Asymptotic Theory of Statistical Estimation [1]

Jiantao Jiao
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Email: jiantao@eecs.berkeley.edu

September 11, 2019

[1] Summary of Chapters in [1]

Contents

Acknowledgment

Thank Russian mathematicians for providing excellent research monographs.

Chapter 1

The Problem of Statistical Estimation

1.1 Formulation of the Problem of Statistical Estimation

Let $(\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta)$ be a statistical experiment generated by the observation X. Let φ be a measurable function from $(\Theta, \mathcal{B})$ into $(\mathscr{Y}, \mathcal{Y})$. Consider the problem of estimating the value of $\varphi(\theta)$ at point θ based on observation X, whose distribution is P_θ. Our only information about θ is that $\theta \in \Theta$. As an estimator for $\varphi(\theta)$ one can choose any function of observations $T(X)$ with values in $(\mathscr{Y}, \mathcal{Y})$. Therefore the following problem arises naturally: how to choose statistic T which would estimate $\varphi(\theta)$ in the best possible manner. However, what is the meaning of the expression "in the best (possible) manner"?

Assume there is on the set $\mathscr{Y} \times \mathscr{Y}$ a real-valued nonnegative function $W(y_1, y_2)$ which we call *loss function* and which has the following meaning: if observation X is distributed according to the distribution P_θ, then utilizing statistic $T(X)$ to estimate $\varphi(\theta)$ yields a loss which is equal to $W(T(X), \varphi(\theta))$. Averaging over all possible values of X, we arrive at the *risk function*:

$$R_W(T; \theta) = \mathbb{E}_\theta W(T(X), \varphi(\theta)), \tag{1.1}$$

which will be chosen as the measure of the quality of statistic T as an estimator of $\varphi(\theta)$ for a given loss function W.

Thus a partial ordering is introduced on the space of all estimators of $\varphi(\theta)$: the estimator T_1 is preferable to T_2 if for all $\theta \in \Theta$, $R_W(T_1; \theta) \le R_W(T_2; \theta)$.

In view of the last definition, estimator T of the function $\varphi(\theta)$ is called *inadmissible* with respect to loss function W if there exists an estimator T^* such that

$$R_W(T^*; \theta) \le R_W(T; \theta), \quad \forall \theta \in \Theta, \qquad R_W(T^*; \theta) < R_W(T; \theta) \quad \text{for some } \theta \in \Theta. \tag{1.2}$$

An estimator which is not inadmissible is called *admissible*.

Although the approach described above is commonly used, it is not free of certain shortcomings. First, many estimators turn out to be uncomparable, and second the choice of loss functions is arbitrary to a substantial degree.

Sometimes it is possible to find estimators which are optimal within a certain class which is smaller than the class of all estimators. One such class is the class of unbiased estimators: an estimator T is called an *unbiased estimator* of function $\varphi(\theta)$ if $\mathbb{E}_\theta T = \varphi(\theta)$ for all $\theta \in \Theta$.

Furthermore, if the initial experiment is invariant with respect to a group of transformations it is natural to confine ourselves to a class of estimators which do not violate the symmetry of the problem. This is called *invariance principle*.

Comparing estimators T_1, T_2 according to their behavior at the "least favorable" points, we arrive at the notion of a minimax estimator. An estimator T_0 is called the *minimax estimator* of $\varphi(\theta)$ in $\Theta_1 \subset \Theta$ relative to loss function W if

$$\sup_{\theta \in \Theta_1} R_W(T_0; \theta) = \inf_T \sup_{\theta \in \Theta_1} R_W(T; \theta), \tag{1.3}$$

where the inf is taken over all estimators T of $\varphi(\theta)$.

If Θ is a subset of a finite-dimensional Euclidean space, then statistical estimation problems based this experiment is called *parametric estimation problems*.

Below we shall mainly deal with parametric problems. Moreover, we shall always assume that Θ is an open subset of a finite-dimensional Euclidean space $\mathbb{R}^k$, and that the family of distributions P_θ and the densities $p(x; \theta) = \frac{dP_\theta}{d\mu}$ are defined on the closure $\bar{\Theta}$ of the set Θ. By $\mathcal{B}$ we shall denote the σ-algebra of Borel subsets of Θ.

In the parametric case it is usually the parameter itself that is estimated (i.e. $\varphi(\theta) = \theta$). In this case the loss function W is defined on the set $\Theta \times \Theta$ and as a rule we shall consider loss functions which possess the following properties:

1. $W(u, v) = w(u - v)$.

2. The function $w(u)$ is defined and is nonnegative on $\mathbb{R}^k$; moreover, $w(0) = 0$ and $w(u)$ is continuous at $u = 0$ but is not identically zero.

3. Function w is symmetric, i.e. $w(u) = w(-u)$.

4. The sets $\{u : w(u) < c\}$ are convex sets for all $c > 0$.

5. The sets $\{u : w(u) < c\}$ are convex sets for all $c > 0$ and are bounded for all $c > 0$ sufficiently small.

The function w will also be called loss functions.

The first three properties are quite natural and do not require additional comments. Property 4 in the case of a one-dimensional parameter set means that function $w(u)$ is non-decreasing on $[0, \infty)$. Denote by $\mathbf{W}$ the class of loss functions satisfying 1-4; the same notation will also be used for the corresponding set of functions w. Denote by $\mathbf{W}'$ the class of functions satisfying 1-5. The notation $\mathbf{W}(\mathbf{W}')$ will be used for the set of functions w which posses a polynomial majorant. Denote by $\mathbf{W}_{e,\alpha}(\mathbf{W}'_{e,\alpha})$ the set of functions w belonging to $\mathbf{W}(\mathbf{W}')$ whose growth as $|u| \to \infty$ is slower than any one of the functions $e^{\epsilon|u|^\alpha}, \epsilon > 0$.

Clearly, all loss functions of the form $|u - v|^r, r > 0$, and the indicator loss function $W_A(u, v) = I(u - v \notin A)$ belong to the class $\mathbf{W}'_p$.

Theorem 1.1.1 (Blackwell's theorem)**.** *Let the family $\{P_\theta, \theta \in \Theta\}$ possess a sufficient statistic T. Let the loss function be of the form $w(u - v)$, where $w(u)$ is a convex function in $\mathbb{R}^k$. Let θ^* be an arbitrary estimator for θ. Then there exists an estimator T^* representable in the form $g(T)$ and such that for all $\theta \in \Theta$,*

$$\mathbb{E}_\theta w(T^* - \theta) \le \mathbb{E}_\theta w(\theta^* - \theta). \tag{1.4}$$

If θ^* is an unbiased estimator for θ, T^* can also be chosen to be unbiased.

Consider again a parametric statistical experiment. Now we assume that θ is a random variable with a known distribution Q on Θ. In such a situation the estimation problem is called the *estimation problem in the Bayesian formulation.* Assume, for simplicity, that Q possesses density q with respect to the Lebesgue measure. If, as before, the losses are measured by means of function w, then the mean loss obtained using estimator T (the so-called *Bayesian risk of estimator T*) is equal to

$$r_w(T) = \mathbb{E}_Q w(T - \theta) = \int_\Theta q(\theta) d\theta \int_{\mathscr{X}} w(T(x) - \theta) P_\theta(dx) = \int_\Theta R_w(T; \theta) q(\theta) d\theta. \tag{1.5}$$

In the Bayesian setup the risk of estimator R is a positive number and one can talk about the best estimator $\tilde{T}$ minimizing risk r_w:

$$r_w(\tilde{T}) = \min_T r_w(T). \tag{1.6}$$

Here we assume the minimum is achievable. The estimator $\tilde{T}$ is called the *Bayesian estimator with respect to loss function w and prior distribution Q.*

Evidently the form of the optimal estimator $\tilde{T}$ depends on the prior density q. On the other hand, one may assume that as the sample size increases to infinite the Bayesian estimator $\tilde{T}$ ceases to depend on the initial distribution Q within a wide class of these distributions (e.g. those Q for which $q > 0$ on Θ). Therefore, for an asymptotic treatment of Bayesian problems the exact knowledge of q is not so obligatory anymore.

1.2 Some Examples

1.2.1 Hodges' and Lehmann's Result

We shall first formulate a simple general criterion due to Lehmann which is useful for proving the minimax property of certain estimators.

Theorem 1.2.1. *Let T_k be a Bayesian estimator with respect to the distribution λ_k on Θ and the loss function W, $k = 1, 2, \ldots$. If the estimator T is such that for all $\theta \in \Theta$,*

$$\mathbb{E}_\theta W(\theta, T) \le \limsup_k \int_\Theta \mathbb{E}_\theta W(\theta, T_k) d\lambda_k(\theta), \tag{1.7}$$

it is minimax.

As a corollary to this theorem we obtain the following result of Hodges and Lehmann.

Theorem 1.2.2. *Let T be an estimator which is Bayesian with respect to W and probability measure λ on Θ. Denote by $\Theta_0 \subset \Theta$ the support of λ. If*

1. $\mathbb{E}_\theta W(\theta, T) \equiv c$ *for all* $\theta \in \Theta_0$,
2. $\mathbb{E}_\theta W(\theta, T) \leq c$ *for all* $\theta \in \Theta$,

then T is minimax estimator.

1.2.2 Estimation of the Mean of a Normal Distribution

Let X_j possess normal distribution $\mathcal{N}(\theta, 1)$ on the real line. Denote by T_k an estimator which is Bayesian with respect to the normal distribution λ_k with mean zero and variance $\sigma_k^2 = k$. Since the loss function is quadratic, then

$$T_k = \frac{nk}{nk+1}\bar{X}. \tag{1.8}$$

Therefore

$$\int_{-\infty}^{\infty} \mathbb{E}_u (T_k - u)^2 d\lambda_k = \frac{k}{nk+1}. \tag{1.9}$$

For all $\theta \in \Theta$,

$$\mathbb{E}_\theta (\bar{X} - \theta)^2 = n^{-1} = \lim_k \int_{-\infty}^{\infty} \mathbb{E}_u (T_k - u)^2 d\lambda_k \tag{1.10}$$

and it follows from Theorem 1.2.1 that $\bar{X}$ is a minimax estimator.

Consequently, also the equivariant estimator $\bar{X}$ is optimal in the class of equivariant estimators. We note immediately that $\bar{X}$ is admissible as well (we can easily show that using results based on Fisher information later). Hence in the problem under consideration, the estimator $\bar{X}$ of parameter θ has all the possible virtues: it is unbiased, admissible, minimax, and equivariant. These properties of $\bar{X}$ are retained also in the case when X_j are normally distributed in $\mathbb{R}^2$.

If, however, X_j are normally distributed in $\mathbb{R}^k, k \geq 3$, with the density $\mathcal{N}(\theta, J)$, then the statistic $\bar{X}$ relative to the loss function $W(\theta, t) = |\theta - t|^2$ loses all of its remarkable properties. It is even inadmissible in this case.

