(n)}) = n$. This means that a single observation costs 1 dollar. Put $\rho_N(q) = \rho_N^{(N)}(\mathbf{Q})$, $N = 0, 1, 2, \ldots$. Now we evaluate $\rho_0(q)$, $\rho_1(q)$, and $\rho_2(q)$. The definition of $\rho_0(q)$ implies that without sampling one accepts the hypothesis H_2 if $0 \leq q \leq 1/2$, and the hypothesis H_1 if $1/2 \leq q \leq 1$. Moreover

$$\rho_0(q) = \begin{cases} 20q, & \text{if } 0 \leq q \leq 1/2, \\ 20(1-q), & \text{if } 1/2 \leq q \leq 1. \end{cases} \tag{4.1.8}$$

This implies that $\rho_0(q) = \rho_0(1-q)$ for $0 \leq g \leq 1$. Since the problem is symmetric, $\rho_j(q) = \rho_j(1-q)$ for $0 \leq q \leq 1$ and for all $j = 1, 2, \ldots$. Thus we need to evaluate $\rho_1(q)$ and $\rho_2(q)$ only for $q \in [0, 1/2]$.

Let $q(x)$ stand for the a priori probability of the event $\{\theta = 1/3\}$ given the observation η is equal to x where either $x = 0$ or $x = 1$. Applying the Bayes formula we obtain

$$q(1) = \frac{q}{q + 2(1-q)}, \qquad q(0) = \frac{2q}{2q + (1-q)}. \tag{4.1.9}$$

It follows from (4.1.8) and (4.1.9) that

$$\rho_0(q(1)) = 20q(1) \quad \text{for } 0 \leq q \leq 1/2,$$
$$\rho_0(q(0)) = \begin{cases} 20q(0), & \text{if } 0 \leq q \leq 1/3, \\ 20(1 - q(0)), & \text{if } 1/3 \leq q \leq 1/2. \end{cases} \tag{4.1.10}$$

The unconditional distribution of the observation η is given by

$$\mathsf{P}\{\eta = 1\} = \frac{1}{3}q + \frac{2}{3}(1-q) = 1 - \mathsf{P}\{\eta = 0\}. \tag{4.1.11}$$

It is clear that

$$\begin{aligned} \mathsf{E}\rho_0(q(\eta)) &= \rho_0(q(1))\mathsf{P}\{\eta = 1\} + \rho_0(q(0))\mathsf{P}\{\eta = 0\} \\ &= \begin{cases} 20q, & \text{if } 0 \le q \le 1/3, \\ 20/3, & \text{if } 1/3 \le q \le 1/2. \end{cases} \end{aligned} \tag{4.1.12}$$

Since $K_1(\xi^{(1)}) = 1$, we derive from (4.1.7), (4.1.8), and (4.1.12) that

$$\rho_1(q) = \min\{\rho_0(q), \mathsf{E}\rho_0(q(\eta)+1)\} = \begin{cases} 20q, & \text{if } 0 \le q \le 23/60, \\ 23/3, & \text{if } 23/60 \le q \le 1/2. \end{cases} \tag{4.1.13}$$

First we consider the case $N \le 1$. Relations (4.1.8) and (4.1.13) imply that if $0 \le q \le 23/60$ or $37/60 \le q \le 1$, then one should make the final decision without sampling, while if $23/60 < q < 37/60$, then it is necessary to take the first observation and one should make the final decision based on a sample consisting of a single observation.

Now let $N \le 2$. To evaluate $p_2(q)$ we note that (4.1.9) implies that there are three pairs of equivalent inequalities, namely

$$\begin{aligned} &q(1) \le 23/60 \text{ and } q \le 46/83 \iff \\ &q(0) \le 23/60 \text{ and } q \le 23/97 \iff \\ &q(0) \ge 37/60 \text{ and } q \ge 37/83. \end{aligned}$$

Using (4.1.11) and (4.1.13) and taking into account the symmetry of the function $\rho_1(q)$ we get

$$\mathsf{E}\rho_1(q(\eta)) = \begin{cases} 20q, & \text{if } 0 \le q \le \frac{23}{97}, \\ \frac{83q+23}{9}, & \text{if } \frac{23}{97} < q \le \frac{37}{83}, \\ \frac{20}{3}, & \text{if } \frac{37}{83} < q \le \frac{1}{2}. \end{cases} \tag{4.1.14}$$

Thus (4.1.7) yields

$$\rho_2(q) = \min\{\rho_0(q), \mathsf{E}\rho_1(q(\eta)+1)\} = \begin{cases} 20q, & \text{if } 0 \le q \le \frac{32}{97}, \\ \frac{83q+32}{9}, & \text{if } \frac{32}{97} < q \le \frac{37}{83}, \\ \frac{23}{3}, & \text{if } \frac{37}{83} < q \le \frac{1}{2}. \end{cases} \tag{4.1.15}$$

Relations (4.1.8), (4.1.13), and (4.1.15) show in the case $N \le 2$ that if

$$0 \le q \le 32/97 \quad \text{or} \quad 65/97 \le q \le 1,$$

then a final decision should be made without sampling, while if $37/83 \le q \le 46/83$, then it is necessary to take the first observation and a final decision should be made based on a sample consisting of a single observation. Finally if

$$32/97 < q < 37/83 \quad \text{or} \quad 46/83 < q < 65/97,$$

then it is necessary to take the second observation and a final decision should be made based on a sample consisting of two observations.

The graphs of the functions $\rho_0(q)$, $\rho_1(q)$, and $\rho_2(q)$ are shown in Figure 4.1.1. The evaluation of the functions $\rho_N(q)$ becomes a time consuming procedure for

large N, since the number of intervals where they change their slopes increases with N. If the size of the sample is not bounded in advance, then the optimal procedure is described in Theorems 4.1.1–4.1.3 (see also Chapter 4 in [15]).

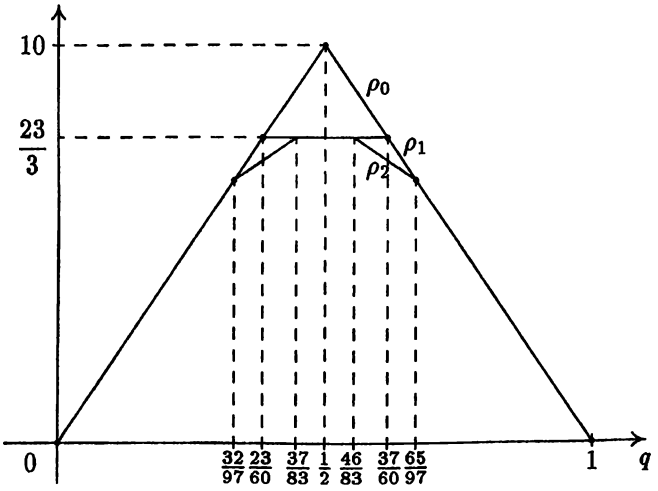

FIGURE 4.1.1. Risk functions $\rho_0(q)$, $\rho_1(q)$, and $\rho_2(q)$

EXAMPLE 4.1.2. This is a continuation of Example 4.1.1. Let the loss functions $A_i(\theta)$ be determined by the numbers $A_1(1/3) = A_2(2/3) = 0$ and

$$A_1(2/3) = A_2(1/3) = 10.$$

As in Example 4.1.1 we have $\rho_j(q) = \rho_j(1-q)$ for $0 \le q \le 1$ and $j = 0, 1, 2, \ldots$. Thus $\rho_0(q) = 10q$ for $0 \le q \le 1/2$ and $\rho_0(q) = 10(1-q)$ for $1/2 \le q \le 1$. Similarly to Example 4.1.1 we get

$$\mathsf{E}\rho_0(q(\eta)) = \begin{cases} 10q, & \text{if } 0 \le q \le 1/3, \\ 10/3, & \text{if } 1/3 < q \le 1/2. \end{cases}$$

Hence (4.1.7) implies that

$$\rho_1(q) = \begin{cases} 10q, & \text{if } 0 \le q \le 13/30, \\ 13/3, & \text{if } 13/30 < q \le 1/2. \end{cases} \tag{4.1.16}$$

Relations (4.1.9) imply that $q(1) \le 13/30 \iff q \le 26/43$, $q(0) \le 13/30 \iff q \le 13/47$ and $q(0) \ge 17/30 \iff q \ge 17/43$. Thus

$$\mathsf{E}\rho_1(q(\eta)) = \begin{cases} 10q, & \text{if } 0 \le q \le \frac{13}{47}, \\ \frac{43q+13}{9}, & \text{if } \frac{13}{47} < q \le \frac{17}{43}, \\ \frac{10}{3}, & \text{if } \frac{17}{43} < q \le \frac{1}{2}. \end{cases}$$

Using (4.1.7) we obtain for $N = 2$ that

$$\rho_2(q) = \begin{cases} 10q, & \text{if } 0 \le q \le 13/30, \\ 13/3, & \text{if } 13/30 < q \le 1/2. \end{cases} \tag{4.1.17}$$

Relations (4.1.16) and (4.1.17) imply that $\rho_1(q) = \rho_2(q)$ for all $q \in [0,1]$. Moreover we derive from (4.1.7) that

$$\rho_3(q) = \min\{\rho_0(q), \mathsf{E}\rho_2(q(\eta)) + 1\} = \min\{\rho_0(q), \mathsf{E}\rho_1(q(\eta)) + 1\} = \rho_2(q)$$

if $0 \le q \le 1$. We conclude by induction that $\rho_N(q) = \rho_1(q)$ for $N = 2, 3, \dots$, that is, the conditions of Theorem 4.1.3 hold.

The above discussion shows that if $13/30 < q < 17/30$, then it is necessary to take the first observation and a decision should be made based on a sample consisting of a single observation, while otherwise a decision can be made without any observation.

In the next section we consider the Wald sequential test for distinguishing two simple hypotheses in the case of general distributions of observations $\xi_1, \xi_2, \dots$.

4.2. Wald sequential tests

Main definitions and notation. Let $\xi = (\xi_1, \xi_2, \dots)$ be a sequence of independent identically distributed random variables with a distribution P_θ where θ is an unknown parameter assuming only two values θ_1 and θ_2. Thus we deal with the case $\Theta = \{\theta_1, \theta_2\}$ and each of the sets $\Theta_1 = \{\theta_1\}$ and $\Theta_2 = \{\theta_2\}$ contains only a single point. Therefore there are two simple hypotheses about the distribution of an observation, namely $H_1\colon \theta = \theta_1$ and $H_2\colon \theta = \theta_2$. Let the distribution P_θ be absolutely continuous with respect to some σ-finite measure μ and denote the density by $p(x;\theta)$. Then the measure $\mathsf{P}_\theta^{(n)}$ determining the distribution of the sample $\xi^{(n)} = (\xi_1, \xi_2, \dots, \xi_n)$ is absolutely continuous with respect to the measure $\mu^n = \mu \times \mu \times \cdots \times \mu$ and its density is $p_n(x^{(n)};\theta) = \prod_{k=1}^n p(x_k;\theta)$, $x^{(n)} = (x_1, x_2, \dots, x_n)$. Let $z_n(x^{(n)})$ be the likelihood ratio

$$z_n(x^{(n)}) = \frac{p_n(x^{(n)};\theta_2)}{p_n(x^{(n)};\theta_1)} = \prod_{k=1}^n \frac{p(x_k;\theta_2)}{p(x_k;\theta_1)} \tag{4.2.1}$$

where $x^{(n)} = (x_1, x_2, \dots, x_n)$ (we agree that $0/0 = 0$). Put

$$\Lambda_n(x^{(n)}) = \ln z_n(x^{(n)}), \qquad \Lambda_n = \Lambda_n(\xi^{(n)}), \qquad z_n = z_n(\xi^{(n)}). \tag{4.2.2}$$

Therefore

$$\Lambda_n = \sum_{k=1}^n \lambda_k, \qquad \lambda_k = \lambda(\xi_k) = \ln \frac{p(\xi_k;\theta_2)}{p(\xi_k;\theta_1)}, \quad k = 1, 2, \dots, n. \tag{4.2.3}$$

Throughout this section we consider sequential tests whose stopping times belong to the class $\mathscr{S}_1$ (in other words, the tests for which it is necessary to take at least one observation) and which depend on two constants a and b such that $0 < a$ and $b < \infty$. Such a test is called a *Wald sequential test* for distinguishing two simple hypotheses H_1 and H_2 if its stopping rule says that a statistician continues the sampling until $-b < \Lambda_n < a$, that is, the Wald stopping time ν is

$$\nu = \inf\{n \ge 0\colon \Lambda_n \notin (-b, a)\}. \tag{4.2.4}$$

If $\nu = n$, then the decision is as follows: the hypothesis H_2 is accepted if $\Lambda_n \ge a$, while the alternative H_1 is accepted if $\Lambda_n \le -b$. The test described above is also called the *Wald sequential test with limit points* $(-b, a)$.