We now present briefly the method of Stein of constructing estimators which are better than $\bar{X}$. The following simple relation is the basis for Stein's construction. Let ξ be a random variable with mean a and variance σ^2. Furthermore, let $\varphi(x)$ be absolutely continuous on $\mathbb{R}^1$ and $\mathbb{E}|\varphi'(\xi)| < \infty$. Then

$$\sigma^2 \mathbb{E}\varphi'(\xi) = \mathbb{E}(\xi - a)\varphi(\xi). \tag{1.11}$$

Indeed, integrating by parts we obtain

$$\mathbb{E}\varphi'(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \varphi'(x) e^{-(x-a)^2/2\sigma^2} dx = -\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \varphi(x) d\left(e^{-\frac{(x-a)^2}{2\sigma^2}}\right) = \sigma^{-2}\mathbb{E}\varphi(\xi)(\xi - a). \tag{1.12}$$

Now let $\xi = (\xi_1, \ldots, \xi_k)$ be a normal random vector in $\mathbb{R}^k$ with mean a and correlation matrix $\sigma^2 J$, where J is the identity matrix. Furthermore, let the function $\varphi = (\varphi_1, \ldots, \varphi_k) : \mathbb{R}^k \to \mathbb{R}^k$ be differentiable and $\mathbb{E}|\frac{\partial \varphi_i(\xi)}{\partial \xi_i}| < \infty$. Under these assumptions the following identity is obtained from Stein's identity:

$$\mathbb{E}\left[\sigma^2 \frac{\partial \varphi_i}{\partial \xi_i}(\xi) - (\xi_i - a_i)\varphi_i(\xi)\right] = 0. \tag{1.13}$$

Return now to the sequence of iid observations $X_1, \ldots, X_n$, where X_j possesses a normal distribution in $\mathbb{R}^k$ with the density $\mathcal{N}(\theta, J)$. An estimator of θ will be sought among the estimators of the form

$$\tilde{\theta}_n = \bar{X} + n^{-1} g(\bar{X}), \tag{1.14}$$

where the function $g(x) = (g_1, \ldots, g_k) : \mathbb{R}^k \to \mathbb{R}^k$. In view of (1.13),

$$\mathbb{E}_\theta |\bar{X} - \theta|^2 - \mathbb{E}_\theta |\bar{X} + n^{-1} g(\bar{X}) - \theta|^2 = -2n^{-1}\mathbb{E}_\theta \langle \bar{X} - \theta, g(\bar{X})\rangle - n^{-2}\mathbb{E}_\theta |g(\bar{X})|^2 \tag{1.15}$$

$$= -2n^{-2}\mathbb{E}_\theta \left(\sum_{i=1}^{k} \frac{\partial g_i}{\partial x_i}(\bar{X})\right) - n^{-2}\mathbb{E}_\theta |g(\bar{X})|^2. \tag{1.16}$$

Assume now that the function g can be represented in the form $g(x) = \nabla \ln \varphi(x)$, where $\varphi(x)$ is a twice differentiable function from $\mathbb{R}^k$ to $\mathbb{R}^1$. Then

$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial x_i}(x) = \sum_{i=1}^{k} \frac{\partial}{\partial x_i}\left(\frac{1}{\varphi(x)}\frac{\partial}{\partial x_i}\varphi(x)\right) = -|g|^2 + \frac{1}{\varphi}\Delta\varphi, \tag{1.17}$$

where $\Delta = \sum_{i=1}^{k} \frac{\partial^2}{\partial x_i^2}$ is the Laplace operator. Consequently, for the above choice of g,

$$\mathbb{E}_\theta|\bar{X} - \theta|^2 - \mathbb{E}_\theta|\bar{X} + n^{-1}g(\bar{X}) - \theta|^2 = n^{-2}\mathbb{E}_\theta|g|^2 - n^{-2}\mathbb{E}_\theta\left[\frac{1}{\varphi(\bar{X})}\Delta\varphi(\bar{X})\right]. \tag{1.18}$$

The right hand side of the last equality is obviously positive provided $\varphi(x)$ is a positive nonconstant superharmonic function–this means that $\Delta\varphi \le 0$. Since there are not superharmonic functions bounded from below on the real line of on a plane which are not constant, the proposed approach does not improve on the estimator $\bar{X}$ in these two cases. However, in spaces of dimension $k \ge 3$, there exist a substantial number of nonnegative superharmonic functions. Consider, for example, the function

$$\varphi_k(x) = \begin{cases} |x|^{-(k-2)} & |x| \ge \sqrt{k-2} \\ (k-2)^{-(k-2)/2} e^{\frac{1}{2}((k-2)-|x|^2)} & |x| \le \sqrt{k-2} \end{cases} \tag{1.19}$$

This function is positive and superharmonic in $\mathbb{R}^k, k \ge 3$,

$$\nabla \ln \varphi_k = \begin{cases} -\frac{k-2}{|x|^2}x & |x| \ge \sqrt{k-2} \\ -x & |x| \le \sqrt{k-2} \end{cases} \tag{1.20}$$

Thus the Stein–James estimator

$$\bar{X} + \frac{1}{n}\nabla \ln \varphi_k(\bar{X}) = \begin{cases} \left(1 - \frac{k-2}{n|\bar{X}|^2}\right)\bar{X} & |x| \ge \sqrt{k-2} \\ \left(1 - \frac{1}{n}\right)\bar{X} & |x| \le \sqrt{k-2} \end{cases} \tag{1.21}$$

is uniformly better than $\bar{X}$. It is worth mentioning, however, that as $n \to \infty$ the improvement is of order n^{-2} while $\mathbb{E}_\theta|\bar{X} - \theta|^2 = k/n$ is of order n^{-1}.

Another improvement on estimator $\bar{X}$ may be obtained by setting $\varphi(x) = |x|^{k-2}$. Here $\varphi(x)$ is a harmonic function and the corresponding estimator is of the form:

$$\tilde{\theta}_n = \left(1 - \frac{k-2}{n|\bar{X}|^2}\right)\bar{X} \tag{1.22}$$

This estimator was suggested by Stein. For Stein's estimator

$$\mathbb{E}_\theta|\bar{X} - \theta|^2 - \mathbb{E}_\theta|\tilde{\theta}_n - \theta|^2 = \left(\frac{k-2}{n}\right)^2 \mathbb{E}_\theta|\bar{X}|^{-2} = \left(\frac{k-2}{2}\right)^2 \frac{1}{(2\pi)^{k/2}} \int_{\mathbb{R}^k} \left|\theta + \frac{x}{\sqrt{n}}\right|^{-2} e^{-|x|^2/2} dx. \tag{1.23}$$

As before, as $n \to \infty$,

$$\mathbb{E}_\theta|\bar{X} - \theta|^2 - \mathbb{E}_\theta|\tilde{\theta}_n - \theta|^2 = O(n^{-2}) \tag{1.24}$$

provided $\theta = 0$. However, for $\theta = 0$, the different equals to $(k-2)/n$. Thus for $\theta = 0$, Stein's estimator is substantially better than $\bar{X}$:

$$\frac{\mathbb{E}_\theta|\bar{X} - \theta|^2}{\mathbb{E}_\theta|\tilde{\theta}_n - \theta|^2} = \frac{k}{2} > 1. \tag{1.25}$$

1.3 Consistency. Methods for Constructing Consistent Estimators

1.3.1 An Existence Theorem

Consider a sequence of statistical experiments $(\mathscr{X}^n, \mathcal{X}^n, P_\theta^n, \theta \in \Theta)$ generated by observations $X^n = (X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ is a sequence of independent observations with values in $(\mathscr{X}, \mathcal{X})$. Would it not be possible to achieve arbitrary precision by increasing indefinitely the number of observations n?

A sequence of statistics $\{T_n(X_1, X_2, \ldots, X_n)\}$ is called a *consistent sequence of estimators* for the value $\varphi(\theta)$ of the function $\varphi : \Theta \to \mathbb{R}^l$ if

$$T_n \overset{P_\theta^n}{\to} \varphi(\theta). \tag{1.26}$$

Below we shall consider more general families of experiments than the sequences of repeated samples and we shall now extend the definition of consistent estimators to this formally more general case.

Consider the family of experiments $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$ generated by observations X^ϵ, where ϵ is a real parameter. For our purposes, it is sufficient to deal with the case when $\epsilon \in (0,1)$ and asymptotic estimation problems are studied for $\epsilon \to 0$. Observe that the case of a sequence of experiments $(\mathscr{X}^n, \mathcal{X}^n, P_\theta^n, \theta \in \Theta)$ is a particular case of this scheme: choose $\epsilon = n^{-1}, n = 1, 2, \ldots$.

The family of statistics $T_\epsilon = T_\epsilon(X^\epsilon)$ is called a *consistent family of estimators* for the value $\varphi(\theta)$ of the function $\varphi : \Theta \to \mathbb{R}^l$ if $T_\epsilon \to \varphi(\theta)$ in $P_\theta^{(\epsilon)}$ probability as $\epsilon \to 0$ for all $\theta \in \Theta$.

A family of statistics $\{T_\epsilon\}$ is called a *uniformly consistent* family of estimators for $\varphi(\theta)$ on the set $K \subset \Theta$ if $T_n \to \varphi(\theta)$ in $P_\theta^{(\epsilon)}$-probability as $\epsilon \to \infty$ uniformly in $\theta \in K$. The latter means that for any $\delta > 0$,

$$\sup_{\theta \in K} P_\theta^{(\epsilon)}(|T_\epsilon(X^\epsilon) - \varphi(\theta)| > \delta) \to 0. \tag{1.27}$$

Observe that since the only information about parameter θ is that it belongs to the set Θ, uniformly consistent estimators in Θ or in any compact set $K \subset \Theta$ are useful.

In this section we shall consider the most commonly employed methods for constructing consistent estimators. We first verify that in the case of repeated sampling, consistent estimators exist under very general assumptions. For repeated experiments, we consider on Θ the Hellinger distances between the measures $\int_A f(x;\theta)d\nu$ and $\int_A f(x;\theta')d\nu$:

$$r_p(\theta;\theta') = \left(\int_{\mathscr{X}} |f^{1/p}(x;\theta) - f^{1/p}(x;\theta')|^p \nu(dx)\right)^{1/p}, \quad p \geq 1. \tag{1.28}$$

Clearly, $0 \leq r_p(\theta;\theta') \leq 1$, and in view of the condition $P_\theta \neq P_{\theta'}$, $\theta \neq \theta'$, $r_p(\theta;\theta') = 0$ if and only if $\theta = \theta'$. Below we shall often use the distances r_1 and r_2.

Theorem 1.3.1. *Let the conditions*

1. $\inf_{\theta':|\theta-\theta'|>\delta} r_1(\theta;\theta') > 0$, *if* $\delta > 0, \theta \in \Theta$
2. $\lim_{\theta' \to \theta} r_1(\theta;\theta') = 0$

be satisfied. Then there exists a sequence $\{T_n\}$ *of estimators consistent for* θ. *If the conditions above are satisfied uniformly in* θ *belong to the compact set* $K \subset \Theta$, *then there exists a sequence of uniformly consistent estimators in* K.

Theorem 1.3.1 is an existence theorem which cannot be used for actual determination of consistent estimators. We now turn to a consideration of practically usable methods.

1.3.2 Method of moments

This method was suggested by K. Pearson and is historically the first general method for construction of estimators. In general terms it can be described as follows:

Let $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$, $\Theta \subset \mathbb{R}^k$ be a family of statistical experiments and let $g_1(\theta), \ldots, g_k(\theta)$ be k real-valued functions on Θ. Assume that there exist consistent estimators $\tilde{g}_1^\epsilon, \ldots, \tilde{g}_k^\epsilon$ for $g_1(\theta), \ldots, g_k(\theta)$. The method of moments recommends that we choose as an estimator for θ the solution of the system of equations

$$g_i(\theta) = \tilde{g}_i^\epsilon, \quad i = 1, 2, \ldots, k. \tag{1.29}$$

To specify this system, consider once more statistical experiment generated by a sequence of iid random variables $X_1, X_2, \ldots$. Assume that X_i are real valued random variables with a finite k-th moment and let $\alpha_\nu(\theta) = \mathbb{E}_\theta X_1^\nu, \nu \leq k$. Denote by a_ν the ν-th sample moment, i.e. $a_\nu = (\sum_{i=1}^n X_i^\nu)/n$. It is known that a_ν is a consistent estimator for α_ν. Thus, in view of that stated above, one can choose as an estimator for θ the solution of the system of equations

$$\alpha_\nu(\theta) = a_\nu, \quad \nu = 1, 2, \ldots, k. \tag{1.30}$$

Theorem 1.3.2. *Let functions* $\alpha_\nu(\theta)$ *possess continuous partial derivatives on* Θ *and let the Jacobian* $\det|\partial\alpha_\nu/\partial\theta_i|, \theta = (\theta_1, \ldots, \theta_k)$ *be different from zero everywhere on* Θ. *Let the method of moments equation possess a unique solution* T_n *with probability approaching* 1 *as* $n \to \infty$. *Then this solution is a consistent estimator for* θ.

1.3.3 Method of Maximum Likelihood

This method, suggested by R. A. Fisher, is one of the commonly used general methods for determination of consistent estimators.

Let $dP_\theta/d\mu = p(x;\theta)$. Let X be the observation. The function $p(X;\theta)$ is called the *likelihood function* corresponding to the experiment; thus $p(X;\theta)$ is a random function of θ defined on $\Theta \subset \mathbb{R}^k$. The statistic $\hat{\theta}$ defined by the relation

$$p(X;\hat{\theta}) = \sup_{\theta\in\bar{\Theta}} p(X;\theta) \tag{1.31}$$

is called the *maximum likelihood estimator* for the parameter θ based on the observation X. Obviously, it may turn out that the maximization has no solution. However, below we shall consider the case when the solution does exist. If there are several maximizers we shall assume, unless otherwise specified, that any one of then is a maximum likelihood estimator.

If $p(X;\theta)$ is a smooth function of θ, and $\hat{\theta} \in \Theta$ then $\hat{\theta}$ is necessarily also a solution for θ of the *likelihood equation*

$$\frac{\partial}{\partial\theta_i} \ln p(X;\theta) = 0, i = 1, 2, \ldots, k, \theta = (\theta_1, \ldots, \theta_k). \tag{1.32}$$

To prove the consistency of the maximum likelihood estimators it is convenient to utilize the following simple general result.

Lemma 1.3.1. *Let $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$ be a family of experiments and let the likelihood functions $p_\epsilon(X^\epsilon;\theta)$ correspond to these experiments. Set*

$$Z_{\epsilon,\theta}(u) = Z_\epsilon(u) = \frac{p_\epsilon(X^\epsilon;\theta+u)}{p_\epsilon(X^\epsilon;\theta)}, u \in U = \Theta - \theta. \tag{1.33}$$

Then in order to have the maximum likelihood estimator $\hat{\theta}_\epsilon$ be consistent it is sufficient that for all $\theta \in \Theta$ and $\gamma > 0$,

$$P_\theta^{(\epsilon)}(\sup_{|u|>\gamma} Z_{\epsilon,\theta}(u) > 1) \to 0 \tag{1.34}$$

as $\epsilon \to 0$. If the last relation is uniform in $\theta \in K \subset \Theta$, then the estimator $\hat{\theta}_\epsilon$ is uniformly consistent in K.

We now turn to the case of independent identically distributed observations $X_1, X_2, \ldots, X_n$, where X_i possesses the density $f(x;\theta)$ with respect to measure ν. The maximum likelihood estimator $\hat{\theta}_n$ is the solution of the equation

$$\prod_{i=1}^n f(X_i;\hat{\theta}_n) = \sup_\theta \prod_{i=1}^n f(X_i;\theta). \tag{1.35}$$

We show that under quite general assumptions $\hat{\theta}_n$ is a consistent estimator.

Theorem 1.3.3. *Let Θ be a bounded open set in $\mathbb{R}^k$, $f(x;\theta)$ be a continuous function of θ on $\bar{\Theta}$ for almost all $x \in \mathscr{X}$ and let the following conditions be fulfilled:*

1. *For all $\theta \in \Theta$ and all $\gamma > 0$,*
$$\inf_{|\theta'-\theta|>\gamma} r_2^2(\theta;\theta') = k_\theta(\gamma) > 0 \tag{1.36}$$
2. *For all $\theta \in \bar{\Theta}$*
$$\left(\int_{\mathscr{X}} \sup_{|h|\le\delta} (f^{1/2}(x;\theta) - f^{1/2}(x;\theta+h))^2 d\nu\right)^{1/2} = \omega_\theta(\delta) \to 0 \tag{1.37}$$
as $\delta \to 0$.

Then for all $\theta \in \Theta$ the estimator $\hat{\theta}_n \to \theta$ as $n \to \infty$ with probability one.

In order to prove the consistency of maximum likelihood estimators in the case of an unbounded parameter set, additional conditions dealing with the interrelation between $f(x;\theta)$ and $f(x;\theta')$ when $|\theta-\theta'| \to \infty$ will be needed. The simplest variant of such a condition is the following.

Theorem 1.3.4. *Let Θ be an open set in $\mathbb{R}^k$, $f(x;\theta)$ be a continuous function of θ for ν-almost all x and let conditions in Theorem 1.3.3 as well as the condition: for all $\theta \in \Theta$,*

$$\lim_{c\to\infty} \int_{\mathscr{X}} \sup_{|u|\ge c} (f^{1/2}(x;\theta) f^{1/2}(x;\theta+u)) d\nu < 1 \tag{1.38}$$

be fulfilled. Then $\hat{\theta}_n$ is a consistent estimator for θ.

1.3.4 Bayesian Estimates

Theorem 1.3.5. *Let Θ be an open bounded set in $\mathbb{R}^k$ and the density $f(x;\theta)$ satisfy the following conditions:*

1. $\inf_{|\theta-\theta'|>\gamma}\int_{\mathscr{X}}(f^{1/2}(x;\theta)-f^{1/2}(x;\theta'))^2d\nu = k_\theta(\gamma)>0$ *for all* $\theta\in\Theta, \gamma>0$.
2. $\int_{\mathscr{X}}(f^{1/2}(x;\theta+h)-f^{1/2}(x;\theta))^2d\nu = O\left(\frac{1}{(\ln h)^2}\right), h\to 0$, *for all* $\theta\in\Theta$.

Then the estimator $\tilde{t}_n$, which is Bayesian relative to the loss function $W(u;\theta)=|u-\theta|^\alpha, \alpha\geq 1$ and the prior density $q(\theta)$ where $q(\theta)$ is continuous and positive on Θ, is a consistent estimator of the parameter θ.

1.4 Inequalities for Probabilities of Large Deviations

1.4.1 Convergence of $\hat{\theta}_\epsilon$ to θ

Let a family of statistical experiments $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta\in\Theta)$ generated by observations X^ϵ be given. Here Θ is an open subset of $\mathbb{R}^k$. Let $p_\epsilon(X^\epsilon;\theta)=\left(\frac{dP_\theta^{(\epsilon)}}{d\nu^\epsilon}\right)(X^\epsilon)$ where ν^ϵ is a measure on $\mathcal{X}^{(\epsilon)}$. Consider the random function

$$\zeta_\epsilon(u)=\frac{p_\epsilon(X^\epsilon;\theta+u)}{p_\epsilon(X^\epsilon;\theta)} \tag{1.39}$$

where θ is the "true" value of the parameter. Above we have seen that (see Lemma 1.3.1) that as long as function $\zeta_\epsilon(u)$ is sufficiently small for large u, the maximum likelihood estimators $\hat{\theta}_\epsilon$ constructed from observations X^ϵ are consistent on Θ. One can expect that by classifying in some manner the decrease of $\zeta_\epsilon(u)$ as $u\to\infty$ and $\epsilon\to 0$ we could pinpoint more closely the nature of convergence of $\hat{\theta}_\epsilon$ to θ. Let for example, for any $\gamma>0$,

$$P_\theta^{(\epsilon)}\left(\sup_{|u|>\gamma}\frac{p_\epsilon(X^\epsilon;\theta+u\epsilon)}{p_\epsilon(X^\epsilon;\theta)}\geq 1\right)\to\infty \tag{1.40}$$

as $\epsilon\to 0$, then evidently also

$$P_\theta^{(\epsilon)}(\epsilon^{-1}|\hat{\theta}_\epsilon-\theta|>\gamma)\to 0 \tag{1.41}$$

as $\epsilon\to 0$.

This sample device is essentially the basis for proofs of all the theorems stated below.

Set

$$Z_{\epsilon,\theta}(u)=Z_\epsilon(u)=\frac{p_\epsilon(X^\epsilon;\theta+\varphi(\epsilon)u)}{p_\epsilon(X^\epsilon;\theta)} \tag{1.42}$$

where $\varphi(\epsilon)$ denotes some matrix nondegenerated normalizing factor; it is also assumed that $|\varphi(\epsilon)|\to 0$ as $\epsilon\to 0$. Thus the function $Z_{\epsilon,\theta}$ is defined on the set $U_\epsilon=(\varphi(\epsilon))^{-1}(\Theta-\theta)$.

Below we shall denote by $\mathbf{G}$ the set of families of functions $\{g_\epsilon(y)\}$ possessing the following properties:

1. For a fixed ϵ, $g_\epsilon(y)$ is monotonically increasing to ∞ as a function of y defined on $[0,\infty)$
2. For any $N>0$,
$$\lim_{y\to\infty,\epsilon\to 0} y^N e^{-g_\epsilon(y)}=0. \tag{1.43}$$

For example, if $g_\epsilon(y)=y^\alpha, \alpha>0$, then $g_\epsilon\in\mathbf{G}$.

We shall agree throughout this section to denote nonnegative constants by the letter B, which the letter b will be reserved for positive constants. When we wish to emphasize the dependence of these constants on certain parameters $a_1, a_2, \ldots$, we shall sometimes write $B(a_1, a_2, \ldots)$.

Theorem 1.4.1. *Let functions $Z_{\epsilon,\theta}(u)$ be continuous with probability 1 and possess the following properties: given an arbitrary compact set $K\subset\Theta$ there correspond to it nonnegative numbers M_1 and m_1 (depending on K) and functions $g_\epsilon^K(y)=g_\epsilon(y), \{g_\epsilon\}\in\mathbf{G}$ such that*

1. *There exist numbers $\alpha>k$ and $m\geq\alpha$ such that for all $\theta\in K$,*
$$\sup_{|u_1|\leq R,|u_2|\leq R}|u_2-u_1|^{-\alpha}\mathbb{E}_\theta^{(\epsilon)}|Z_{\epsilon,\theta}^{1/m}(u_2)-Z_{\epsilon,\theta}^{1/m}(u_1)|^m\leq M_1(1+R^{m_1}). \tag{1.44}$$

2. For all $u \in U_\epsilon, \theta \in K$,

$$\mathbb{E}_\theta^{(\epsilon)} Z_{\epsilon,\theta}^{1/2}(u) \le e^{-g_\epsilon(|u|)}. \tag{1.45}$$

Then the maximum likelihood estimator $\hat{\theta}_\epsilon$ is consistent and for all ϵ sufficiently small, $0 < \epsilon < \epsilon_0$,

$$\sup_{\theta \in K} P_\theta^{(\epsilon)}(|(\varphi(\epsilon))^{-1}(\hat{\theta}_\epsilon - \theta)| > H) \le B_0 e^{-b_0 g_\epsilon(H)}, \tag{1.46}$$

where $B_0, b_0 > 0$ are constants.

This theorem, as well as Theorem 1.4.2 below, may appear to be exceedingly cumbersome, involving conditions which are difficult to verify. In the next subsection we shall, however, illustrate the usefulness of these theorems by applying them to sequences of independent homogeneous observations. Both Theorem 1.4.1 and 1.4.2 play a significant role in the subsequent chapters as well.

Corollary 1.4.1. *Under the conditions of Theorem 1.4.1 for any $N > 0$ we have*

$$\lim_{H \to \infty, \epsilon \to 0} H^N \sup_{\theta \in K} P_\theta^{(\epsilon)}(|(\varphi(\epsilon))^{-1}(\hat{\theta}_\epsilon - \theta)| > H) = 0. \tag{1.47}$$

Corollary 1.4.2. *Under the conditions of Theorem 1.4.1 for any function $w \in \mathbf{W}_p$,*

$$\limsup_{\epsilon \to 0} \mathbb{E}_\theta^{(\epsilon)} w((\varphi(\epsilon))^{-1}(\hat{\theta}_\epsilon - \theta)) < \infty. \tag{1.48}$$

If we replace the maximum likelihood estimator by a Bayesian one, condition (1) of Theorem 1.4.1 may be substantially weakened, the only requirement being that α is positive.

Theorem 1.4.2. *Let function $Z_{\epsilon,\theta}(u)$ possess the following properties: given a compact set $K \subset \Theta$ to which numbers $M_1 > 0, m_1 \ge 0$ and functions $g_\epsilon^K(y) \in \mathbf{G}$ correspond such that*

1. For some $\alpha > 0$ and all $\theta \in K$,

$$\sup_{|u_1| \le R, |u_2| \le R} |u_2 - u_1|^{-\alpha} \mathbb{E}_\theta^{(\epsilon)} |Z_{\epsilon,\theta}^{1/2}(u_2) - Z_{\epsilon,\theta}^{1/2}(u_1)|^2 \le M_1(1 + R^{m_1}). \tag{1.49}$$

2. For all $u \in U_\epsilon, \theta \in K$,

$$\mathbb{E}_\theta^{(\epsilon)} Z_{\epsilon,\theta}^{1/2}(u) \le e^{-g_\epsilon(|u|)}. \tag{1.50}$$

Let $\{\tilde{t}_\epsilon\}$ be a family of estimators. Bayesian with respect to a prior density q—continuous and positive on K and possessing in Θ a polynomial majorant—and a loss function $W_\epsilon(u, v) = l((\varphi(\epsilon))^{-1}(u - v))$ where

1. $l \in \mathbf{W}_p'$

2. there exist numbers $\gamma > 0, H_0 \ge 0$ such that for $H \ge H_0$,

$$\sup\{l(u) : |u| \le H^\gamma\} - \inf\{l(u) : |u| \ge H\} \le 0. \tag{1.51}$$

Then for any N,

$$\lim_{H \to \infty, \epsilon \to 0} H^N \sup_{\theta \in K} P_\theta^{(\epsilon)}(|(\varphi(\epsilon))^{-1}(\tilde{t}_\epsilon - \theta)| > H) = 0. \tag{1.52}$$

If in addition $l(u) = \tau(|u|)$ then for all ϵ sufficiently small, $0 < \epsilon < \epsilon_0$,

$$\sup_{\theta \in K} P_\theta^{(\epsilon)}(|(\varphi(\epsilon))^{-1}(\tilde{t}_\epsilon - \theta)| > H) \le B_0 e^{-b_0 g_\epsilon(H)}. \tag{1.53}$$

Corollary 1.4.3. *Under the conditions of Theorem 1.4.2 for any function $w \in \mathbf{W}_p$,*

$$\limsup_{\epsilon \to 0} \mathbb{E}_\theta^{(\epsilon)} w((\varphi(\epsilon))^{-1}(\tilde{t}_\epsilon - \theta)) < \infty. \tag{1.54}$$

1.4.2 Some Basic Theorems and Lemmas

In this subsection we shall consider a sequence of statistical experiments $(\mathscr{X}^n, \mathcal{X}^n, P_\theta^n, \theta \in \Theta)$ where Θ is an open subset in $\mathbb{R}^k$ generated by a sequence of homogeneous independent observations $X_1, X_2, \ldots, X_n$ with common density $f(x;\theta)$ with respect to measure ν. Based on the results of the preceding section we shall continue the study of consistent estimators $\{T_n\}$ for θ with the aim of determining the rate of convergence of $\{T_n\}$ to θ for certain classes of estimators $\{T_n\}$. It will be shown that the rate of convergence depends on the asymptotic behavior of the Hellinger distance

$$r_2(\theta;\theta+h) = \left(\int_{\mathscr{X}} |f^{1/2}(x,\theta+h) - f^{1/2}(x,\theta)|^2 d\nu\right)^{1/2}, \tag{1.55}$$

as $h \to 0$.