The above test is often called a *Wald sequential likelihood ratio test* or *sequential likelihood ratio test* or *Wald test.* For a Wald test we denote by α and β the *type* I and *type* II *error probabilities*, respectively, that is,

$$\alpha = \mathsf{P}_{\theta_1}\{\Lambda_\nu \geq a\}, \qquad \beta = \mathsf{P}_{\theta_2}\{\Lambda_\nu \leq -b\} \tag{4.2.5}$$

where $\mathsf{P}_{\theta_i}\{\cdot\}$ are the probabilities of events if the distribution of $\xi_1, \xi_2, \ldots$ is determined by the measure P_{θ_i}. The pair of numbers (α, β) is called the *power of a test.* We say that a *Wald sequential test is terminated with probability one during a finite time* if $\mathsf{P}_{\theta_i}\{\nu < \infty\} = 1$ for all $i = 1, 2$.

Inequalities for error probabilities of a Wald test. The following two results establish relationships between the power (α, β) of a Wald test and its limit points $(-b, a)$.

LEMMA 4.2.1. *If a Wald sequential test of power (α, β) and with limit points $(-b, a)$ is terminated with probability one during a finite time, then*

$$A \leq \frac{1-\beta}{\alpha}, \qquad B \geq \frac{\beta}{1-\alpha} \tag{4.2.6}$$

where $0 < A < 1 < B < \infty$ and $\ln B = -b$, $\ln A = a$.

PROOF. Consider the sets

$$W_n = \Big\{x^{(n)}\colon B < z_k\big(x^{(k)}\big) < A \text{ for all } k = 1, 2, \ldots, n-1, \; z_n\big(x^{(n)}\big) \geq A\Big\}, \tag{4.2.7}$$

$$V_n = \Big\{x^{(n)}\colon B < z_k\big(x^{(k)}\big) < A \text{ for all } k = 1, 2, \ldots, n-1, \; z_n\big(x^{(n)}\big) \leq B\Big\} \tag{4.2.8}$$

where $n = 1, 2, \ldots$, $x^{(n)} = (x_1, x_2, \ldots, x_n)$, and $z_n(x^{(n)})$ is the likelihood ratio (4.2.1). Using (4.2.2)–(4.2.4) and (4.2.7)–(4.2.8) we get

$$\{\Lambda_\nu \geq a\} = \bigcup_{n=1}^{\infty} \Big\{\xi^{(n)} \in W_n\Big\}, \qquad \{\Lambda_\nu \leq b\} = \bigcup_{n=1}^{\infty} \Big\{\xi^{(n)} \in V_n\Big\}. \tag{4.2.9}$$

Relations (4.2.5) and (4.2.9) imply that

$$\alpha = \sum_{n=1}^{\infty} \int_{W_n} \mathsf{P}_{\theta_1}^{(n)}(dx) = \sum_{n=1}^{\infty} \int_{W_n} \frac{1}{z_n(x)}\, \mathsf{P}_{\theta_2}^{(n)}(dx) \leq \frac{1}{A} \sum_{n=1}^{\infty} \int_{W_n} \mathsf{P}_{\theta_2}^{(n)}(dx) = \frac{1-\beta}{A},$$

that is, the first inequality in (4.2.6) is proved. Here we used for $i = 1$ and $i = 2$ that

$$\sum_{n=1}^{\infty} \int_{W_n} \mathsf{P}_{\theta_i}^{(n)}(dx) + \sum_{n=1}^{\infty} \int_{V_n} \mathsf{P}_{\theta_i}^{(n)}(dx) = \sum_{n=1}^{\infty} \mathsf{P}_{\theta_i}\{\nu = n\} = \mathsf{P}_{\theta_i}\{\nu < \infty\} = 1 \tag{4.2.10}$$

by the assumptions of the lemma. Similarly, relations (4.2.5) and (4.2.9) imply

$$1-\alpha=\sum_{n=1}^{\infty}\int_{V_n}\mathsf{P}_{\theta_1}^{(n)}(dx)=\sum_{n=1}^{\infty}\int_{V_n}\frac{1}{z_n(x)}\,\mathsf{P}_{\theta_2}^{(n)}(dx)\geq\frac{1}{B}\sum_{n=1}^{\infty}\int_{V_n}\mathsf{P}_{\theta_2}^{(n)}(dx)=\frac{\beta}{B},$$

that is, the second inequality in (4.2.6) is also proved. □

LEMMA 4.2.2. *Let $A=(1-\beta)/\alpha$, $B=\beta/(1-\alpha)$ and $-b=\ln B$, $a=\ln A$. If a Wald sequential test with limit points $(-b,a)$ is terminated with probability one during a finite time and its power is (α',β'), then*

$$\alpha'\leq\frac{\alpha}{1-\beta},\qquad\beta'\leq\frac{\beta}{1-\alpha}, \tag{4.2.11}$$

$$\alpha'+\beta'\leq\alpha+\beta. \tag{4.2.12}$$

PROOF. According to Lemma 4.2.1,

$$A=\frac{1-\beta}{\alpha}\leq\frac{1-\beta'}{\alpha'},\qquad B=\frac{\beta}{1-\alpha}\geq\frac{\beta'}{1-\alpha'}. \tag{4.2.13}$$

This implies

$$\alpha'\leq(1-\beta')\frac{\alpha}{1-\beta}\leq\frac{\alpha}{1-\beta},\qquad\beta'\leq(1-\alpha')\frac{\beta}{1-\alpha}\leq\frac{\beta}{1-\alpha},$$

that is, inequalities (4.2.11) hold. Moreover (4.2.13) yields

$$\alpha'(1-\beta)\leq\alpha(1-\beta'),\qquad\beta'(1-\alpha)\leq\beta(1-\alpha').$$

Combining these inequalities we obtain

$$\alpha'-\alpha'\beta+\beta'-\alpha\beta'\leq\alpha-\alpha\beta'+\beta-\alpha'\beta,$$

whence inequality (4.2.12) follows. □

REMARK 4.2.1. Lemma 4.2.2 for small α and β implies that the power of a Wald test with limit points $\big(\ln(\beta/(1-\alpha)),\ln((1-\beta)/\alpha)\big)$ is (α',β') where α' is close to α, β' is close to β, and always $\alpha'+\beta'\leq\alpha+\beta$, that is, $\beta'<\beta$ if $\alpha'>\alpha$ and $\alpha'<\alpha$ if $\beta'>\beta$.

Properties of the stopping time of a Wald sequential test. The following result contains sufficient conditions for the finiteness of the Wald stopping time ν and its moment generating function.

LEMMA 4.2.3. *Let ν be the stopping time of a Wald sequential test with limit points $(-b,a)$. Let $\mathsf{P}_\theta\{|\lambda_1|>0\}>0$ where either $\theta=\theta_1$ or $\theta=\theta_2$, and $\lambda_1=\lambda_1(\xi_1)$. Then*

a) $\mathsf{P}_\theta\{\nu<\infty\}=1$;

b) $\mathsf{E}_\theta e^{t\nu}<\infty$ *for all $t<t_0$ where t_0 is a positive number and E_θ is the expectation with respect to the distribution P_θ.*

PROOF. Let m and k be fixed integer numbers such that $m > k$ and $r = [m/k]$ where $[c]$ is the integer part of a number c. Consider the random variables

$$T_1 = \Lambda_k, \qquad T_i = \Lambda_{ik} - \Lambda_{(i-1)k}, \quad i = 2, 3, \ldots, r.$$

If $\nu > m$, then $\Lambda_i \in (-b, a)$ for all $i = 1, 2, \ldots, m$. In particular, this inclusion holds for $i = k, 2k, \ldots, rk$. Thus

$$|T_i| < b \vee a = c, \qquad i = 1, 2, \ldots, r.$$

Since the random variables $T_1, T_2, \ldots, T_r$ are independent and identically distributed, we have

$$\mathsf{P}_\theta\{\nu > m\} \le \mathsf{P}_\theta\{|T_i| < c \text{ for all } i = 1, 2, \ldots, r\} = (\mathsf{P}_\theta\{|T_1| < c\})^r. \tag{4.2.14}$$

Note that $\mathsf{P}_\theta\{|\lambda_1| > 0\} > 0$, whence it follows that there exists a positive number h such that either $\mathsf{P}_\theta\{\lambda_1 > h\} > 0$ or $\mathsf{P}_\theta\{\lambda_1 < -h\} > 0$. If k is greater than c/h, then

$$\begin{aligned}
\mathsf{P}_\theta\{|T_1| \ge c\} &= \mathsf{P}_\theta\{|\lambda_1 + \cdots + \lambda_k| \ge c\} \\
&\ge \mathsf{P}_\theta\left\{\lambda_i \ge \frac{c}{k} \text{ for all } i = 1, 2, \ldots, k\right\} \\
&\quad + \mathsf{P}_\theta\left\{\lambda_i \le -\frac{c}{k} \text{ for all } i = 1, 2, \ldots, k\right\} \\
&\ge (\mathsf{P}_\theta\{\lambda_1 > h\})^k + (\mathsf{P}_\theta\{\lambda_1 < -h\})^k > 0.
\end{aligned}$$

This yields $\mathsf{P}_\theta\{|T_1| < c\} < 1$. In view of (4.2.14) we therefore get

$$\lim_{m \to \infty} \mathsf{P}_\theta\{\nu > m\} = \mathsf{P}_\theta\{\nu = \infty\} = 0,$$

that is, $\mathsf{P}_\theta\{\nu < \infty\} = 1$ and statement a) is proved.