Theorem 1.4.3. *Let Θ be a bounded interval in $\mathbb{R}^1$, $f(x;\theta)$ be a continuous function of θ on $\bar{\Theta}$ for ν-almost all x and let the following conditions be satisfied:*

1. *There exists a number $\alpha > 1$ such that*

$$\sup_{\theta\in\Theta}\sup_h |h|^{-\alpha} r_2^2(\theta;\theta+h) = A < \infty \tag{1.56}$$

2. *For any compact set K there corresponds a positive number $a(K) = a > 0$ such that*

$$r_2^2(\theta;\theta+h) \geq \frac{a|h|^\alpha}{1+|h|^\alpha}, \theta \in K. \tag{1.57}$$

Then the maximum likelihood estimator $\hat{\theta}_n$ is defined, is consistent, and

$$\sup_{\theta\in K} P_\theta(n^{1/\alpha}|\hat{\theta}_n - \theta| > H) \leq B_0 e^{-b_0 a H^\alpha}, \tag{1.58}$$

where the positive constants B_0, b_0 do not depend on K.

A version of Theorem 1.4.3 for an unbounded interval can be stated as follows.

Theorem 1.4.4. *Let Θ be an interval in $\mathbb{R}^1$ (not necessarily bounded), $f(x;\theta)$ be a continuous function of θ for ν-almost all x and let the following conditions be satisfied: there exist numbers $\alpha > 1$ and $\gamma > 0$ such that*

1. *$\sup_{|\theta|<R}\sup_h |h|^{-\alpha} r_2^2(\theta;\theta+h) \leq M(1+R^m)$, where M, m are constants.*
2. *For any compact set $K \subset \Theta$ there corresponds a number $a(K) = a$ such that*

$$r_2^2(\theta;\theta+h) \geq \frac{a|h|^\alpha}{1+|h|^\alpha}, \theta \in K. \tag{1.59}$$

3. *For any compact set $K \subset \Theta$ there corresponds a number $c = c(K)$ such that*

$$\sup_{|h|>R} \int_{\mathscr{X}} f^{1/2}(x,\theta) f^{1/2}(x,\theta+h) d\nu \leq cR^{-\gamma}, \theta \in K. \tag{1.60}$$

Then a maximum likelihood estimator is defined, is consistent, and for all $n \geq n_0(\gamma)$ the following inequalities are satisfied: for any number $\Lambda > 0$ there correspond positive constants B_i, b_i (dependent also on $\alpha, M, m, a, \gamma, c$) such that

$$\sup_{\theta\in\Theta} P_\theta(n^{1/\alpha}|\hat{\theta}_n - \theta| > H) \leq \begin{cases} B_1 e^{-b_1 a H^\alpha} & H < \Lambda n^{1/\alpha} \\ B_2 e^{-b_2 c n \ln \frac{H}{n^{1/\alpha}}} & H \geq \Lambda n^{1/\alpha} \end{cases} \tag{1.61}$$

Remark 1.4.1. *The last inequality in Theorem 1.4.4 is equivalent to the following:*

$$\sup_{\theta\in\Theta} P_\theta(|\hat{\theta}_n - \theta| > \delta) \leq \begin{cases} B_1 e^{-b_1 a n \delta^\alpha} & \delta < \Lambda \\ B_2 e^{-b_2 c n \ln \delta} & \delta \geq \Lambda \end{cases} \tag{1.62}$$

We now present two theorems on Bayesian estimators. In these theorems Θ is an open subset of $\mathbb{R}^k$. Bayesian estimators are constructed with respect to a positive continuous prior density q on Θ possessing a polynomial majorant on Θ. These theorems are analogous to Theorems 1.4.3 and 1.4.4. The first deals with case of a bounded set Θ and the second with an unbounded one. However, transition to Bayesian estimators allows us to substantially weaken the restrictions on f whenever the dimension k of the parameter set is greater than 1.

Theorem 1.4.5. *Let Θ be an open bounded set in $\mathbb{R}^k$. Let the following conditions be satisfied: there exists a number $\alpha > 0$ such that*

1. $\sup_{\theta\in\Theta} |h|^{-\alpha} r_2^2(\theta;\theta+h) = A < \infty$
2. *For any compact set K there corresponds a positive number $a(K) = a > 0$ such that*

$$r_2^2(\theta;\theta+h) \geq \frac{a|h|^\alpha}{1+|h|^\alpha}, \theta \in K. \tag{1.63}$$

Finally let $\{\tilde{t}_n\}$ be a sequence of estimators which are Bayesian with respect to prior density q and the loss function $W_n(u,v) = l(n^{1/\alpha}|u-v|)$, where $l \in \mathbf{W}'_p$. Then

$$\sup_{\theta\in K} P_\theta(n^{1/\alpha}|\tilde{t}_n - \theta| > H) \leq Be^{-baH^\alpha}. \tag{1.64}$$

Here the constants B, b are positive and depend only on A, α and the diameter of Θ.

Theorem 1.4.6. *Let Θ be an open set in $\mathbb{R}^k$. Let the following conditions be satisfied: there exist numbers $\alpha > 0, \gamma > 0$ such that*

1. $$\sup_{|\theta|\leq R} |h|^{-\alpha} r_2^2(\theta;\theta+h) \leq M_1(1+R^{m_1}), \tag{1.65}$$

   where M_1, m_1 are positive numbers.
2. *For any compact set $K \subset \Theta$ there corresponds a positive number $a(K) = a > 0$ such that*

$$r_2^2(\theta;\theta+h) \geq \frac{a|h|^\alpha}{1+|h|^\alpha}, \theta \in K. \tag{1.66}$$

3. *For any compact set $K \subset \Theta$ there corresponds a positive number $c(K) = c > 0$ such that*

$$\sup_{|h|>R} \int_{\mathscr{X}} f^{1/2}(x;\theta) f^{1/2}(f;\theta+h) d\nu \leq cR^{-\gamma}, \theta \in K. \tag{1.67}$$

Finally, let $\{\tilde{t}_n\}$ be a sequence of estimators which are Bayesian with respect to prior density q and the loss function $W_n(u,v) = l(n^{1/\alpha}|u-v|)$, where $l \in \mathbf{W}_p$. Then, for any number $\Lambda > 0$ there correspond positive constants B_1, B_2, b_1, b_2 such that for all $n > n_0(m_1, \gamma)$,

$$\sup_{\theta\in K} P_\theta(n^{1/\alpha}|\tilde{t}_n - \theta| > H) \leq \begin{cases} B_1 e^{-b_1 a H^\alpha} & H \leq \Lambda n^{1/\alpha} \\ B_2 e^{-b_2 cn \ln \frac{H}{n^{1/\alpha}}} & H > \Lambda n^{1/\alpha} \end{cases} \tag{1.68}$$

Theorem 1.4.7. *Let conditions 1-3 of Theorem 1.4.6 be satisfied. Furthermore, let $\{\tilde{t}_n\}$ be a sequence of estimators which are Bayesian with respect to prior density q and the loss function $W_n(u,v) = l(n^{1/\alpha}|u-v|)$, where $l \in \mathbf{W}'_p$, and, moreover, for some $\gamma_0 > 0, H_0 > 0, H > H_0$ the inequality*

$$\sup\{l(u) : |u| \leq H^{\gamma_0}\} - \inf\{l(u) : |u| > H\} \leq 0 \tag{1.69}$$

be fulfilled. Then for all $n \geq n_0$,

$$\sup_{\theta\in K} P_\theta(n^{1/\alpha}|t_n - \theta| > H) \leq B_N H^{-N}, \tag{1.70}$$

whatever the number $N > 0$ is.

One can formulate for maximum likelihood estimators theorems which are similar to Theorems 1.4.3 and 1.4.4 also in the case when $\Theta \subset \mathbb{R}^k, k > 1$; in these cases in place of the Hellinger distance r_2 one should take the distance

$$r_m(\theta; \theta + h) = \left(\int_{\mathscr{X}} |f^{1/2}(x, \theta + h) - f^{1/2}(x, \theta)|^m d\nu \right)^{1/m}, \tag{1.71}$$

$m > k$ and require that

$$r_m^m(\theta; \theta + h) \geq \frac{a|h|^\alpha}{1 + |h|^\alpha}, \alpha > k. \tag{1.72}$$

We shall not dwell on this point in detail, but present the following result which is somewhat different from the preceding theorems.

Theorem 1.4.8. *Let Θ be a bounded open convex set in $\mathbb{R}^k$, $f(x; \theta)$ be a continuous function of θ on $\bar{\Theta}$ for ν-almost all x and let the following conditions be satisfied:*

1. *There exists a number $\alpha > 0$ such that*

$$\int_{\mathscr{X}} \sup_{\theta_1, \theta_2 \in \Theta} \frac{(f^{1/2}(x; \theta_2) - f^{1/2}(x; \theta_1))^2}{|\theta_1 - \theta_2|^\alpha} d\nu = A < \infty. \tag{1.73}$$

2. *There exists a number $\beta > 0$ such that for all $\theta \in \Theta$,*

$$r_2^2(\theta; \theta + h) \geq \frac{a(\theta)|h|^\beta}{1 + |h|^\beta}, \tag{1.74}$$

where $a(\theta) > 0$.

Then the maximum likelihood estimator $\hat{\theta}_n$ is defined, it is consistent and for any $\lambda \leq 1/\beta$,

$$P_\theta(n^\lambda |\hat{\theta}_n - \theta| > H) \leq B n^{\beta - 1 - (2\alpha)^{-1}} e^{-n^{1-\lambda\beta} b a(\theta) H^\beta}. \tag{1.75}$$

Here the positive constants B, b depend only on A, α, β and the diameter of the set Θ.

1.4.3 Examples

Example 1.4.1. *Let $(X_1, X_2, \ldots, X_n)$ be a sample from the normal distribution $\mathcal{N}(a, \sigma^2)$, where $a, -\infty < a < \infty$ is an unknown parameter. The conditions of Theorem 1.4.4 are fulfilled; here $\alpha = 2$. The maximum likelihood estimator for a is $\bar{X}$ and for this estimator the inequality of Theorem 1.4.4 is satisfied with $\alpha = 2$. Since $\mathcal{L}(\bar{X}) = \mathcal{N}(a, \sigma^2/n)$, this inequality can be substantially improved.*

Example 1.4.2. *Now let the variables $X_i \sim \mathrm{Bern}(p), 0 < p < 1$. The conditions of Theorem 1.4.3 are satisfied if $0 < p_0 \leq p \leq p_1 < 1$; here $\alpha = 2$. The maximum likelihood estimator for p once again is $\bar{X}$. Since the set of values the parameter p can take is compact, we have*

$$P_p(\sqrt{n}|\bar{X} - p| > H) \leq B_1 e^{-b_1 H^2}. \tag{1.76}$$

This bound can be substantially refined using Hoeffding's inequality. This inequality applied in our case asserts that

$$P_p(\sqrt{n}|\bar{X} - p| > H) \leq 2e^{-2H^2}, 0 < p < 1. \tag{1.77}$$

Example 1.4.3. *Let θ be a location parameter. This means that $\mathscr{X} = \Theta = \mathbb{R}^k$ and the distribution P_θ in $\mathbb{R}^k$ possesses a density of the form $f(x - \theta)$ with respect to the Lebesgue measure λ. In this case we always have*

$$\int_{\mathbb{R}^k} |f^{1/2}(x + h) - f^{1/2}(x)|^2 dx = r(h) \geq \frac{a|h|^2}{1 + |h|^2}, a > 0. \tag{1.78}$$

Indeed, denote by $\varphi(t)$ the Fourier transform of the function $f^{1/2}(x)$. It follows from the Parseval equality that

$$\liminf_{|h| \to 0} \frac{r(h)}{|h|^2} = \liminf_{|h| \to 0} \frac{4}{|h|^2} \int \sin^2(1/2\langle t, h\rangle) |\varphi(t)|^2 dt = \inf_{|h|=1} \int \langle t, h\rangle^2 |\varphi(t)|^2 dt = \mu \geq 0. \tag{1.79}$$

However, $\lambda\{t : |\varphi(t)| > 0\} > 0$ so that the quadratic form on the right hand side is positive definite and $\mu > 0$ (the case $\mu = \infty$ is not excluded.) Moreover, it is clear that $r(h) \to 2$ as $|h| \to \infty$.