Relation (4.2.14) for $t \ge 0$ implies that

$$\begin{aligned}
\mathsf{E}_\theta e^{t\nu} &= \sum_{n=1}^{\infty} e^{tn} \mathsf{P}_\theta\{\nu = n\} = \sum_{r=1}^{\infty} \sum_{n=(r-1)k+1}^{kr} e^{tn} \mathsf{P}_\theta\{\nu = n\} \\
&\le \sum_{r=1}^{\infty} e^{tkr} \mathsf{P}_\theta\{(r-1)k < \nu \le kr\} \le \sum_{r=1}^{\infty} e^{tkr} \mathsf{P}_\theta\{\nu > (r-1)k\} \\
&\le \sum_{r=1}^{\infty} e^{tkr} \gamma^{r-1}
\end{aligned} \tag{4.2.15}$$

where $\gamma = \mathsf{P}_\theta\{|T_1| < c\}$ and k is such that $\gamma < 1$. It follows from (4.2.15) that $\mathsf{E}_\theta e^{t\nu} < \infty$ for $\gamma e^{tk} < 1$, thus for $t \in [0, t_0)$ where $t_0 = k^{-1} \ln \gamma^{-1}$.

If $t < 0$, then the equality $\mathsf{P}_\theta\{\nu \ge 0\} = 1$ implies that $\mathsf{E}_\theta e^{t\nu} \le e^{t \cdot 0} = 1$ and statement b) is also proved. □

REMARK 4.2.2. Lemma 4.2.3 holds in the case where θ is different from both θ_1 and θ_2 but if the random variables $\xi_1, \xi_2, \ldots$ are independent and identically distributed with respect to the measure P_θ. If the conditions of Lemma 4.2.3 are satisfied for both $\theta = \theta_1$ and $\theta = \theta_2$, then statement a) implies that the Wald sequential test with limit points $(-b, a)$ is terminated during a finite time with probability one.

REMARK 4.2.3. The assumption that a Wald sequential test is terminated during a finite time with probability one ($\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0$, $i = 1, 2$) is not too restrictive. Moreover this assumption is quite natural, since otherwise $\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} = 0$ or $\mathsf{P}_{\theta_i}\{\lambda_1 = 0\} = 1$ for $i = 1, 2$ and the measures P_{θ_1} and P_{θ_2} coincide. The hypotheses H_1 and H_2 are indistinguishable in this case (see Section 1.1), that is, the problem of distinguishing the hypotheses H_1 and H_2 makes no sense.

In what follows we need the following auxiliary result.

LEMMA 4.2.4. *Let ν be the stopping time of a Wald sequential test with limit points $(-b, a)$. If $\mathsf{E}_\theta \nu < \infty$ and $\mathsf{E}_\theta|\lambda_1| < \infty$ for $\theta = \theta_1$ or $\theta = \theta_2$, then*

$$\mathsf{E}_\theta \Lambda_\nu = \mathsf{E}_\theta \nu \mathsf{E}_\theta \lambda_1. \tag{4.2.16}$$

PROOF. Consider random variables $\eta_i = I\{\nu > i-1\}$ where $i = 1, 2, \ldots$. Here $I(A)$ is the indicator of an event A, that is, $\eta_i = 1$ if a decision is not made by observations $\xi_1, \xi_2, \ldots, \xi_{i-1}$. The random variable η_i is a function of $\xi_1, \xi_2, \ldots, \xi_{i-1}$ and does not depend on ξ_i, thus it does not depend on $\lambda_i = \lambda(\xi_i)$. It is clear that

$$\Lambda_\nu = \lambda_1\eta_1 + \lambda_2\eta_2 + \ldots,$$

whence

$$\mathsf{E}_\theta \Lambda_\nu = \mathsf{E}_\theta \sum_{i=1}^{\infty} \lambda_i\eta_i = \sum_{i=1}^{\infty} \mathsf{E}_\theta \lambda_i\eta_i = \sum_{i=1}^{\infty} \mathsf{E}_\theta \lambda_i \mathsf{E}_\theta \eta_i$$

$$= \mathsf{E}_\theta \lambda_1 \sum_{i=1}^{\infty} \mathsf{E}_\theta \eta_i = \mathsf{E}_\theta \lambda_1 \sum_{i=1}^{\infty} \mathsf{P}_\theta\{\nu \geq i\} = \mathsf{E}_\theta \lambda_1 \mathsf{E}_\theta \nu.$$

We interchanged the summation and expectation in the preceding relation, since

$$\sum_{i=1}^{\infty} \mathsf{E}_\theta|\lambda_i\eta_i| = \mathsf{E}_\theta|\lambda_1| \sum_{i=1}^{\infty} \mathsf{P}_\theta\{\nu \geq i\} = \mathsf{E}_\theta|\lambda_1| \mathsf{E}_\theta \nu < \infty.$$

We also used the obvious equality

$$\mathsf{E}_\theta \nu = \sum_{i=1}^{\infty} \mathsf{P}_\theta\{\nu \geq i\}.$$

Thus equality (4.2.16) is proved. □

Equality (4.2.16) is called the *Wald identity*. It also holds in the case where the assumptions of Lemma 4.2.4 are satisfied for some θ different from both θ_1 and θ_2 (see Remark 4.2.2).

The expectation of the stopping time of a Wald sequential test. The following result contains a lower bound of expectations $\mathsf{E}_{\theta_i}\nu$ for $i = 1, 2$.

LEMMA 4.2.5. *Let ν be the stopping time of a Wald sequential test of power (α, β) and with limit points $(-b, a)$. If $\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0$ and $\mathsf{E}_{\theta_i}|\lambda_1| < \infty$ for $i = 1, 2$, then*

$$\mathsf{E}_{\theta_1}\nu \geq -H(\alpha|1-\beta)/\mathsf{E}_{\theta_1}\lambda_1, \tag{4.2.17}$$

$$\mathsf{E}_{\theta_2}\nu \geq H(\beta|1-\alpha)/\mathsf{E}_{\theta_2}\lambda_1 \tag{4.2.18}$$

where

$$H(x|y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y}, \qquad 0 \le x,\ y \le 1, \tag{4.2.19}$$

is a relative entropy of the distribution $(x, 1-x)$ *with respect to the distribution* $(y, 1-y)$.

PROOF. By the assumptions of the lemma $\mathsf{P}_{\theta_i}(\lambda_1 \neq 0) > 0$ for $i = 1, 2$. Thus $\mathsf{E}_{\theta_i}\nu < \infty$ for $i = 1, 2$ according to statement b) of Lemma 4.2.3, whence

$$\mathsf{E}_{\theta_i}\Lambda_\nu = \mathsf{E}_{\theta_i}\nu\mathsf{E}_{\theta_i}\lambda_1, \qquad i = 1, 2, \tag{4.2.20}$$

in view of Lemma 4.2.4. Now we conclude that the Wald test is terminated during a finite time with probability one by Lemma 4.2.3 and Remark 4.2.2, since

$$\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0, \qquad i = 1, 2.$$

This together with (4.2.9) and (4.2.10) implies

$$\mathsf{P}_{\theta_i}\{\Lambda_\nu \ge a\} + \mathsf{P}_{\theta_i}\{\Lambda_\nu \le -b\} = \mathsf{P}_{\theta_i}\{\nu < \infty\} = 1, \qquad i = 1, 2. \tag{4.2.21}$$

Taking into account (4.2.21) for $i = 1$ we derive from (4.2.5) that

$$\begin{aligned} \mathsf{E}_{\theta_1}\Lambda_\nu &= \mathsf{P}_{\theta_1}\{\Lambda_\nu \le -b\}\mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \le -b\} \\ &\quad + \mathsf{P}_{\theta_1}\{\Lambda_\nu \ge a\}\mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \ge a\} \\ &= (1-\alpha)\mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \le -b\} + \alpha\mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \ge a\}. \end{aligned} \tag{4.2.22}$$

The Jensen inequality and definition (4.2.8) of the set V_n imply that

$$\begin{aligned} \mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \le -b\} &\le \ln \mathsf{E}_{\theta_1}\left\{e^{\Lambda_\nu}/\Lambda_\nu \le -b\right\} \\ &= \ln\frac{1}{1-\alpha}\sum_{n=1}^{\infty}\int_{V_n} e^{\Lambda_n(x)}\,\mathsf{P}_{\theta_1}^{(n)}(dx) \\ &= \ln\frac{1}{1-\alpha}\sum_{n=1}^{\infty}\int_{V_n}\mathsf{P}_{\theta_2}^{(n)}(dx) = \ln\frac{\beta}{1-\alpha}. \end{aligned} \tag{4.2.23}$$

Similarly we get

$$\begin{aligned} \mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \ge a\} \le \ln \mathsf{E}_{\theta_1}\left\{e^{\Lambda_\nu}/\Lambda_\nu \ge a\right\} &= \ln\frac{1}{\alpha}\sum_{n=1}^{\infty}\int_{W_n} e^{\Lambda_n(x)}\,\mathsf{P}_{\theta_1}^{(n)}(dx) \\ = \ln\frac{1}{\alpha}\sum_{n=1}^{\infty}\int_{W_n}\mathsf{P}_{\theta_2}^{(n)}(dx) &= \ln\frac{1-\beta}{\alpha}. \end{aligned} \tag{4.2.24}$$

It follows from equality (4.2.22) and inequalities (4.2.23) and (4.2.24) that

$$\mathsf{E}_{\theta_1}\nu \le -H(\alpha|1-\beta) \tag{4.2.25}$$

where $H(x|y)$ is the function defined by (4.2.19).

Since $\mathsf{P}_{\theta_1}(\lambda_1 \neq 0) > 0$, we have $\mathsf{E}_{\theta_1}\lambda_1 < 0$. Indeed, applying the elementary inequality $\ln x \le x - 1$, $x > 0$, we get

$$\begin{aligned} \mathsf{E}_{\theta_1}\lambda_1 = \int \ln\frac{p(x;\theta_2)}{p(x;\theta_1)}\,\mathsf{P}_{\theta_1}(dx) &< \int\left(\frac{p(x;\theta_2)}{p(x;\theta_1)} - 1\right)\mathsf{P}_{\theta_1}(dx) \\ = \int p(x;\theta_2)\,\mu(dx) - \int p(x;\theta_1)\,\mu(dx) &= 0 \end{aligned} \tag{4.2.26}$$

in view of

$$\mathsf{P}_{\theta_1}\{\lambda_1 = 0\} = \mathsf{P}_{\theta_1}\left\{x: \ln\frac{p(x;\theta_2)}{p(x;\theta_1)} = 0\right\} = \mathsf{P}_{\theta_1}\left\{x: \frac{p(x;\theta_2)}{p(x;\theta_1)} = 1\right\} < 1.$$

The lower bound (4.2.17) for $\mathsf{E}_{\theta_1}\nu$ follows from (4.2.20) for $i = 1$ and inequality (4.2.25) by considering the sign of $\mathsf{E}_{\theta_1}\lambda_1$.

Similarly to inequality (4.2.26) we obtain from $\mathsf{P}_{\theta_2}\{\lambda_1 \neq 0\} > 0$ that $\mathsf{E}_{\theta_2}\lambda_1 > 0$. Further we use the Jensen inequality and definition (4.2.8) of the set V_n and obtain

$$\begin{aligned}\mathsf{E}_{\theta_2}\{\Lambda_\nu/\Lambda_\nu \le -b\} &= -\mathsf{E}_{\theta_2}\{-\Lambda_\nu/\Lambda_\nu \le -b\} \ge -\ln \mathsf{E}_{\theta_2}\{e^{-\Lambda_\nu}/\Lambda_\nu \le -b\}\\ &= -\ln\frac{1}{\beta}\sum_{n=1}^{\infty}\int_{V_n} e^{-\Lambda_\nu(x)}\,\mathsf{P}_{\theta_2}^{(n)}(dx)\\ &= -\ln\frac{1}{\beta}\sum_{n=1}^{\infty}\int_{V_n}\mathsf{P}_{\theta_1}^{(n)}(dx) = -\ln\frac{1-\alpha}{\beta} = \ln\frac{\beta}{1-\alpha}.\end{aligned} \tag{4.2.27}$$

Following the same reasoning we get

$$\mathsf{E}_{\theta_2}\{\Lambda_\nu/\Lambda_\nu \ge a\} \ge \ln\frac{1-\beta}{\alpha}. \tag{4.2.28}$$

Now (4.2.22), (4.2.27), and (4.2.28) imply

$$\mathsf{E}_{\theta_2}\Lambda_\nu \ge H(\beta|1-\alpha). \tag{4.2.29}$$

Since $\mathsf{E}_{\theta_2}\lambda_1$ is positive, equality (4.2.20) for $i = 2$ and inequality (4.2.29) imply the lower bound (4.2.18). □

REMARK 4.2.4. The relative entropy of a measure P with respect to a measure $\widetilde{\mathsf{P}}$ is defined in Section 2.3 and is denoted there by $I(\mathsf{P}|\widetilde{\mathsf{P}})$ for arbitrary probability measures P and $\widetilde{\mathsf{P}}$. If P is a measure concentrated at two points with probabilities x and $1-x$, while $\widetilde{\mathsf{P}}$ is a measure concentrated at the same points with probabilities y and $1-y$, then Definition 2.3.1 and equality (4.2.19) imply that $I(\mathsf{P}|\widetilde{\mathsf{P}}) = H(x|y)$.

REMARK 4.2.5. If the assumptions of Lemma 4.2.5 are satisfied, then considering the sign of $\mathsf{E}_{\theta_i}\lambda_1$, $i = 1, 2$, one can rewrite inequalities (4.2.17) and (4.2.18) in the form

$$\mathsf{E}_{\theta_1}\nu \ge H(\alpha|1-\beta)|\mathsf{E}_{\theta_1}\lambda_1|^{-1}, \tag{4.2.30}$$

$$\mathsf{E}_{\theta_2}\nu \ge H(\beta|1-\alpha)|\mathsf{E}_{\theta_2}\lambda_1|^{-1}, \tag{4.2.31}$$

that is, inequality (4.2.31) follows from (4.2.30) if we interchange the hypotheses $H_1\colon \theta = \theta_1$ and $H_2\colon \theta = \theta_2$. The latter means that θ_2 substitutes θ_1, β substitutes α, and α substitutes β. In a similar manner, inequality (4.2.31) follows from (4.2.30).

REMARK 4.2.6. One can evaluate approximate values of the expectations $\mathsf{E}_{\theta_i}\nu$, $i = 1, 2$, as follows. Let $\mathsf{E}_{\theta_i}|\lambda_1| < \infty$ and $\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0$ for $i = 1, 2$. Then $\mathsf{E}_{\theta_i}\nu < \infty$ for $i = 1, 2$ by Lemma 4.2.3 and

$$\mathsf{E}_{\theta_i}\Lambda_\nu = \mathsf{E}_{\theta_i}\nu\mathsf{E}_{\theta_i}\lambda_1, \qquad i = 1, 2, \tag{4.2.32}$$

by Lemma 4.2.4. On the other hand

$$\begin{aligned}\mathsf{E}_{\theta_1}\Lambda_\nu &= \mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \geq a\}\mathsf{P}_{\theta_1}\{\Lambda_\nu \geq a\} + \mathsf{E}_{\theta_1}\{\Lambda_\nu/\Lambda_\nu \leq -b\}\mathsf{P}_{\theta_1}\{\Lambda_\nu \leq -b\} \\ &\approx a\alpha - b(1-\alpha)\end{aligned}$$

where the approximation appears, since we neglect the exits of Λ_ν from the interval $(-b, a)$. This approximation and equality (4.2.32) yield

$$\mathsf{E}_{\theta_1}\nu \approx (a\alpha - b(1-\alpha))/\mathsf{E}_{\theta_1}\lambda_1. \tag{4.2.33}$$

Similarly we obtain

$$\mathsf{E}_{\theta_2}\nu \approx (a(1-\beta) - b\beta)/\mathsf{E}_{\theta_2}\lambda_1. \tag{4.2.34}$$

If

$$a = \ln\frac{1-\beta}{\alpha} \quad \text{and} \quad b = \ln\frac{1-\alpha}{\beta},$$

then approximations (4.2.33) and (4.2.34) for $\mathsf{E}_{\theta_1}\nu$ and $\mathsf{E}_{\theta_2}\nu$ coincide with the lower bounds in (4.2.17) and (4.2.18), respectively, in the case of a Wald sequential test with limit points $(-b, a)$.

EXAMPLE 4.2.1. Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables whose distribution under the hypothesis H_i is $\mathcal{N}(\theta_i, \sigma^2)$ normal, $i = 1, 2$, where $\theta_1 < \theta_2$ and the variance σ^2 is known. Then

$$\Lambda_n = \frac{\theta_2 - \theta_1}{2\sigma^2}\left(2\sum_{i=1}^{n}\xi_i - n(\theta_2 + \theta_1)\right).$$

It is clear that $\mathsf{P}_{\theta_i}(\lambda_1 = 0) = 0$ for $i = 1, 2$. Let ν be the stopping time of a Wald sequential test of power (α, β) with limit points $(-b, a)$. Since

$$\mathsf{E}_{\theta_1}\lambda_1 = -\frac{(\theta_2 - \theta_1)^2}{2\sigma^2} \quad \text{and} \quad \mathsf{E}_{\theta_2}\lambda_1 = \frac{(\theta_2 - \theta_1)^2}{2\sigma^2},$$

approximations (4.2.33) and (4.2.34) become of the form

$$\mathsf{E}_{\theta_1}\nu \approx \frac{2(b(1-\alpha) - a\alpha)\sigma^2}{(\theta_2 - \theta_1)^2}, \tag{4.2.35}$$

$$\mathsf{E}_{\theta_2}\nu \approx \frac{2(a(1-\beta) - b\beta)\sigma^2}{(\theta_2 - \theta_1)^2}. \tag{4.2.36}$$

Consider a test with a fixed nonrandom size n of the sample and whose power is (α, β). For example, the Neyman–Pearson test $\delta_n^{+,\alpha}$ of level α satisfies these conditions. Then

$$\alpha = \mathsf{P}_{\theta_1}\{\Lambda_n > c_\alpha\} = \mathsf{P}_{\theta_1}\left\{\sum_{i=1}^{n}(\xi_i - \theta_1) > z_{1-\alpha}\sigma\sqrt{n}\right\}$$

where z_p is a p-quantile of the $\mathcal{N}(0,1)$ law, that is, $\Phi(z_p) = p$. The test $\delta_n^{+,\alpha}$ has the type II error probability β if

$$\beta = \mathsf{P}_{\theta_2}\left\{\sum_{i=1}^{n}(\xi_i - \theta_1) < z_{1-\alpha}\sigma\sqrt{n}\right\}$$
$$= \mathsf{P}_{\theta_2}\left\{\sum_{i=1}^{n}(\xi_i - \theta_2) < z_{1-\alpha}\sigma\sqrt{n} - n(\theta_2 - \theta_1)\right\}$$

or, equivalently, if

$$z_{1-\alpha}\sigma - \sqrt{n}(\theta_2 - \theta_1) = z_\beta\sigma.$$

Since $z_{1-\alpha} = -z_\alpha$, we have

$$n = \frac{\sigma^2(z_\alpha + z_\beta)^2}{(\theta_2 - \theta_1)^2}. \tag{4.2.37}$$

Relations (4.2.35)–(4.2.37) provide the following approximations:

$$\frac{\mathsf{E}_{\theta_1}\nu}{n} \approx \frac{2(b(1-\alpha) - a\alpha)}{(z_\alpha + z_\beta)^2}, \qquad \frac{\mathsf{E}_{\theta_2}\nu}{n} \approx \frac{2(a(1-\beta) - b\beta)}{(z_\alpha + z_\beta)^2}.$$

If $\alpha = \beta = 0.05$, then $z_\alpha = z_\beta \approx 1.6449$ and $a = -b \approx 2.9444$, whence

$$\frac{\mathsf{E}_{\theta_i}\nu}{n} \approx 0.4897, \qquad i = 1,2. \tag{4.2.38}$$

It is seen from equalities (4.2.38) that the above Wald sequential test of power $(0.05; 0.05)$ requires two times less observations than the Neyman–Pearson test of the same power $(0.05; 0.05)$ and with a nonrandom size of the sample.

The fundamental identity of sequential analysis. First we prove two auxiliary results.

LEMMA 4.2.6. *Let ζ be a random variable defined on the main probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and such that:*

a) $\mathsf{P}\{\zeta > 0\} > 0$ *and* $\mathsf{P}\{\zeta < 0\} > 0$;

b) $\varphi(t) = \mathsf{E}e^{t\zeta}$ *exists for all* $t \in (-\infty, \infty)$;

c) $\mathsf{E}\zeta \neq 0$.

Then there is a unique number $\tau \neq 0$ such that $\varphi(\tau) = 1$, and moreover $\tau < 0$ if $\mathsf{E}\zeta > 0$ and $\tau > 0$ if $\mathsf{E}\zeta < 0$.

PROOF. The condition $\mathsf{P}\{\zeta > 0\} > 0$ implies that there is a constant $c > 0$ such that $\mathsf{P}\{\zeta > c\} > 0$. Thus for all $t > 0$

$$\varphi(t) = \mathsf{E}e^{t\zeta} \geq \mathsf{E}I(\zeta > c)e^{t\zeta} > e^{tc}\mathsf{P}\{\zeta > c\},$$

whence $\varphi(t) \to \infty$ as $t \to \infty$. Similarly, the condition $\mathsf{P}\{\zeta < 0\} > 0$ implies that $\varphi(t) \to \infty$ as $t \to -\infty$. Moreover, $\varphi(0) = 1$ and $\varphi'(0) = \mathsf{E}\zeta \neq 0$. If $\mathsf{E}\zeta > 0$, then $\varphi'(0) > 0$ and thus there is $\tau < 0$ such that $\varphi(\tau) = 1$. Similarly, if $\mathsf{E}\zeta < 0$, then $\varphi'(0) < 0$ and thus there is $\tau > 1$ such that $\varphi(\tau) = 1$. It is easy to show that $\varphi''(t) = \mathsf{E}\zeta^2 e^{t\zeta} > 0$, hence the function $\varphi(t)$ is strictly convex. The latter property means, in particular, that φ has a unique minimum. Therefore the solution τ is unique. □

LEMMA 4.2.7. *Let ν be the stopping time of a Wald sequential test with limit points $(-b, a)$. Assume that $\mathsf{P}_\theta\{\nu < \infty\} = 1$ where either $\theta = \theta_1$ or $\theta = \theta_2$. Then*

$$\mathsf{E}_\theta e^{t\Lambda_\nu}(\varphi(t))^{-\nu} = \mathsf{P}\{\nu < \infty / H_t\} \tag{4.2.39}$$

for t such that $\varphi(t) = \mathsf{E}_\theta e^{t\lambda_1} < \infty$ where $\mathsf{P}\{\cdot / H_t\}$ is the conditional probability given H_t where the hypothesis H_t is that the random variables $\xi_1, \xi_2, \ldots$ are independent, identically distributed, and whose density is

$$p(x/H_t) = \frac{e^{t\lambda(x)}}{\varphi(t)} p_\theta(x). \tag{4.2.40}$$

PROOF. Taking into account $\mathsf{P}_\theta\{\nu < \infty\} = 1$ we conclude that

$$\begin{aligned}\mathsf{E}_\theta e^{t\Lambda_\nu}(\varphi(t))^{-\nu} &= \sum_{n=1}^{\infty} \int_{V_n \cup W_n} (\varphi(t))^{-n} \exp\left\{t\Lambda_n\left(x^{(n)}\right)\right\} \prod_{i=1}^{n} p_\theta(x_i)\, \mu^n\left(dx^{(n)}\right) \\ &= \sum_{n=1}^{\infty} \int_{V_n \cup W_n} \prod_{i=1}^{n} p_\theta(x_i/H_t)\, \mu^n\left(dx^{(n)}\right) = \mathsf{P}\{\nu < \infty / H_t\}\end{aligned}$$

where V_n and W_n are the sets defined by (4.2.7) and (4.2.8), respectively, while $p(x/H_t)$ is the density of the distribution given by (4.2.40). Thus equality (4.2.39) is proved. □

LEMMA 4.2.8. *Let ν be the stopping time of a Wald sequential test with limit points $(-b, a)$. Assume that $\mathsf{P}_\theta\{\lambda_1 \neq 0\} > 0$ where either $\theta = \theta_1$ or $\theta = \theta_2$. Then*

$$\mathsf{E}_\theta e^{t\Lambda_\nu}(\varphi(t))^{-\nu} = 1 \tag{4.2.41}$$

for t such that $\varphi(t) = \mathsf{E}_\theta e^{t\lambda_1} < \infty$.