Example 1.4.4. *Let $(X_1, X_2, \dots, X_n)$ be a sample from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, where θ is the parameter to be estimated. It is a particular case of the location model with $f(x) = 1$ for $|x| < 1/2$ and $f(x) = 0$ for $|x| > 1/2$. It is easy to see that*

$$\int |f^{1/2}(x+h) - f^{1/2}(x)|^2 dx = 2h, \tag{1.80}$$

so that $\alpha = 1$. The statistic $t_n = (X_{\max} + X_{\min})/2$ is a Bayesian estimator with respect to the quadratic loss function and the uniform prior distribution (Pitman's estimator) and in view of Theorem 1.4.5,

$$\mathbb{P}_\theta(n|t_n - \theta| > H) \le B_1 e^{-b_1 H}, H < n/2. \tag{1.81}$$

Taking into account the exact form of the distribution of statistic t_n, one can write for $n \ge 2$,

$$\mathbb{P}_\theta(n|t_n - \theta| > H) = 2\int_H^{n/2} \left(1 - \frac{2u}{n}\right)^{n-1} du \le 2\int_H^\infty e^{-u} du = 2e^{-H}. \tag{1.82}$$

Example 1.4.5. *Let $(X_1, X_2, \dots, X_n)$ be a sample from a population with a Gamma distribution with unknown location parameter θ, i.e., let X_j possess on the real line the probability density equal to*

$$\frac{(x-\theta)^{p-1} e^{\theta - x}}{\Gamma(p)}, \theta \le x < \infty, \tag{1.83}$$

where $p > 0$ is known. Once again the conditions of Example 1.4.3 are satisfied and

$$r(h) \sim \begin{cases} h^2 & p > 2 \\ h^p & p < 2 \end{cases} \tag{1.84}$$

Example 1.4.6. *Experiments with a finite Fisher information amount serve as a source for a large number of examples. We will discuss this case in more details later.*

1.5 Lower Bounds on the Risk Function

Let a repeated sample of size n, $X_1, X_2, \dots, X_n$ with density function $f(x;\theta)$ with respect to some measure ν be given. AS usual, let $\theta \in \Theta \subset \mathbb{R}^k$. If $T_n = T_n(X_1, X_2, \dots, X_n)$ be an estimator for θ, we set

$$S^{(m)}(T_n; \theta) = \mathbb{E}_\theta |T_n - \theta|^m. \tag{1.85}$$

Theorem 1.5.1. *Let for all $\theta \in \Theta$,*

$$r_2^2(\theta; \theta') = r_2^2 = \int_{\mathscr{X}} (\sqrt{f(x;\theta)} - \sqrt{f(x;\theta')})^2 \nu(dx) \le K_1(\theta)|\theta - \theta'|^\alpha, K_1 > 0, \alpha > 0 \tag{1.86}$$

as long as $|\theta - \theta'| \le h_1(\theta), h_1 > 0$. Denote by j a vector of unit length in $\mathbb{R}^k$. Then for any sequence of estimators $\{T_n\}$,

$$\liminf_{n\to\infty} n^{m/\alpha} \left(S^{(m)}(T_n;\theta) + S^{(m)}(T_n; \theta + (2nK_1(\theta))^{-1/\alpha} j)\right) > 0 \tag{1.87}$$

for all $\theta \in \Theta$ and $m \ge 1$.

Theorem 1.5.1 establishes the asymptotic lower bound on the risk function for arbitrary estimators; a comparison of this result with Theorems 1.4.3-1.4.7 shows that Bayesian estimators and maximum likelihood estimators are optimal as far as the order of decrease of the risk is concerned under power (polynomial) loss functions.

If we denote

$$\tilde{S}^{(m)}(T;\theta) = \inf_{|j|=1} S^{(m)}(T; \theta + (2nK_1(\theta))^{-1/\alpha} j), \tag{1.88}$$

and if $n > (2K_1 h_1^\alpha)^{-1}$, then we could show

$$\inf_{T_n} \left(S^{(1)}(T_n;\theta) + \tilde{S}^{(1)}(T_n;\theta)\right) \ge \frac{2^{-5}}{(2K_1 n)^{1/\alpha}}. \tag{1.89}$$

It implies that

$$\inf_{T_n} \left(S^{(m)}(T_n;\theta) + \tilde{S}^{(m)}(T_n;\theta)\right) \ge 2^{-m+1} \frac{2^{-5m}}{(2K_1 n)^{m/\alpha}}. \tag{1.90}$$

In the last step we have used the elementary inequality

$$a^m + b^m \ge 2^{1-m}(a+b)^m, a > 0, b > 0, m \ge 1. \tag{1.91}$$

1.6 Regular Statistical Experiments. The Cramer-Rao Inequality

1.6.1 Regular Statistical Experiments

Consider the statistical experiment $(\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta)$, where Θ is an open subset of $\mathbb{R}^k$ and all the measures P_θ are absolutely continuous with respect to measure ν on $\mathcal{X}$; moreover $\frac{dP_\theta}{d\nu} = p(x;\theta)$.

Let $p(x;\theta)$ be a continuous function of θ on Θ for ν-almost all x. Assume that for ν-almost all x the density $p(x;u)$ is differentiable at the point $u = \theta$ and all the integrals

$$I_{jj}(\theta) = \int_{\mathscr{X}} \left| \frac{\partial p(x;\theta)}{\partial \theta_j} \right|^2 \frac{\nu(dx)}{p(x;\theta)} < \infty. \tag{1.92}$$

Here the integration is carried out over those x for which $p(x;\theta) \neq 0$, so that

$$I_{jj}(\theta) = \mathbb{E}_\theta \left[\frac{\partial p(X;\theta)}{\partial \theta_j} \frac{1}{p(X;\theta)} \right]^2. \tag{1.93}$$

Let us agree to interpret all the integrals of the form

$$\int_{\mathscr{X}} \frac{a(x;\theta)}{p(x;\theta)} \nu(dx) \tag{1.94}$$

as the integrals over the set $\{x : p(x;\theta) \neq 0\}$, i.e. as

$$\mathbb{E}_\theta \left[\frac{a(X;\theta)}{p^2(X;\theta)} \right]. \tag{1.95}$$

The Cauchy-Schwarz inequality implies that together with $I_{jj}(\theta)$ all the following integrals

$$I_{ij}(\theta) = \int_{\mathscr{X}} \frac{\partial p(x;\theta)}{\partial \theta_i} \frac{\partial p(x;\theta)}{\partial \theta_j} \frac{\nu(dx)}{p(x;\theta)}, i, j = 1, 2, \ldots, k \tag{1.96}$$

are convergent. The matrix $I(\theta)$ whose ij-th entry is $I_{ij}(\theta)$ is called *Fisher's information matrix*.

Denote by $L_2(\nu)$ the Hilbert space of functions, square integrable with respect to measure ν on $\mathscr{X}$ with the scalar product $(\varphi, \psi)_\nu = \int_{\mathscr{X}} \varphi(x)\psi(x)\nu(dx)$ and the norm $\|\varphi\|_\nu$. Note that Fisher's information amount can be written as

$$I(\theta) = 4 \int_{\mathscr{X}} \left(\frac{\partial}{\partial \theta} p^{1/2}(x;\theta) \right)^2 d\nu = 4 \left\| \frac{\partial}{\partial \theta} p^{1/2} \right\|_\nu^2, \tag{1.97}$$

and the information matrix as

$$I(\theta) = 4 \int_{\mathscr{X}} \frac{\partial}{\partial \theta} p^{1/2}(x;\theta) \left(\frac{\partial}{\partial \theta} p^{1/2}(x;\theta) \right)^T d\nu. \tag{1.98}$$

We say that experiment possesses at point $\theta \in \Theta$ *Fisher's finite information* if the function $p^{1/2}(\cdot;u)$ is differentiable at point $u = \theta$ in $L_2(\nu)$.

Set, for brevity, $p^{1/2}(x;\theta) = g(x;\theta)$. The differentiability of function g in $L_2(\nu)$ means the following. There exists a function $\psi : \mathscr{X} \times \Theta \to \mathbb{R}^k$ such that

$$\int_{\mathscr{X}} |\psi(x;u)|^2 d\nu = \|\psi(\cdot;u)\|_\nu^2 < \infty, \tag{1.99}$$

$$\int_{\mathscr{X}} |g(x;\theta+h) - g(x;\theta) - \langle \psi(x;\theta), h \rangle|^2 d\nu = o(|h|^2), h \to 0. \tag{1.100}$$

The matrix

$$I(\theta) = 4 \int_{\mathscr{X}} \psi(x;\theta)(\psi(x;\theta))^T d\nu \tag{1.101}$$

will be called as before Fisher information matrix, and the integral

$$I(\theta) = 4 \int_{\mathscr{X}} |\psi(x;\theta)|^2 d\nu, \Theta \subset \mathbb{R}^1, \tag{1.102}$$

will be called Fisher's information amount (measure).

If in addition the density $p(x;u)$ is differentiable at point θ, then of course

$$\psi(x;\theta) = \frac{\partial}{\partial\theta} p^{1/2}(x;\theta) \tag{1.103}$$

and, for example,

$$I(\theta) = 4\int_{\mathscr{X}} |\psi(x;\theta)|^2 d\nu = \int_{\mathscr{X}} \frac{\left|\frac{\partial}{\partial\theta} p(x;\theta)\right|^2}{p(x;\theta)} d\nu. \tag{1.104}$$

Since our intention is to retain the classical expression also in the case when then density p is not differentiable in the usual sense we shall adhere to the following convention. Let $c(s)$ be a differentiable function from $\mathbb{R}^1$ into $\mathbb{R}^1$. Formal differentiation yields

$$\frac{\partial}{\partial\theta} c(p(x;\theta)) = c'(p(x;\theta)) \frac{\partial}{\partial\theta} p(x;\theta) \tag{1.105}$$

$$= 2c'(p(x;\theta)) p^{1/2}(x;\theta) \frac{\partial}{\partial\theta} p^{1/2}(x;\theta) \tag{1.106}$$

$$= 2c'(p(x;\theta)) p^{1/2}(x;\theta)\psi(x;\theta). \tag{1.107}$$

The expression on the right hand side is defined provided the function $g(x;\theta) = p^{1/2}(x;\theta)$ is differentiable in the mean square and in this case we shall set by definition

$$\frac{\partial}{\partial\theta} c(p) = 2c'(p) p^{1/2}\psi. \tag{1.108}$$

For example, we have

$$\frac{\partial}{\partial\theta} p(x;\theta) = 2p^{1/2}(x;\theta)\psi(x;\theta) \tag{1.109}$$

$$\frac{\partial}{\partial\theta} \ln p(x;\theta) = 2p^{-1/2}(x;\theta)\psi(x;\theta) \tag{1.110}$$

and so on. Utilizing this convention in the case of an experiment with finite Fisher information at point θ we can then use for Fisher information matrix the notation

$$I(\theta) = \int_{\mathscr{X}} \frac{\partial p}{\partial\theta} \left(\frac{\partial p}{\partial\theta}\right)^T p^{-1}(x;\theta) d\nu. \tag{1.111}$$

We shall adhere to this notation below.