PROOF. The condition $\mathsf{P}_\theta\{\lambda_1 \neq 0\} > 0$ implies that $\mathsf{P}_\theta\{\nu < \infty\} = 1$ by Lemma 4.2.3 and that $\mathsf{P}_\theta\{\lambda(\xi_1) \neq 0/H_t\} > 0$, since

$$\mathsf{P}_\theta\{\lambda(\xi_1) \neq 0/H_t\} = \frac{1}{\varphi(t)} \int_{\{\lambda(x) \neq 0\}} e^{t\lambda(x)}\, \mathsf{P}_\theta(dx) \neq 0.$$

Applying Lemma 4.2.3 once more we get

$$\mathsf{P}_\theta\{\nu < \infty / H_t\} = 1.$$

Now equality (4.2.41) follows from (4.2.39). □

Equality (4.2.41) is called the *fundamental identity of sequential analysis.*

REMARK 4.2.7. Equalities (4.2.39) and (4.2.41) hold for t such that $\varphi(t) < \infty$. If $\mathsf{P}_\theta\{\lambda_1 \neq 0\} > 0$, then it follows from Lemma 4.2.3 that there is $t_0 > 0$ such that $\varphi(t) < \infty$ for all $t < t_0$.

REMARK 4.2.8. Lemmas 4.2.7 and 4.2.8 hold for measures P_θ if θ is different from both θ_1 and θ_2.

The fundamental identity of sequential analysis can be applied to the problem of distinguishing composite hypotheses.

Let $\xi_1, \xi_2, \dots$ be independent identically distributed random variables with a distribution P_θ depending on an unknown parameter $\theta \in \Theta$. Assume that P_θ is absolutely continuous with respect to some σ-finite measure μ and the density is $p(x;\theta)$. Let $\Theta = \Theta_1 \cup \Theta_2$ and $\Theta_1 \cap \Theta_2 = \varnothing$. Consider the problem of distinguishing the hypotheses $H_1\colon \theta \in \Theta_1$ and $H_2\colon \theta \in \Theta_2$ by observations $\xi_1, \xi_2, \dots$ with the help of a Wald sequential test that distinguishes two simple hypotheses $H_1'\colon \theta = \theta_1$ and $H_2'\colon \theta = \theta_2$ where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ are some fixed points. Let the random variable $\zeta = \lambda_1$ satisfy the assumptions of Lemma 4.2.6 concerning the measure P_θ and let $\theta \in \Theta$ be an arbitrary fixed point. According to Lemma 4.2.6, there is a number $\tau(\theta) \neq 0$ such that $\varphi_\theta(\tau(\theta)) = 1$ where $\varphi_\theta(t) = \mathsf{E}_\theta e^{t\lambda_1}$. Further

$$\mathsf{E}_\theta e^{\tau(\theta)\Lambda_\nu} = 1 \tag{4.2.42}$$

by Lemma 4.2.8 and Remark 4.2.8.

The function $\beta(\theta) = \mathsf{P}_\theta\{\Lambda_\nu \leq -b\}$ is the probability to accept the hypothesis H_1 if the parameter is θ, that is, $\beta(\theta)$ is the power function of the Wald test for distinguishing the hypotheses H_1 and H_2. The function $\beta(\theta)$ is also called the *operating characteristic of the test* in sequential analysis.

Put

$$\mathsf{E}_\theta^* = \mathsf{E}_\theta\{e^{\tau(\theta)\Lambda_\nu}/\Lambda_\nu \leq -b\}, \qquad \mathsf{E}_\theta^{**} = \mathsf{E}_\theta\{e^{\tau(\theta)\Lambda_\nu}/\Lambda_\nu \geq a\}. \tag{4.2.43}$$

It follows from equalities (4.2.42), (4.2.43) and (4.2.21) for $\theta_i = \theta$ that

$$1 = \mathsf{E}_\theta e^{\tau(\theta)\Lambda_\nu} = \beta(\theta)\mathsf{E}_\theta^* + (1 - \beta(\theta))\mathsf{E}_\theta^{**}. \tag{4.2.44}$$

Using the approximations

$$\mathsf{E}_\theta^* \approx e^{-b\tau(\theta)}, \qquad \mathsf{E}_\theta^{**} \approx e^{a\tau(\theta)}$$

we derive from (4.2.44) an approximation for the operating characteristic:

$$\beta(\theta) \approx \frac{1 - e^{a\tau(\theta)}}{e^{-b\tau(\theta)} - e^{a\tau(\theta)}}, \qquad \theta \in \Theta.$$

Similarly one can obtain approximations for $\mathsf{E}_\theta\nu$, $\theta \in \Theta$.

More details on sequential Wald tests and their properties as well as on the other sequential tests can be found in [**15, 51, 54**]. The sequential analysis is described in [**13, 46**].

4.3. The optimality of a sequential Wald test

The main theorem. As in the preceding section we consider the problem of distinguishing two simple hypotheses $H_1\colon \theta = \theta_1$ and $H_2\colon \theta = \theta_2$ by observations $\xi_1, \xi_2, \dots$ where $\xi_1, \xi_2, \dots$ are independent identically distributed random variables whose distribution P_θ depends on an unknown parameter θ. Moreover we assume that their distribution possesses the density $p(x;\theta)$ with respect to a σ-finite measure μ. Throughout this section we also assume that $\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0$ for $i = 1, 2$.

Generally speaking, we consider a sequential test for distinguishing the hypotheses H_1 and H_2 by observations $\xi^{(\nu)} = (\xi_1, \xi_2, \dots, \xi_\nu)$ where ν is a *stopping*

time. A stopping time can be either random or deterministic. Sequential tests are also called *sequential decision procedures*. *Decision functions* $d_\nu = d_\nu(\xi^{(\nu)})$ are defined as follows: if $\nu = n$, then $d_n(x^{(n)})$ assumes only two values d_1 and d_2. If $d_n(x^{(n)}) = d_1$, then the hypothesis H_1 is accepted, while H_2 is accepted if $d_n(x^{(n)}) = d_2$. We define the type I and type II *error probabilities* for a sequential test with a decision function d_ν by

$$\alpha_1(d_\nu) = \mathsf{P}_{\theta_1}\{d_\nu = d_2\}, \qquad \alpha_2(d_\nu) = \mathsf{P}_{\theta_2}\{d_\nu = d_1\}. \tag{4.3.1}$$

Note that

$$\alpha_1 = \alpha_1(d_\nu^*) = \mathsf{P}_{\theta_1}\{d_\nu^* = d_2\} = \mathsf{P}_{\theta_1}\{\Lambda_\nu \geq a\}, \tag{4.3.2}$$

$$\alpha_2 = \alpha_2(d_\nu^*) = \mathsf{P}_{\theta_2}\{d_\nu^* = d_1\} = \mathsf{P}_{\theta_2}\{\Lambda_\nu \leq -b\} \tag{4.3.3}$$

for a Wald test of power (α_1, α_2) with limit points $(-b, a)$ and decision function d_ν^*.

The following result asserts that a sequential Wald test is optimal in the sense that $\mathsf{E}_{\theta_1}\nu$ and $\mathsf{E}_{\theta_2}\nu$ are minimal for it.

THEOREM 4.3.1. *The sequential Wald test of power* (α_1, α_2) *minimizes both expectations* $\mathsf{E}_{\theta_1}\nu$ *and* $\mathsf{E}_{\theta_2}\nu$ *in the set of all tests (including nonsequential tests) such that* $\mathsf{E}_{\theta_1}\nu$ *and* $\mathsf{E}_{\theta_2}\nu$ *are finite and*

$$\mathsf{P}_{\theta_1}\{d_\nu = d_2\} \leq \alpha_1, \qquad \mathsf{P}_{\theta_2}\{d_\nu = d_1\} \leq \alpha_2. \tag{4.3.4}$$

To prove Theorem 4.3.1 we consider an auxiliary Bayes problem and use it to show that the Wald test is optimal.

An auxiliary problem. Consider the following sequential Bayes problem for distinguishing the hypotheses $H_1\colon \theta = \theta_1$ and $H_2\colon \theta = \theta_2$. Let $w_i > 0$ be the loss caused by a wrong decision given the hypothesis H_i is true and let the loss caused by a correct decision be zero. Assume that the cost of every observation is $c > 0$. The risk of the sequential test δ when making a decision given the hypothesis H_i is true equals

$$\alpha_i w_i + c\mathsf{E}_{\theta_i}\nu, \qquad i = 1, 2,$$

where ν is the stopping time of the sequential test δ and α_1 and α_2 are the type I and type II error probabilities, respectively. The risk of a test includes the mean loss caused by making a decision and the mean cost per observation. Let $q = \mathsf{P}\{\theta = \theta_1\}$ and $1 - q = \mathsf{P}\{\theta = \theta_2\}$ be a priori probabilities of the hypotheses H_1 and H_2. Then the (unconditional) risk of the test δ is

$$r(q, \delta) = q(\alpha_1 w_1 + c\mathsf{E}_{\theta_1}\nu) + (1 - q)(\alpha_2 w_2 + c\mathsf{E}_{\theta_2}\nu). \tag{4.3.5}$$

DEFINITION 4.3.1. A sequential test δ^* is called *q-Bayes* if $r(q, \delta^*) \leq r(q, \delta)$ for all tests δ where $q \in [0, 1]$ is given and $r(q, \delta)$ is the risk of a test δ defined by (4.3.5).

DEFINITION 4.3.2. A sequential test δ^* is called *Bayes* if $r(q, \delta^*) \leq r(q, \delta)$ for all tests δ and all $q \in [0, 1]$.

The q-Bayes test for the above auxiliary problem is described in the following result.

LEMMA 4.3.1. *Let $q' \le q''$ be solutions of the equations*

$$r(q', \delta_1) = \rho(q'), \qquad r(q'', \delta_2) = \rho(q''), \tag{4.3.6}$$

respectively (provided the solutions exist), where

$$\rho(q) = \inf_{\delta \in \mathscr{S}_1} r(q, \delta), \tag{4.3.7}$$

$\mathscr{S}_1$ is the class of tests for which it is necessary to take at least one observation, and δ_i is the test rejecting the hypothesis H_i without sampling. Put

$$q' = q'' = \frac{w_2}{w_1 + w_2} \tag{4.3.8}$$

for the case where equations (4.3.6) have no solutions. If $0 < q' < q'' < 1$, then for all $q \in (q', q'')$ the Wald sequential test with limit points

$$-b = \ln\left(\frac{q}{1-q}\,\frac{1-q''}{q''}\right), \qquad a = \ln\left(\frac{q}{1-q}\,\frac{1-q'}{q'}\right) \tag{4.3.9}$$

is q-Bayes.

PROOF. *Step* I. First we find q for which the better decision is the one made without sampling. We get from (4.3.5) that

$$r(q, \delta_1) = qw_1, \qquad r(q, \delta_2) = (1-q)w_2.$$

Further, (4.3.5) and (4.3.7) for all $\lambda \in (0,1)$ and all $q_1, q_2 \in [0,1]$ imply that

$$\rho(\lambda q_1 + (1-\lambda)q_2) = \inf_{\delta \in \mathscr{S}_1} [\lambda r(q_1, \delta) + (1-\lambda) r(q_2, \delta)] \ge \lambda \rho(q_1) + (1-\lambda)\rho(q_2),$$

that is, $\rho(q)$ is a convex function. Since $\rho(q) \ge 0$, the function $\rho(q)$ is continuous in the interval $(0,1)$.

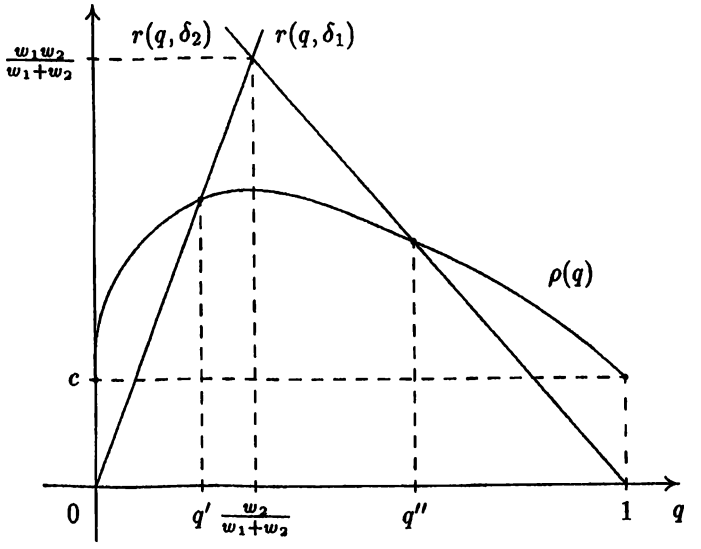

FIGURE 4.3.1. Graphs of the functions $\rho(q)$, $r(q, \delta_1)$, and $r(q, \delta_2)$

The graphs of the functions $\rho(q)$, $r(q,\delta_1)$, and $r(q,\delta_2)$ are shown in Figure 4.3.1. If

$$\rho\left(\frac{w_1w_2}{w_1+w_2}\right) < \frac{w_1w_2}{w_1+w_2} = r\left(\frac{w_2}{w_1+w_2},\delta_1\right) = r\left(\frac{w_2}{w_1+w_2},\delta_2\right), \tag{4.3.10}$$

then solutions q' and q'' of equations (4.3.6) exist. Otherwise we accept convention (4.3.8). Assume that relation (4.3.10) holds. Then $0 < q' < q'' < 1$ and the test δ_1 minimizes $r(q,\delta)$ if and only if $q \le q'$, while δ_2 minimizes $r(q,\delta)$ if and only if $q' \ge q''$. This implies that the unique optimal decision on the first step is as follows: if $q \le q'$, then the hypothesis H_1 is rejected and the hypothesis H_2 is accepted without sampling; if $q \ge q''$, then the hypothesis H_2 is rejected and the hypothesis H_1 is accepted without sampling; if $q' < q < q''$, then it is necessary to take the first observation ξ_1.

Step II. We use induction to complete the proof. Let $q' < q < q''$ and n observations $\xi_1 = x_1$, $\xi_2 = x_2$, ..., $\xi_n = x_n$ be given. Then the procedure is the same as that described in Step I above. The loss nc caused by making n observations does not change the problem, since further observations cannot reimburse this loss. If the probability that the hypothesis H_1 is true does not exceed q' or is not less than q'', then we terminate the sampling; otherwise it is necessary to take one more observation ξ_{n+1}. According to the Bayes formula the probability that the hypothesis H_1 is true given $\xi_1 = x_1$, $\xi_2 = x_2$, ..., $\xi_n = x_n$ equals

$$q(x^{(n)}) = \frac{qp_n(x^{(n)};\theta_1)}{qp_n(x^{(n)}\theta_1) + (1-q)p_n(x^{(n)};\theta_2)}$$

where $x^{(n)} = (x_1, x_2, \ldots, x_n)$ and $p_n(x^{(n)};\theta)$ is the density of the vector

$$\xi^{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)$$

with respect to the measure μ^n. Thus we keep sampling if $q' < q(x^{(n)}) < q''$, that is, if

$$e^{-b} < z_n(x^{(n)}) = \frac{p_n(x^{(n)};\theta_2)}{p_n(x^{(n)};\theta_1)} < e^a$$

where b and a are the constants defined by (4.3.9). If $z_n(x^{(n)}) < e^{-b}$, then the hypothesis H_1 is accepted, while if $z_n(x^{(n)}) > e^a$, then the hypothesis H_2 is accepted. Thus we proved for $q' < q < q''$ that the q-Bayes test coincides with the Wald sequential test with limit points (4.3.9). □

REMARK 4.3.1. In Step I of the proof of Lemma 4.3.1 we determined the q-Bayes procedure (now we denote it by δ^*) as follows: $\delta^* = \delta_1$ if $q < q'$, $\delta^* = \delta_2$ if $q > q''$, and δ^* requires the first observation if $q' < q < q''$. The test δ_1 minimizes risk (4.3.5) if $q = q'$. However δ_1 is not a unique optimal test, since there exists $\delta \in \mathscr{S}_1$ such that $r(q',\delta) = \rho(q')$. If $q = q'$ and it is necessary to take an observation ξ_1, then we showed in Step II of the proof of Lemma 4.3.1 that there is a test in $\mathscr{S}_1$ that minimizes the risk. This means that it makes no difference how one constructs the test in the case of $q = q'$. The same is true for the case of $q = q''$. Moreover this also is true for the next steps. This therefore proves that if $q' \le q \le q''$, then the test coinciding with the Wald sequential test with limit points (4.3.9) is q-Bayes.

A relationship between the auxiliary and main problems is established in the following result.

LEMMA 4.3.2. *For all* $0 < q_0' < q_0'' < 1$ *there are numbers* $w \in (0,1)$ *and* $c > 0$ *such that the Bayes solution of the auxiliary problem with* $w_1 = 1-w$, $w_2 = w$, *and with a priori probability* $q \in (q_0', q_0'')$ *is the Wald sequential test with limit points* $(-b_0, a_0)$ *such that*

$$-b_0 = \ln\left(\frac{q}{1-q}\frac{1-q_0''}{q_0''}\right), \qquad a_0 = \ln\left(\frac{q}{1-q}\frac{1-q_0'}{q_0'}\right).$$

PROOF. *Step* I. Let $q'(w,c)$ and $q''(w,c)$ be solutions of equations (4.3.6) where $r(q,\delta)$ is defined by (4.3.5) for $w_1 = 1-w$ and $w_2 = w$. Thus we need to find w and c such that $q'(w,c) = q_0'$ and $q_0''(w,c) = q_0''$. Given a fixed w let

$$q'(c) = q'(w,c), \qquad q''(c) = q''(w,c).$$

Let $c_0 = c_0(w)$ be the minimal number c such that $q'(c) = q''(c)$. Then $q'(c)$ and $q''(c)$ for $0 < c < c_0$ are defined from the equations

$$(1-w)q' = \rho(q',c), \qquad w(1-q'') = \rho(q'',c)$$

where $\rho(q,c)$ stands for $\rho(q)$ defined in (4.3.7).

Given a fixed q the function $\rho(q,c)$ of the argument c is such that

1) $\rho(q,c)$ is continuous with respect to c;
2) $\rho(q,c)$ increases with respect to c, since for any $\delta \in \mathscr{S}_1$ the risk increases with respect to c and the minimal risk $\rho(q,c)$ is attained for the test $\delta \in \mathscr{S}_1$;
3) $\rho(q,c) \to 0$ as $c \to 0$.

The latter property holds, since the type I and type II error probabilities for samples with fixed size n can be arbitrarily small if n is sufficiently large.

The above properties of the function ρ imply that for $0 < c < c_0$ the function $q'(c)$ (respectively, $q''(c)$) is continuous, increasing (respectively, decreasing), and $q'(c) \to 0$ (respectively, $q''(c) \to 1$) as $c \to 0$. On the other hand,

$$q''(c) - q'(c) \to 0 \quad \text{as } c \to c_0,$$

so that both functions $q'(c)$ and $q''(c)$ approach the solution $q'(c) = q''(c) = w$ of the equation $q'(1-w) = (1-q')w$. The above properties also imply that for fixed w the function

$$\lambda(c) = \frac{q'(c)}{1-q'(c)} \cdot \frac{1-q''(c)}{q''(c)}$$

is continuous, increasing, and varying from 0 to 1 as c is varying from 0 to

$$c_0 = c_0(w).$$

Step II. Put

$$\lambda(w,c) = \frac{q'(w,c)}{1-q'(w,c)} \cdot \frac{1-q''(w,c)}{q''(w,c)}, \qquad \gamma(w,c) = \frac{q_0''(w,c)}{1-q_0''(w,c)}.$$

We prove that there are w and c such that

$$\lambda(w,c) = \frac{q_0'}{1-q_0'} \cdot \frac{1-q_0''}{q_0''} = \lambda_0, \qquad \gamma(w,c) = \frac{q_0''}{1-q_0''} = \gamma_0.$$

We proved in Step I that for all fixed w there exists a unique $c = c(w)$ such that $\lambda(w, c) = \lambda_0$. In Step III below we prove that the function $\gamma(w) = \gamma(w, c(w))$ is a one-to-one correspondence between $w \in (0,1)$ and $\gamma \in (0, \infty)$. Therefore there exists a unique number $w \in (0,1)$ such that $\gamma(w) = \gamma_0$. This will complete the proof of the lemma.