Moreover, below we shall always use the very same notation $\frac{\partial}{\partial\theta}$ for the ordinary derivatives as well as for the derivatives in the mean square. In general in order to construct a well developed theory of experiments wit finite Fisher information it is necessary to impose certain smoothness conditions on the family $\{p(x;\theta)\}$. We shall utilize the following definition:

Definition 1.6.1 (Regular Statistical Experiment)**.** *A statistical experiment $\mathscr{E}$ is called* regular *in Θ if*

1. *$p(x;\theta)$ is a continuous function on Θ for ν-almost all x*
2. *$\mathscr{E}$ possess finite Fisher information at each point $\theta \in \Theta$*
3. *the function $\psi(\cdot;\theta)$ is continuous in the space $L_2(\nu)$.*

Note that conditions 1-3 are not totally independent, for example, if $\Theta \subset \mathbb{R}^1$ then conditions 2-3 imply condition 1. Namely if the density $p(x;\theta)$ satisfies conditions 2-3 it can be modified on sets of ν-measure zero (these sets may depend on θ) in such a manner that it becomes a continuous function of θ. Indeed, the measure ν may be considered to be a probability measure so that $p^{1/2}(x;\theta)$ is a random function of θ satisfying the condition

$$\mathbb{E}\left(p^{1/2}(\cdot;\theta+h) - p^{1/2}(\cdot;\theta)\right)^2 \le Bh^2, \tag{1.112}$$

where B is a constant. In view of Kolmogorov's continuity criterion, there exists a modification of $p^{1/2}(\cdot;\theta)$ which is continuous with probability one.

It should be verified that the "regularity" property of the experiment and its information matrix do not depend on the choice of the measure ν dominating the family $\{P_\theta\}$. Let μ be another such measure and let $q(x;\theta) = \frac{dP_\theta}{d\mu}$. Then $q(x;\theta) = p(x;\theta)\gamma(x)$ where γ does not depend on θ. Indeed, we have

$$\frac{dP_\theta}{d(\mu+\nu)}(x) = p(x;\theta)\frac{d\nu}{d(\mu+\nu)}(x) = q(x;\theta)\frac{d\mu}{d(\mu+\nu)}(x). \tag{1.113}$$

Consequently, functions p, q either both satisfy conditions in Definition 1.6.1 or both don't satisfy these conditions. Moreover,

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\sqrt{p}\left(\frac{\partial}{\partial\theta}\sqrt{p}\right)^T\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\sqrt{q}\left(\frac{\partial}{\partial\theta}\sqrt{q}\right)^T\right] \tag{1.114}$$

The following lemma is a simple corollary of Definition 1.6.1.

Lemma 1.6.1. *Let $\mathscr{E} = \{\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta\}$ be a regular experiment. Then*

1. *The matrix $I(\theta)$ is continuous on Θ (i.e. all of the functions $I_{ij}(\theta)$ are continuous on Θ)*
2. *The integrals $I_{ij}(\theta)$ converge uniformly on an arbitrary compact set $K \subset \Theta$, i.e.*

$$\lim_{A\to\infty}\sup_{\theta\in K}\int_{\mathscr{X}}\frac{\left|\frac{\partial p}{\partial\theta_i}\frac{\partial p}{\partial\theta_j}\right|}{p(x;\theta)}I\left(x : \frac{\left|\frac{\partial p}{\partial\theta_i}\frac{\partial p}{\partial\theta_j}\right|}{p(x;\theta)} > A\right)d\nu = 0. \tag{1.115}$$

3. *If the interval $\{t + su | 0 \le s \le 1\}$ belongs to Θ then*

$$g(x;t+u) - g(x;t) = \int_0^1 \left\langle \frac{\partial g}{\partial t}(x;t+su), u\right\rangle ds \tag{1.116}$$

in $L_2(\nu)$ (i.e. the integral on the right hand side is the limit in $L_2(\nu)$ of the Riemann sums).

Lemma 1.6.2. *Let $(\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta)$ be a regular experiment and the statistic $T : \mathscr{X} \to \mathbb{R}^1$ be such that the function $\mathbb{E}_u T^2$ be bounded in the neighbourhood of a point $\theta \in \Theta$. Then the function $\mathbb{E}_u T$ is continuously differentiable in a neighborhood of θ and*

$$\frac{\partial}{\partial u}\mathbb{E}_u T = \frac{\partial}{\partial u}\int_{\mathscr{X}} T(x)p(x;u)\nu(dx) = \int_{\mathscr{X}} T(x)\frac{\partial}{\partial u}p(x;u)\nu(dx). \tag{1.117}$$

Remark 1.6.1. *It follows from (1.117) and the Fubini theorem that for all bounded T we have*

$$\int_{\mathscr{X}} T(x)[p(x;\theta+u) - p(x;\theta)]d\nu = \int_{\mathscr{X}} T(x)\int_0^1 \frac{\partial}{\partial\theta}p(x;\theta+su)ds, \tag{1.118}$$

provided the interval $\{\theta + su : 0 \le s \le 1\} \subset \Theta$. Consequently for ν-almost all x,

$$p(x;\theta+u) - p(x;\theta) = \int_0^1 \left\langle \frac{\partial}{\partial\theta}p(x;\theta+su), u\right\rangle ds. \tag{1.119}$$

Setting $T(x) \equiv 1$, we obtain

$$\int_{\mathscr{X}} \frac{\partial}{\partial\theta}p(x;\theta)d\nu = \frac{\partial}{\partial\theta}\int_{\mathscr{X}} p(x;\theta)d\nu = \frac{\partial}{\partial\theta}1 = 0. \tag{1.120}$$

Theorem 1.6.1. *Let $\mathscr{E}_1 = \{\mathscr{X}_1, \mathcal{X}_1, P_{\theta_1}, \theta \in \Theta\}$ and $\mathscr{E}_2 = \{\mathscr{X}_2, \mathcal{X}_2, P_{\theta_2}, \theta \in \Theta\}$ be regular statistical experiments with Fisher informations $I_1(\theta), I_2(\theta)$ respectively. Then the experiment $\mathscr{E} = \mathscr{E}_1 \times \mathscr{E}_2$ is also regular and moreover*

$$I(\theta) = I_1(\theta) + I_2(\theta), \tag{1.121}$$

here $I(\theta)$ is the Fisher information of experiment $\mathscr{E}$.

Theorem 1.6.2. *Let $\mathscr{E} = \{\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta\}$ be a regular experiment. Let experiment $\{\mathscr{Y}, \mathcal{Y}, Q_\theta, \theta \in \Theta\}$ generated by the statistic $T : \mathscr{X} \to \mathscr{Y}$ be a part of the experiment $\mathscr{E}$. Then $\mathscr{E}$ possess finite Fisher information at all points $\theta \in \Theta$ and, moreover,*

$$\tilde{I}(\theta) \le I(\theta), \tag{1.122}$$

where $\tilde{I}(\theta), I(\theta)$ are Fisher information matrices for experiments $\{\mathscr{Y}, \mathcal{Y}, Q_\theta, \theta \in \Theta\}, \{\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta\}$, respectively. If the density $p(x;\theta)$ is a continuously differentiable function of θ for ν-almost all x, then the inequality (1.122) is valid for all $\theta \in \Theta$ if and only if T is a sufficient statistic for $\{P_\theta\}$.

1.6.2 The Cramer-Rao Inequality

Theorem 1.6.3 (The Cramer-Rao Inequality)**.** *Let $(\mathscr{X}, \mathcal{X}, P_\theta, \theta \in \Theta)$ be a regular experiment with an information matrix $I(\theta) > 0, \Theta \subset \mathbb{R}^k$. If the statistic $T = (T_1, T_2, \ldots, T_k) : \mathscr{X} \to \mathbb{R}^k$ is such that the risk $\mathbb{E}_u|T-u|^2$ is bounded in the neighborhood of the point $\theta \in \Theta$, then the bias*

$$d(u) = \mathbb{E}_u T - u \tag{1.123}$$

is continuously differentiable in a neighborhood of θ and the following matrix inequality is satisfied:

$$\mathbb{E}_\theta (T-\theta)(T-\theta)^T \geq \left(J + \frac{\partial d(\theta)}{\partial \theta}\right) I^{-1}(\theta) \left(J + \frac{\partial d(\theta)}{\partial \theta}\right)^T + d(\theta) d(\theta)^T, \tag{1.124}$$

where J is the identity matrix.

In particular, if $\Theta \subset \mathbb{R}^1$ and $I(\theta) > 0$, then

$$\mathbb{E}_\theta |T-\theta|^2 \geq \frac{(1+d'(\theta))^2}{I(\theta)} + d^2(\theta). \tag{1.125}$$

The Cramer-Rao inequality allows us to show that the existence of consistent estimators is connected, in a certain sense, with unlimited inflow of Fisher information. Consider, for example, a sequence of regular statistical experiments $\mathscr{E}^{(n)}$ where the parameter set Θ is a bounded interval (a, b) on the real line. A sequence of consistent estimators $\{T_n\}$ for $\theta \in \Theta$ may exist only if for an arbitrary $[\alpha, \beta] \subset (a, b)$,

$$\inf_{[\alpha,\beta]} I(\theta; \mathscr{E}^{(n)}) \to \infty \tag{1.126}$$

as $n \to \infty$.

Indeed, let the sequence $T_n \to \theta$ in probability for all $\theta \in \Theta$. Since the set Θ is bounded one can assume that $|T_n| \leq \max(|a|+1, |b|+1)$. Therefore $\mathbb{E}_\theta |T_n - \theta|^2 \to 0$ for all θ. Assume that for some interval $[\alpha, \beta]$, $I(\theta; \mathscr{E}^{(n)}) < M$ for all $\theta \in [\alpha, \beta]$. In view of the Cramer-Rao inequality,

$$\frac{(1+d_n'(\theta))^2}{M} + d_n^2(\theta) \to 0, \theta \in [\alpha, \beta]. \tag{1.127}$$

Here $d_n(\theta)$ is the bias of the estimator T_n. In this case $d_n(\theta) \to 0, \theta \in \Theta, d_n'(\theta) \to -1$ for all $\theta \in [\alpha, \beta]$. However, in view of Lemma 1.6.2,

$$|d_n'(\theta)| = \left| \int T \frac{\partial}{\partial \theta} \frac{dP_\theta^{(n)}}{d\nu} d\nu - 1 \right| \leq \left(\mathbb{E}_\theta T^2\right)^{1/2} I^{1/2}(\theta; \mathscr{E}^{(n)}) + 1 \leq \sqrt{M}(|a|+|b|+2)^{1/2} + 1. \tag{1.128}$$

Lebesgue's theorem yields

$$\lim_n \int_\alpha^\tau d_n'(u) du = \int_\alpha^\tau \lim_n d_n'(u) du = -(\tau - \alpha), \alpha < \tau \leq \beta. \tag{1.129}$$

However, in this case

$$d_n(\tau) = \int_\alpha^\tau d_n'(u) du + d_n(\alpha) \to -(\tau - \alpha) \neq 0. \tag{1.130}$$

The contradiction obtained proves our assertion.

The Cramer-Rao inequality is exact in the sense that there are cases when equality is attained. However, equality in the Cramer-Rao inequality is possible only for some special families of distributions. We shall exemplify this using a family with a one-dimensional parameter. In view of Lemma 1.6.2, for the estimator T,

$$d'(\theta) + 1 = \frac{d}{d\theta} \mathbb{E}_\theta T = \int_{\mathscr{X}} (T - \mathbb{E}_\theta T) \frac{\partial p(x;\theta)}{\partial \theta} \nu(dx) \tag{1.131}$$

and from the Cauchy-Schwarz inequality we have

$$\mathbb{E}_\theta (T - \mathbb{E}_\theta T)^2 \geq \frac{(1+d'(\theta))^2}{I(\theta)}. \tag{1.132}$$

Moreover, $\mathbb{E}_\theta (T - \mathbb{E}_\theta T)^2 = \mathbb{E}_\theta (T-\theta)^2 - d^2(\theta)$. Thus equality in the Cramer-Rao inequality is attainable only if

$$\mathbb{E}_\theta (T - \mathbb{E}_\theta T)^2 = \frac{(1+d'(\theta))^2}{I(\theta)} = \frac{(\partial \mathbb{E}_\theta T / \partial \theta)^2}{I(\theta)}. \tag{1.133}$$

As it is known, the equality in the Cauchy-Schwarz inequality is satisfied only if

$$T - \mathbb{E}_\theta T = k(\theta)\frac{\partial}{\partial\theta}\ln p(x;\theta). \tag{1.134}$$

Analogous results are also valid for a multi-dimensional parameter.

Estimators for which the equality sign holds in the Cramer-Rao inequality are called *efficient*. They possess a number of nice properties. First, it follows from Theorem 1.6.2 it is a sufficient estimator for θ. Second, under certain conditions an efficient estimator is admissible with respect to a quadratic loss function. Namely, the following result is valid.