Step III. According to Lemma 4.3.1 for the auxiliary problem with $w_1 = 1 - w$ and $w_2 = w$, the cost per observation $c = c(w)$, and the a priori probability $q = q'(w, c(w))$, there exists a sequential q-Bayes test δ' which is a Wald sequential test with limit points $(-b', 0)$ where

$$-b' = \ln\left(\frac{q'(w,c(w))}{1-q'(w,c(w))} \cdot \frac{1-q''(w,c(w))}{q''(w,c(w))}\right) = \ln \lambda(w, c(w)) = \ln \lambda_0.$$

Further let δ'' be the Wald sequential test for the auxiliary problem with constants $w_1 = 1 - w$, $w_2 = w$, $c = c(w)$, and $q = q''(w, c(w))$ that is a Wald sequential test with limit points $(0, a'')$ where

$$a'' = \ln\left(\frac{q''(w,c(w))}{1-q''(w,c(w))} \cdot \frac{1-q'(w,c(w))}{q'(w,c(w))}\right) = \ln \frac{1}{\lambda_0}.$$

Then the error probabilities α_1' and α_2' and the expectations $\mathsf{E}_{\theta_1}\nu'$ and $\mathsf{E}_{\theta_2}\nu'$ of the test δ' as well as error probabilities α_1'' and α_2'' and expectations $\mathsf{E}_{\theta_1}\nu''$ and $\mathsf{E}_{\theta_2}\nu''$ of the test δ'' depend on w and c through λ_0 but not through γ. Thus they are fixed numbers for a fixed λ_0. The Bayes risks for $q' = q'(w, c(w))$ and $q'' = q''(w, c(w))$ are equal to

$$\rho(q') = r(q', \delta'), \qquad \rho = (q'')r(q'', \delta''),$$

respectively. Relations (4.3.6) imply that

$$r(q', \delta_1) = r(q', \delta'), \qquad r(q'', \delta_2) = r(q'', \delta'').$$

The latter equalities can be rewritten as

$$q'(1-w) = q'[\alpha_1'(1-w) + c\mathsf{E}_{\theta_1}\nu'] + (1-q')[\alpha_2' w + c\mathsf{E}_{\theta_2}\nu'],$$
$$(1-q'')w = q''[\alpha_1''(1-w) + c\mathsf{E}_{\theta_1}\nu''] + (1-q'')[\alpha_2'' w + c\mathsf{E}_{\theta_2}\nu''].$$

Using

$$\frac{q'}{1-q'} = \lambda_0\gamma, \qquad \frac{q''}{1-q''} = \gamma$$

in the latter equalities and excluding c we obtain

$$\{\lambda_0\gamma(1-\alpha_1') - w[\lambda_0\gamma(1-\alpha_1') + \alpha_2']\}(\gamma\mathsf{E}_{\theta_1}\nu'' + \mathsf{E}_{\theta_2}\nu'')$$
$$= \{-\gamma\alpha_1'' + w[(1-\alpha_2'') + \gamma\alpha_1'']\}(\lambda_0\gamma\mathsf{E}_{\theta_1}\nu' + \mathsf{E}_{\theta_2}\nu').$$

This equation is linear with respect to w, thus it has a solution $w \in (0,1)$ for all $\gamma > 0$. Collecting all the terms on one side of this equality we obtain a polynomial of the second degree with respect to γ such that the coefficient of γ^2 and constant term have different signs if $w \in (0,1)$. Thus there exists a unique positive solution γ which is the desired one-to-one correspondence between γ and w. □

REMARK 4.3.2. Property 3) of the function $\rho(q, c)$ mentioned in Step I of the proof of Lemma 4.3.2 follows from the following reasoning. By assumption

$$\mathsf{P}_{\theta_i}\{\lambda_1 \neq 0\} > 0$$

for $i = 1$ and $i = 2$. Then $\mathsf{E}_{\theta_1}\lambda_1 < 0$ and $\mathsf{E}_{\theta_2}\lambda_1 > 0$ (see (4.2.26)). If $\mathsf{E}_{\theta_1}\lambda_1 > -\infty$, then the Khinchine law of large numbers implies that $n^{-1}\Lambda_n \to \mathsf{E}_{\theta_1}\lambda_1$ as $n \to \infty$ in probability P_{θ_1}. Further if $\mathsf{E}_{\theta_1}\lambda_1 = -\infty$, then one can prove that $n^{-1}\Lambda_n \to -\infty$ as $n \to \infty$ in probability P_{θ_1}. Thus $\Lambda_n \to -\infty$ as $n \to \infty$ in probability P_{θ_1}. Then the type I error probabilities $\alpha_1(\delta_n^q)$ approach zero as $n \to \infty$ for the Bayes test δ_n^q constructed from a sample of size n by inequalities (2.3.73). Similarly we obtain that $\alpha_2(\delta_n^q) \to 0$ as $n \to \infty$. Thus $\rho(q, c) \to 0$ as $c \to 0$.

Proof of the main theorem. Now we use Lemmas 4.3.1 and 4.3.2 to prove the main result that the Wald test is optimal.

PROOF OF THEOREM 4.3.1. Consider the Wald sequential test of power (α_1, α_2) with limit points $(-b, a)$ where $a > 0$ and $b > 0$. Let ν be the stopping time of this test. Consider an arbitrary number q of the interval $(0, 1)$ and put

$$q' = \frac{q}{e^a(1-q)+q}, \qquad q'' = \frac{q}{e^{-b}(1-q)+q}.$$

The numbers q' and q'' satisfy (4.3.9) and moreover $0 < q' < q < q'' < 1$. According to Lemma 4.3.2 there are numbers $w \in (0, 1)$ and $c > 0$ such that this test is a Bayes solution of the auxiliary problem for which a priori probabilities of the hypotheses H_1 and H_2 are q and $1-q$, the loss due to a wrong decision is $w_1 = 1-w$ and $w_2 = w$, respectively, and the cost per observation is c. Consider an arbitrary test δ^* (not necessarily sequential) with error probabilities α_1^* and α_2^* and the stopping time ν^* where $\alpha_i^* \leq \alpha_i$ and $\mathsf{E}_{\theta_i}\nu^* < \infty$ for $i = 1, 2$. Again by Lemma 4.3.2

$$\begin{aligned} &q[(1-w)\alpha_1 + c\mathsf{E}_{\theta_1}\nu] + (1-q)[w\alpha_2 + c\mathsf{E}_{\theta_2}\nu] \\ &\quad \leq q[(1-w)\alpha_1^* + c\mathsf{E}_{\theta_1}\nu^*] + (1-q)[w\alpha_2^* + c\mathsf{E}_{\theta_2}\nu^*] \\ &\quad \leq q[(1-w)\alpha_1 + c\mathsf{E}_{\theta_1}\nu^*] + (1-q)[w\alpha_2 + c\mathsf{E}_{\theta_2}\nu^*] \end{aligned} \tag{4.3.11}$$

where the latter inequality holds, since $\alpha_1^* \leq \alpha_1$ and $\alpha_2^* \leq \alpha_2$ by condition. Then inequalities (4.3.11) imply

$$q\mathsf{E}_{\theta_1}\nu + (1-q)\mathsf{E}_{\theta_2}\nu \leq q\mathsf{E}_{\theta_1}\nu^* + (1-q)\mathsf{E}_{\theta_2}\nu^*. \tag{4.3.12}$$

Since (4.3.12) holds for all $q \in (0, 1)$, we pass to the limit as $q \to 0$ and obtain from (4.3.12) that $\mathsf{E}_{\theta_2}\nu \leq \mathsf{E}_{\theta_2}\nu^*$. Similarly we pass to the limit as $q \to 1$ and obtain from (4.3.12) that $\mathsf{E}_{\theta_1}\nu \leq \mathsf{E}_{\theta_1}\nu^*$. □

REMARK 4.3.3. In the proof above we constructed a q-Bayes sequential test for distinguishing the hypotheses H_1 and H_2 with a priori distribution $(q, 1-q)$ of the hypotheses and for the loss matrix

$$A = \begin{pmatrix} 0 & w_1 \\ w_2 & 0 \end{pmatrix}, \qquad w_1 > 0,\ w_2 > 0.$$

The general case of the problem of constructing the Wald sequential tests is reduced to the solution of the Bellman equation (also known as the optimality equation in dynamic programming) (see Section 4.1 and [**13, 15, 46**]).

References to Part 2

1. R. R. Bahadur, *Some limit theorems in statistics*, SIAM, Philadelphia, PA, 1971.
2. J.-R. Barra, *Notions fondamentales de statistique mathématique*, Dunod, Paris, 1971; English transl., *Mathematical basis of statistics*, Academic Press, New York–London, 1981.
3. P. Billingsley, *Convergence of probability measures*, Wiley, New York–London–Sydney, 1968.
4. D. Blackwell and M. A. Girshick, *Theory of games and statistical decisions*, Wiley and Chapman and Hall, New York and London, 1954.
5. M. V. Boldin, G. I. Simonova, and Yu. N. Tyurin, *Sign-based methods in linear statistical models*, "Nauka", Moscow, 1997; English transl., Amer. Math. Soc., Providence, RI, 1997.
6. L. N. Bol'shev and N. V. Smirnov, *Tables of mathematical statistics*, "Nauka", Moscow, 1965. (Russian)
7. A. A. Borovkov, *Mathematical statistics. Estimation of parameters. Testing of hypotheses*, "Nauka", Moscow, 1984; English transl., *Mathematical statistics*, Gordon & Breach, Amsterdam, 1998.
8. ———, *Mathematical statistics. Supplementary chapters*, "Nauka", Moscow, 1984; English transl., Gordon & Breach, Amsterdam, 1998.
9. ———, *Mathematical statistics*, "Nauka", Sibirskoe otdelenie RAN, Novosibirsk, 1997. (Russian)
10. A. A. Borovkov and A. A. Mogul'skiĭ, *Large deviations and the testing of statistical hypotheses*, Proceedings of the Institute of Mathematics, 19, "Nauka", Sibirskoe otdelenie RAN, Novosibirsk, 1992. (Russian)
11. N. N. Čhencov, *Statistical decision rules and optimal inference*, "Nauka", Moscow, 1972; English transl., Amer. Math. Soc., Providence, RI, 1982.
12. D. M. Chibisov, *Certain tests of the chi-square type for continuous distributions*, Teor. Veroyatnost. i Primenen. **16** (1971), no. 1, 3–20; English transl. in Theor. Probability Appl. **16** (1971), no. 1, 1–22.
13. Y. S. Chow, H. Robbins, and D. Siegmund, *The theory of optimal stopping*, Corrected reprint of the 1971 original, Dover, New York, 1991.
14. H. Cramér, *Mathematical methods of statistics*, reprint of the 1946 original, Princeton Univ. Press, Princeton, NJ, 1999.
15. M. H. DeGroot, *Optimal statistical decisions*, McGraw-Hill, New York–London–Sydney, 1970.
16. D. Dugué, *Traité statistique théorique et appliquée: analyse aléatoire, algèbre aléatoire*, Masson et Cie, Paris, 1958. (French)
17. R. S. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer-Verlag, Berlin, 1985.
18. W. Feller, *An introduction to probability theory and its applications*, Third edition, vol. 1, Wiley, New York–London–Sydney, 1968; vol. 2, 1971.
19. I. I. Gikhman, *An introduction to the general theory of measure and integral*, Donetsk University Press, Donetsk, 1971. (Russian)
20. B. V. Gnedenko and A. N. Kolmogorov, *Limit distributions for sums of independent random variables*, Gostekhizdat, Leningrad–Moscow, 1949; English transl., Addison-Wesley, Reading, MA, 1968.
21. P. E. Greenwood and A. N. Shiryayev, *Contiguity and the statistical invariance principle*, Gordon & Breach, New York, 1985.
22. J. Hájek and Z. Šidák, *Theory of rank tests*, Academic Press and Academia Publishing House of the Czechoslovak Academy of Sciences, New York–London and Prague, 1967.
23. P. R. Halmos, *Measure theory*, Van Nostrand, New York, 1950.
24. P.-L. Hennequin and A. Tortrat, *Théorie des probabilités et quelques applications*, Masson et Cie, Éditeurs, Paris, 1965. (French)

25. I. A. Ibragimov and R. Z. Khas'minskiĭ, *Statistical estimation. Asymptotic theory*, "Nauka", Moscow, 1979; English transl., Springer-Verlag, New York–Berlin, 1981.
26. G. I. Ivchenko and Yu. I. Medvedev, *Mathematical statistics*, "Vysshaya shkola", Moscow, 1984. (Russian)
27. ———, *Decomposable statistics and hypotheses testing for grouped data*, Teor. Veroyatnost. i Primenen. **25** (1980), no. 3, 549–560; English transl. in Theory Probab. Appl. **25** (1981), no. 3, 540–551.
28. L. Jacod and A. N. Shiryaev, *Limit theorems for stochastic processes*, 2nd edition, Berlin, Springer-Verlag, 2003.
29. M. G. Kendall and A. Stuart, *The advanced theory of statistics. Inference and relationship*, vol. 2, Griffin, London, 1961.
30. N. Kligene and L. Telksnis, *Methods of detecting instants of change of random processes properties*, Avtomat. i Telemekh. (1983), no. 10, 5–56; English transl. in Automat. Remote Control **44** (1984), no. 10, part 1, 1241–1283.
31. A. N. Kolmogorov and S. V. Fomin, *Introductory real analysis*, "Nauka", Moscow, 1968; English transl., Prentice-Hall, New York, 1970.
32. A. M. Kolodiĭ, *Basics of the general theory of measure and integral*, Volgograd University Press, Volgograd, 1999. (Russian)
33. S. Kullback, *Information theory and statistics*, Dover, New York, 1968.
34. E. L. Lehmann, *Theory of point estimation*, Wiley, New York, 1983.
35. F. Liese and I. Vajda, *Convex statistical distances*, Teubner, Leipzig, 1987.
36. Yu. N. Lin'kov, *The asymptotic distinguishability of two simple statistical hypotheses*, Preprint 86.45, Institute of Mathematics of Ukrainian Academy of Sciences, Kiev, 1986.
37. ———, *Asymptotic statistical methods for stochastic processes*, "Naukova dumka", Kiev, 1993; English transl., Amer. Math. Soc., Providence, RI, 2001.
38. ———, *Lectures on mathematical statistics*, vol. 1, "Istoki", Donetsk, 1999; English transl., Part 1 of this book.
39. ———, *Large deviation theorems for extended random variables and some applications*, J. Math. Sci. **93** (1999), no. 4, 563–573.
40. G. V. Martynov, *Omega-square tests*, "Nauka", Moscow, 1978. (Russian)
41. I. P. Natanson, *Theory of functions of real variable*, GITTL, Moscow, 1957; English transl., vol. 1, Ungar, New York, 1955; vol. 2, 1961.
42. J. Neveu, *Mathematical foundations of the calculus of probability*, Masson et Cie, Éditeurs, Paris, 1964; English transl., Holden-Day, San Francisco–London–Amsterdam, 1965.
43. C. R. Rao, *Statistical inference and its applications*, Wiley, New York–London–Sydney, 1965.
44. R. T. Rockafellar, *Convex analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
45. G. G. Roussas, *Contiguity of probability measures: some applications in statistics*, Cambridge Univ. Press, London–New York, 1972.
46. A. N. Shiryaev, *Optimal stopping rules*, "Nauka", Moscow, 1976; English transl., Springer-Verlag, New York–Heidelberg, 1978.
47. ———, *Probability*, "Nauka", Moscow, 1989; English transl., Springer-Verlag, New York, 1996.
48. A. V. Skorokhod, *Random processes with independent increments*, "Nauka", Moscow, 1964; English transl., Kluwer, Dordrecht, 1991.
49. J.-L. Soler, *Notion de liberté en statistique mathématique*, Thèse de Docteur de Troiseime Cycle, Université de Grenoble, 1970. (French)
50. F. P. Tarasenko, *Nonparametric statistics*, Tomsk University Press, Tomsk, 1976. (Russian)
51. A. Wald, *Sequential analysis*, Wiley and Chapman and Hall, New York and London, 1947.
52. A. Wald, *Statistical decision functions*, Wiley and Chapman and Hall, New York and London, 1950.
53. S. S. Wilks, *Mathematical statistics*, Wiley, New York–London, 1967.
54. S. Zacks, *The theory of statistical inference*, Wiley, New York–London–Sydney, 1971.

Index

Titles in This Series

229 **Yu. N. Lin'kov,** Lectures in mathematical statistics, 2005

228 **D. Zhelobenko,** Principal structures and methods of representation theory, 2005

227 **Takahiro Kawai and Yoshitsugu Takei,** Algebraic analysis of singular perturbation theory, 2005

226 **V. M. Manuilov and E. V. Troitsky,** Hilbert C^*-modules, 2005

225 **S. M. Natanzon,** Moduli of Riemann surfaces, real algebraic curves, and their superanalogues, 2004

224 **Ichiro Shigekawa,** Stochastic analysis, 2004

223 **Masatoshi Noumi,** Painlevé equations through symmetry, 2004

222 **G. G. Magaril-Il'yaev and V. M. Tikhomirov,** Convex analysis: Theory and applications, 2003

221 **Katsuei Kenmotsu,** Surfaces with constant mean curvature, 2003

220 **I. M. Gelfand, S. G. Gindikin, and M. I. Graev,** Selected topics in integral geometry, 2003

219 **S. V. Kerov,** Asymptotic representation theory of the symmetric group and its applications to analysis, 2003

218 **Kenji Ueno,** Algebraic geometry 3: Further study of schemes, 2003

217 **Masaki Kashiwara,** D-modules and microlocal calculus, 2003

216 **G. V. Badalyan,** Quasipower series and quasianalytic classes of functions, 2002

215 **Tatsuo Kimura,** Introduction to prehomogeneous vector spaces, 2003

214 **L. Š. Grinblat,** Algebras of sets and combinatorics, 2002

213 **V. N. Sachkov and V. E. Tarakanov,** Combinatorics of nonnegative matrices, 2002

212 **A. V. Mel'nikov, S. N. Volkov, and M. L. Nechaev,** Mathematics of financial obligations, 2002

211 **Takeo Ohsawa,** Analysis of several complex variables, 2002

210 **Toshitake Kohno,** Conformal field theory and topology, 2002

209 **Yasumasa Nishiura,** Far-from-equilibrium dynamics, 2002

208 **Yukio Matsumoto,** An introduction to Morse theory, 2002

207 **Ken'ichi Ohshika,** Discrete groups, 2002

206 **Yuji Shimizu and Kenji Ueno,** Advances in moduli theory, 2002

205 **Seiki Nishikawa,** Variational problems in geometry, 2001

204 **A. M. Vinogradov,** Cohomological analysis of partial differential equations and Secondary Calculus, 2001

203 **Te Sun Han and Kingo Kobayashi,** Mathematics of information and coding, 2002

202 **V. P. Maslov and G. A. Omel'yanov,** Geometric asymptotics for nonlinear PDE. I, 2001

201 **Shigeyuki Morita,** Geometry of differential forms, 2001

200 **V. V. Prasolov and V. M. Tikhomirov,** Geometry, 2001

199 **Shigeyuki Morita,** Geometry of characteristic classes, 2001

198 **V. A. Smirnov,** Simplicial and operad methods in algebraic topology, 2001

197 **Kenji Ueno,** Algebraic geometry 2: Sheaves and cohomology, 2001

196 **Yu. N. Lin'kov,** Asymptotic statistical methods for stochastic processes, 2001

195 **Minoru Wakimoto,** Infinite-dimensional Lie algebras, 2001

194 **Valery B. Nevzorov,** Records: Mathematical theory, 2001

193 **Toshio Nishino,** Function theory in several complex variables, 2001

192 **Yu. P. Solovyov and E. V. Troitsky,** C^*-algebras and elliptic operators in differential topology, 2001

191 **Shun-ichi Amari and Hiroshi Nagaoka,** Methods of information geometry, 2000

190 **Alexander N. Starkov,** Dynamical systems on homogeneous spaces, 2000

189 **Mitsuru Ikawa,** Hyperbolic partial differential equations and wave phenomena, 2000

188 **V. V. Buldygin and Yu. V. Kozachenko,** Metric characterization of random variables and random processes, 2000

187 **A. V. Fursikov,** Optimal control of distributed systems. Theory and applications, 2000

186 **Kazuya Kato, Nobushige Kurokawa, and Takeshi Saito,** Number theory 1: Fermat's dream, 2000

185 **Kenji Ueno,** Algebraic Geometry 1: From algebraic varieties to schemes, 1999

184 **A. V. Mel'nikov,** Financial markets, 1999

183 **Hajime Sato,** Algebraic topology: an intuitive approach, 1999

182 **I. S. Krasil'shchik and A. M. Vinogradov, Editors,** Symmetries and conservation laws for differential equations of mathematical physics, 1999

181 **Ya. G. Berkovich and E. M. Zhmud',** Characters of finite groups. Part 2, 1999

180 **A. A. Milyutin and N. P. Osmolovskii,** Calculus of variations and optimal control, 1998

179 **V. E. Voskresenskiĭ,** Algebraic groups and their birational invariants, 1998

178 **Mitsuo Morimoto,** Analytic functionals on the sphere, 1998

177 **Satoru Igari,** Real analysis—with an introduction to wavelet theory, 1998

176 **L. M. Lerman and Ya. L. Umanskiy,** Four-dimensional integrable Hamiltonian systems with simple singular points (topological aspects), 1998

175 **S. K. Godunov,** Modern aspects of linear algebra, 1998

174 **Ya-Zhe Chen and Lan-Cheng Wu,** Second order elliptic equations and elliptic systems, 1998

173 **Yu. A. Davydov, M. A. Lifshits, and N. V. Smorodina,** Local properties of distributions of stochastic functionals, 1998

172 **Ya. G. Berkovich and E. M. Zhmud',** Characters of finite groups. Part 1, 1998

171 **E. M. Landis,** Second order equations of elliptic and parabolic type, 1998

170 **Viktor Prasolov and Yuri Solovyev,** Elliptic functions and elliptic integrals, 1997

169 **S. K. Godunov,** Ordinary differential equations with constant coefficient, 1997

168 **Junjiro Noguchi,** Introduction to complex analysis, 1998

167 **Masaya Yamaguti, Masayoshi Hata, and Jun Kigami,** Mathematics of fractals, 1997

166 **Kenji Ueno,** An introduction to algebraic geometry, 1997

165 **V. V. Ishkhanov, B. B. Lur'e, and D. K. Faddeev,** The embedding problem in Galois theory, 1997

164 **E. I. Gordon,** Nonstandard methods in commutative harmonic analysis, 1997

163 **A. Ya. Dorogovtsev, D. S. Silvestrov, A. V. Skorokhod, and M. I. Yadrenko,** Probability theory: Collection of problems, 1997

162 **M. V. Boldin, G. I. Simonova, and Yu. N. Tyurin,** Sign-based methods in linear statistical models, 1997

161 **Michael Blank,** Discreteness and continuity in problems of chaotic dynamics, 1997

This volume is intended for the advanced study of topics in mathematical statistics. The first part of the book is devoted to sampling theory (from one-dimensional and multidimensional distributions), asymptotic properties of sampling, parameter estimation, sufficient statistics, and statistical estimates. The second part is devoted to hypothesis testing and includes the discussion of families of statistical hypotheses that can be asymptotically distinguished. In particular, the author describes goodness-of-fit and sequential statistical criteria (Kolmogorov, Pearson, Smirnov, and Wald) and studies their main properties.

For additional information
and updates on this book, visit
www.ams.org/bookpages/mmono-229

ISBN 0-8218-3732-X

9 780821 837320

MMONO/229