Theorem 1.6.4. *Let $\mathscr{E} = (\mathscr{X}, \mathcal{X}, P_\theta)$ be a regular experiment with one dimensional parameter set $\Theta = (a, b), -\infty \le a < b \le \infty$. If the integrals*

$$\int_a I(u)du, \quad \int^b I(u)du, \quad I(u) = I(u;\mathscr{E}), \tag{1.135}$$

are both divergent, then the efficient estimator T_0 for parameter θ is admissible with respect to a quadratic loss function.

Example 1.6.1. *In the one dimensional Gaussian mean estimation problem $\mathcal{N}(\theta, \sigma^2)$, we have the Fisher information*

$$I(\theta) = n\sigma^{-2}, \tag{1.136}$$

where σ^2 is the variance of variable X_i, and in view of Theorem 1.6.4 the efficient estimator $\bar{X}$ is also admissible. Note that if the parameter set Θ is not the whole real line, but, for example, the ray $(0, \infty)$, then the conditions of Theorem 1.6.4 are no longer satisfied (the integral $\int_0 n/\sigma^{-2}d\theta$ is convergent) and as before the efficient estimator $\bar{X}$ is inadmissible (and is worse than the estimator $\max(0, \bar{X})$.)

Example 1.6.2. *Let us consider the* $\mathrm{Bern}(\theta)$ *example. The sample average $\bar{X}$ is an efficient estimator, and $I(\theta) = \frac{1}{\theta(1-\theta)}$. Since*

$$\int_0 I(u)du = \int_0 \frac{1}{u(1-u)}du = \infty, \quad \int^1 I(u)du = \int^1 \frac{1}{u(1-u)}du = \infty, \tag{1.137}$$

we know $\bar{X}$ is an admissible estimator provided $\Theta = (0, 1)$.

1.6.3 Bounds on the Hellinger distance $r_2^2(\theta;\theta')$ in Regular Experiments

Consider a regular experiment and assume that Θ is a convex subset of $\mathbb{R}^k$.

Theorem 1.6.5. *The following inequality is valid:*

$$r_2^2(\theta;\theta+h) = \int_{\mathscr{X}} |p^{1/2}(x;\theta+h) - p^{1/2}(x;\theta)|^2 d\nu \le \frac{|h|^2}{4}\int_0^1 \mathrm{tr} I(\theta+sh)ds. \tag{1.138}$$

Moreover, if K is a compact subset of Θ then we have uniformly in K

$$\liminf_{h\to 0} |h|^{-2}\int_{\mathscr{X}} |p^{1/2}(x;\theta+h) - p^{1/2}(x;\theta)|^2 d\nu \ge \frac{1}{4}\inf_{|u|=1}\langle I(\theta)u, u\rangle. \tag{1.139}$$

Let a sequence of iid observations $X_1, X_2, \ldots, X_n$ generate regular experiments with probability density $f(x;\theta), \theta \in \Theta \subset \mathbb{R}^k$ and information matrix $I(\theta)$. Let Θ be a convex bounded set and the matrix $I(\theta)$ be strictly positive at each point θ. Assume also that for any compact $K \subset \Theta$ and all $\delta > 0$

$$\inf_{\theta\in K}\inf_{|h|\ge\delta}\int_{\mathscr{X}} |f^{1/2}(x;\theta+h) - f^{1/2}(x;\theta)|^2 d\nu > 0. \tag{1.140}$$

It follows from Theorem 1.6.5 and the last inequality that to any compact set $K \subset \Theta$ there correspond two constants $a, A > 0$ such that for all $\theta \in K$,

$$\frac{a|h|^2}{1+|h|^2} \le \int_{\mathscr{X}} |f^{1/2}(x;\theta+h) - f^{1/2}(x;\theta)|^2 d\nu \le A|h|^2. \tag{1.141}$$

This, together with the theorems on large deviations of maximum likelihood estimators and Bayesian estimators as well as lower bounds on the risk function, implies that the rate of convergence of the best estimators to the value θ is of order $n^{-1/2}$.

For example, if $\Theta \subset \mathbb{R}^1$ and $\hat{\theta}_n$ is a maximum likelihood estimator then it follows from Theorem 1.4.2 that for all $p > 0$,

$$\limsup_{n\to\infty}\sup_{\theta\in K}\mathbb{E}_\theta|\sqrt{n}(\hat{\theta}_n - \theta)|^p < \infty. \tag{1.142}$$

1.7 Approximating Estimators by Means of Sums of Independent Random Variables

Let $X_1, X_2, \dots$ be iid observations with a density $f(x;\theta)$ with respect to the measure $\nu, \Theta \subset \mathbb{R}^k$. If for $h \to 0$,

$$\int_{\mathscr{X}} |f^{1/2}(x;\theta+h) - f^{1/2}(x;\theta)|^2 d\nu \sim |h|^\alpha, \tag{1.143}$$

then results obtained in the large deviations of estimators allow us to hope that under some additional restrictions the distribution of random variables

$$n^{1/\alpha}(\hat\theta_n - \theta), \quad n^{1/\alpha}(\tilde t_n - \theta) \tag{1.144}$$

converges to a proper limit distribution as $n \to \infty$. Similar theorems will indeed be proved in the subsequent chapters. Here as an example we shall present some results along these lines for regular experiments.

Let experiments $\mathscr{E}_n$ generated by $X_1, X_2, \dots, X_n$ be regular. Let $I(\theta)$ denote the Fisher information matrix, set $l(x;\theta) = \ln f(x;\theta)$.

Below we shall assume that Θ is a convex bounded open subset of $\mathbb{R}^k$ and that uniformly with respect to $\theta \in K \subset \Theta$,

$$\inf_{|h|\ge\delta} \int_{\mathscr{X}} |f^{1/2}(x;\theta+h) - f^{1/2}(x;\theta)|^2 d\nu > 0. \tag{1.145}$$

Theorem 1.7.1. *Let the function $l(x;\theta)$ for all x be twice differentiable in θ. Assume, furthermore, that there exist a number $\delta, 0 < \delta \le 1$, such that for any compact set $K \subset \Theta$,*

$$\sup_{\theta\in K} \mathbb{E}_\theta \left| \frac{\partial}{\partial\theta} l(X_1;\theta) \right|^{2+\delta} < \infty \tag{1.146}$$

$$\sup_{\theta\in K} \mathbb{E}_\theta \left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} l(X_1;\theta) \right|^{1+\delta} < \infty \tag{1.147}$$

$$\mathbb{E}_\theta \left(\sup_{\theta,\theta'\in K} \left| \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(X_1;\theta) - \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(X_1;\theta') \right| |\theta-\theta'|^{-\delta} \right) < \infty. \tag{1.148}$$

Then there exists a number $\epsilon > 0$ such that with P_θ probability one uniformly in $\theta \in K$ as $n \to \infty$, we have

$$\sqrt{n} I(\theta)(\hat\theta_n - \theta) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\partial}{\partial\theta} l(X_j;\theta) + O(n^{-\epsilon}). \tag{1.149}$$

Theorem 1.7.2. *Let the conditions of Theorem 1.7.1 be satisfied. If $\tilde t_n$ is a Bayesian estimator with respect to the prior density $q(\theta) > 0$ continuous on Θ and a convex loss function $w(u), u \in \mathbf{W}_p$, then there exists $\epsilon > 0$ such that with P_θ-probability one as $n \to \infty$ we have*

$$\sqrt{n} I(\theta)(\tilde t_n - \theta) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{\partial}{\partial\theta} l(X_j;\theta) + O(n^{-\epsilon}). \tag{1.150}$$

Theorems 1.7.1 and 1.7.2 imply that in the regular case the estimators $\hat\theta_n, \tilde t_n$ are well approximated, in terms of sums of independent random variables $\frac{\partial l(X_j;\theta)}{\partial\theta}$ and, in particular, an appeal to limit theorems to sums of independent random variables allows us to obtain analogous limit theorems for the estimators $\hat\theta_n, \tilde t_n$.

For example, it follows directly from the central limit theorem that under certain conditions the differences $\sqrt{n}(\hat\theta_n - \theta), \sqrt{n}(\tilde t_n - \theta)$ are asymptotically normal with mean zero and correlation matrix $I^{-1}(\theta)$.

If θ is a one-dimensional parameter, the law of the iterated logarithm for $\hat\theta_n, \tilde t_n$ follows from the law of the iterated logarithm for sums of independent random variables. We have

$$P_\theta \left(\limsup_n (\hat\theta_n - \theta) \sqrt{\frac{nI(\theta)}{2\ln\ln n}} = -\liminf_n (\hat\theta_n - \theta) \sqrt{\frac{nI(\theta)}{2\ln\ln n}} = 1 \right) = 1 \tag{1.151}$$

$$P_\theta \left(\limsup_n (\tilde t_n - \theta) \sqrt{\frac{nI(\theta)}{2\ln\ln n}} = -\liminf_n (\tilde t_n - \theta) \sqrt{\frac{nI(\theta)}{2\ln\ln n}} = 1 \right) = 1 \tag{1.152}$$

1.8 Asymptotic Efficiency

1.8.1 Basic Definition

The term asymptotic efficient estimator was introduced by R. Fisher to designate consistent asymptotically normal estimators with asymptotically minimal variance. The motivation was that estimators of this kind should be preferable from the asymptotic point of view. The program outlined by R. Fisher consisted in showing that

1. if $\hat{\theta}_n$ is a maximum likelihood estimator then under some natural regularity conditions the different $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with parameter $(0, I^{-1}(\theta))$;

2. if T_n is an asymptotically normal sequence of estimators then

$$\liminf_{n\to\infty} \mathbb{E}_\theta \left(\sqrt{n}(T_n - \theta)\right)^2 \geq I^{-1}(\theta), \theta \in \mathbb{R}^1. \tag{1.153}$$

Should these conjectures be verified to be true, one could then indeed consider maximum likelihood estimators to be asymptotically best (in the class of asymptotically normal estimators). However, the program as stated above cannot be realized for the simple reason that estimators with minimal variance do not exist.

Indeed, let T_n be an arbitrary sequence of estimators and let the difference $\sqrt{n}(T_n - \theta)$ be asymptotically normal with parameters $(0, \sigma^2(\theta))$. If $\sigma^2(\theta_0) \neq 0$, define the estimator

$$\tilde{T}_n = \begin{cases} T_n & |T_n - \theta_0| > n^{-1/4} \\ \theta_0 & |T_n - \theta_0| \leq n^{-1/4} \end{cases} \tag{1.154}$$

It is easy to verify that the sequence $\sqrt{n}(\tilde{T}_n - \theta)$ is asymptotically normal with parameters $(0, \tilde{\sigma}^2(\theta))$, where $\tilde{\sigma}^2(\theta) = \sigma^2(\theta)$ if $\theta \neq \theta_0$ and $\tilde{\sigma}^2(\theta_0) = 0 < \sigma^2(\theta_0)$.

In particular, applying this method of improving estimators to maximum likelihood estimators one can construct (under regularity conditions) a sequence of asymptotically normal estimators $\{T_n\}$ such that

$$\lim_{n\to\infty} \mathbb{E}_\theta \left(\sqrt{n}(T_n - \theta)\right)^2 \leq I^{-1}(\theta), \tag{1.155}$$

and at certain points

$$\lim_{n\to\infty} \mathbb{E}_\theta \left(\sqrt{n}(T_n - \theta)\right)^2 < I^{-1}(\theta). \tag{1.156}$$

Estimators $\{T_n\}$ such that for $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{E}_\theta \left(\sqrt{n}(T_n - \theta)\right)^2 \leq I^{-1}(\theta), \tag{1.157}$$

and for at least one point θ the strict inequality is valid are called *superefficient with respect to a quadratic loss function*, and points θ at which the strict inequality holds are called *points of superefficiency*. Thus if $X_1, X_2, \ldots$ are iid $\mathcal{N}(\theta, 1)$ random variables, then $\bar{X}$ is the maximum likelihood estimator for θ and the estimator

$$T_n = \begin{cases} \bar{X} & |\bar{X}| > n^{-1/4} \\ \alpha\bar{X} & |\bar{X}| \leq n^{-1/4} \end{cases}, \tag{1.158}$$

$|\alpha| < 1$ is a superefficient estimator with superefficiency point $\theta = 0$. Chronologically, it is the first example of a superefficient estimator and is due to Hodges.

Various authors proposed several definitions of the notion of an asymptotically efficient estimator which retain the asymptotic efficiency of maximum likelihood estimators. We shall now present some of the definitions.

Returning to the $\mathcal{N}(\theta, 1)$ model, we note that for all θ, $\mathbb{E}_\theta \left(\sqrt{n}(\bar{X} - \theta)\right)^2 \leq 1$, while one can find a sequence of values of parameter $\theta_n \to 0$ such that

$$\liminf_{n\to\infty} \mathbb{E}_{\theta_n} \left(\sqrt{n}(T_n - \theta_n)\right)^2 > 1. \tag{1.159}$$

(for example one can choose $\theta_n = c/\sqrt{n}, c \neq 0$.) Thus in the vicinity of the point of superefficiency there are located points θ_n where the estimator T_n behaves worse than $\bar{X}$. The first definition of asymptotic efficiency is based precisely on the minimax property which $\bar{X}$ (but not T_n) possesses.

The merit of this definition is that it is in complete correspondence with the principle of comparing estimators based on their risk functions. Unlike Fisher's definition, it does not limit the class of competing estimators to asymptotically normal estimators and therefore makes sense also for nonregular statistical problems.

Definition 1.8.1. *Let $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$ be a family of statistical experiments. A family of estimators $\tilde{\theta}_\epsilon$ is called w_ϵ-asymptotically efficient in $K \subset \Theta$ (asymptotically efficient with respect to the family of loss functions w_ϵ) if for any nonempty open set $U \subset K$ the relation*

$$\liminf_{\epsilon\to 0}\left[\inf_{T_\epsilon}\sup_{u\in U}\mathbb{E}_u^{(\epsilon)}w_\epsilon(T_\epsilon-u)-\sup_{u\in U}\mathbb{E}_u^{(\epsilon)}w_\epsilon(\tilde{\theta}_\epsilon-u)\right]=0 \tag{1.160}$$

is satisfied.

As this definition plays a basic role in our further exposition we shall therefore present a few comments.

1. Should the square bracket in the left hand side of (1.160) vanish, it would mean that $\tilde{\theta}_\epsilon$ is a minimax estimator in U for the loss function w_ϵ. Therefore one can assert that an estimator is called asymptotically efficient in Θ if it is asymptotically minimax in any nonempty and (independent of ϵ) open set $U \subset \Theta$.

2. Let $w_\epsilon(x) = w(x) \in \mathbf{W}$ and $|w(x)| < c$ so that the loss function does not depend on ϵ and is bounded. Then clearly for any uniformly consistent in K estimator $\tilde{\theta}_\epsilon$,

$$\lim_{\epsilon\to 0}\sup_{u\in K}\mathbb{E}_u^{(\epsilon)}w_\epsilon(\tilde{\theta}_\epsilon-u)=0. \tag{1.161}$$

   This implies (1.160) and thus any uniformly consistent estimator will be an asymptotically efficient estimator for these w. Obviously such loss functions are not of great interest. In order to take into account more subtle differences between the estimators it is necessary that the loss function itself depend on ϵ. For example, for regular experiment generated by iid observations, it is natural to set $w_n(x) = w(\sqrt{n}x), w \in \mathbf{W}$.

3. Definition 1.8.1 can be localized in the following manner. The estimation $\tilde{\theta}_\epsilon$ is called w_ϵ-asymptotically efficient at point $\theta \in \Theta$ provided

$$\lim_{\delta\to 0}\liminf_{\epsilon\to 0}\left[\inf_{T_\epsilon}\sup_{|u-\theta|<\delta}\mathbb{E}_u^{(\epsilon)}w_\epsilon(T_\epsilon-u)-\sup_{|u-\theta|<\delta}\mathbb{E}_u^{(\epsilon)}w_\epsilon(\tilde{\theta}_\epsilon-u)\right]=0 \tag{1.162}$$

   Obviously, asymptotic efficiency in K implies asymptotic efficiency at any interior point $\theta \in K$.

In relation to Definition 1.8.1, Fisher's program is now realized: in a wide class of problems asymptotically efficient estimators exist, and under regularity conditions such estimators are maximum likelihood estimators.

Observe that the left hand side of (1.160) is bounded from above by 0 for any estimator $\tilde{\theta}_\epsilon$. Therefore in order to prove asymptotic efficiency of an estimator θ_ϵ^* it is sufficient first to verify that uniformly in $U \subset K$ the limit

$$\lim_{\epsilon\to 0}\mathbb{E}_u^{(\epsilon)}w_\epsilon(\theta_\epsilon^*-u)=L(u) \tag{1.163}$$

exist and next to prove that for any estimator T_ϵ and any nonempty open set $U \subset K$ the inequality

$$\liminf_{\epsilon\to 0}\sup_{u\in U}\mathbb{E}_u^{(\epsilon)}w_\epsilon(T_\epsilon-u)\geq\sup_{u\in U}L(u) \tag{1.164}$$

is valid.

A general method for obtaining lower bounds of the type (1.164) is given by

Theorem 1.8.1. *Let $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$, $\Theta \subset \mathbb{R}^k$ be a family of statistical experiments and let $\tilde{t}_{\epsilon,q}$ be a Bayesian estimator of parameter θ with respect to loss function w_ϵ and the prior density q. Assume that for any continuous prior density which is positive at point $u \in K$ the relation*

$$\lim_{\epsilon\to 0}\mathbb{E}_u^{(\epsilon)}w_\epsilon(\tilde{t}_{\epsilon,q}-u)=L(u) \tag{1.165}$$

is satisfied and the function $L(u)$ is continuous and bounded in K.

Then for any estimator T_ϵ and any nonempty open set $U \subset K$ the relation (1.164) is fulfilled.

Remark 1.8.1. *It is sufficient to check that (1.165) is fulfilled not for all prior density q, but only for a certain subclass Q provide that for any $U \subset K$,*

$$\sup_U L(u)=\sup_{q\in Q}\int_U L(u)q(u)du. \tag{1.166}$$

For example, one may choose $Q = \{q_\delta(u;t)\}$, where

$$q_\delta(u;t)=\begin{cases}\frac{1}{\lambda\{u:|u-t|\leq\delta\}} & |u-t|\leq\delta\\ 0 & |u-t|>\delta\end{cases} \tag{1.167}$$

Corollary 1.8.1. *If the conditions of Theorem 1.8.1 are fulfilled and for some estimator θ_ϵ^* uniformly in $u \in K$ equality (1.163) is satisfied, then the estimator θ_ϵ^* is w_ϵ asymptotically efficient in K.*

Corollary 1.8.2. *If the conditions of Theorem 1.8.1 are satisfied for $w_\epsilon(x) = (\varphi(\epsilon)x)^2$, $\Theta \subset \mathbb{R}^1$, then for a w_ϵ-asymptotically efficient estimator θ_ϵ^* in K it is necessary that for any density $q(u), u \in K$ the equality*

$$\lim_{\epsilon\to 0} \int_K \mathbb{E}_u^{(\epsilon)} \left(\varphi(\epsilon)(\theta_\epsilon^* - \tilde{t}_{\epsilon,q})\right)^2 q(u)du = 0 \tag{1.168}$$

will be fulfilled.

1.8.2 Examples

Example 1.8.1. *Consider a sequence of iid observations $X_1, X_2, \ldots, X_n$ with density $f(x;\theta)$ satisfying the conditions of Theorems 1.7.1 and 1.7.2. In view of Theorem 1.7.2 there exists the limit*

$$\lim_{n\to\infty} \mathbb{E}_u w(\sqrt{n}(\tilde{t}_{n,q_\delta} - u)) = \sqrt{\frac{\det I(u)}{(2\pi)^k}} \int_{\mathbb{R}^k} w(y) e^{-\frac{1}{2}\langle I(u)y,y\rangle} dy = L(u) \tag{1.169}$$

and hence for any sequence of estimators $\{T_n\}$ the inequality

$$\liminf_{n\to\infty} \sup_{u\in U} \mathbb{E}_u w(\sqrt{n}(T_n - u)) \geq \sup_{u\in U} L(u) \tag{1.170}$$

is valid for all $U \subset \Theta$. In view of Theorems 1.7.1 and 1.7.2 for Bayesian and maximum likelihood estimators, the equality is attained in the last inequality and these estimators become asymptotically efficient.

Example 1.8.2. *Let $X_1, X_2, \ldots, X_n$ be iid random variables with uniform distribution on $[\theta - 1/2, \theta + 1/2]$. If one chooses q_δ for the prior distribution, the posterior distribution after n observations will be uniform on the interval*

$$[\alpha_n, \beta_n] = [\theta - \delta, \theta + \delta] \cap [\max X_i - 1/2, \min X_i + 1/2]. \tag{1.171}$$

Therefore, it w is an even, increasing function on $[0,\infty)$ the Bayes estimator $t_{n,q_\delta} = \frac{1}{2}(\alpha_n + \beta_n)$. Since

$$\lim_{n\to\infty} P_u(n(1/2 + u - \max X_i) > s_1, n(\min X_i + 1/2 - u) > s_2) = e^{-s_1 - s_2}, \tag{1.172}$$

we have for any δ,

$$\lim_n \mathbb{E}_u w(n(t_{n,q_\delta} - u)) = \mathbb{E} w(\frac{\tau_1 + \tau_2}{2}), \tag{1.173}$$

where $\tau_1, -\tau_2$ are iid random Exp(1) *random variables.*

Thus in this example the classical estimator for θ, which is $\tilde{t}_n = \frac{1}{2}(\max X_i + \min X_i)$ is again asymptotically efficient. However, in this case Fisher's information is not defined and the asymptotic distribution of $n(\tilde{t}_n - \theta)$ is not normal.

1.8.3 Bahadur's Asymptotic Efficiency

Consider once more the family $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$. Let T_ϵ be an estimator for θ constructed from this experiment.

R. Bahadur suggested measuring the asymptotic efficiency of the estimators $\{T_\epsilon\}$ by the magnitude of concentration of the estimator in the interval of a fixed (independent of ϵ) length with center at θ, i.e., by the magnitude of probability

$$P_\theta^{(\epsilon)}(|T_\epsilon - \theta| < \gamma) \tag{1.174}$$

As it follows from the result of inequalities for probabilities of large deviations, under quite general assumptions the probabilities

$$P_\theta^{(\epsilon)}(|T_\epsilon - \theta| \geq \gamma) \tag{1.175}$$

for a fixed γ in the case of "nice" estimators decrease exponentially in ϵ^{-1}. If $\epsilon^{-1/\alpha}$ is the "correct" normalizing factor then it is natural to select as the measure of asymptotic efficiency the upper bound over all T_ϵ of the expressions

$$\liminf_{\gamma\to 0} \liminf_{\epsilon\to 0} \frac{\epsilon}{\gamma^\alpha} \ln P_\theta^{(\epsilon)}(|T_\epsilon - \theta| \geq \gamma). \tag{1.176}$$

The following asymptotic inequality in the iid observation model serves as the basis for calculations connected with efficient estimators in the Bahadur sense.

Theorem 1.8.2. *Let $\{T_n\}$ be a consistent estimator for θ. Then for all $\gamma' > \gamma$,*

$$\liminf_{n\to\infty} \frac{1}{n} \ln P_\theta(|T_n - \theta| > \gamma) \geq -\int_{\mathscr{X}} f(x, \theta + \gamma') \ln \frac{f(x, \theta + \gamma')}{f(x;\theta)} \nu(dx). \tag{1.177}$$

Theorem 1.8.2 actually is a particular case of Stein's lemma in hypothesis testing. It provides the bound on the probability of the first kind for testing hypothesis θ against the alternative $\theta + \gamma'$ by means of a test constructed from T_n.

Inequality (1.177) results in the following bound on asymptotic efficiency in the Bahadur sense: for any estimator T_n,

$$\liminf_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{\gamma^\alpha n} \ln P_\theta(|T_n - \theta| > \gamma) \geq \lim_{\gamma\to 0} \frac{1}{\gamma^\alpha} \int_{\mathscr{X}} f(x, \theta + \gamma) \ln \frac{f(x,\theta)}{f(x, \theta + \gamma)} \nu(dx). \tag{1.178}$$

Assume now $f(x;\theta)$ possess finite Fisher information $I(\theta)$. Then $\alpha = 2$ should be taken in (1.178); moreover under the assumption the normal distribution ought to play a prominent role since the estimators which are commonly used possess asymptotically a normal distribution.

Following Bahadur we shall call the quantity $\tau_n = \tau_n(\theta, \gamma, T_n)$ defined by the equality

$$P_\theta(|T_n - \theta| > \gamma) = \sqrt{\frac{2}{\pi}} \int_{\gamma/\tau_n}^{\infty} e^{-u^2/2} du = P(|\xi| \geq \frac{\gamma}{\tau_n}), \tag{1.179}$$

where $\mathcal{L}(\xi) = \mathcal{N}(0, 1)$, the *efficient standard deviation of the estimator* T_n. Clearly the sequence $\{T_n\}$ is consistent if and only if $\tau_n \to 0$ as $n \to \infty$ for any $\gamma > 0$ and $\theta \in Theta$. If the variable T_n is normal with mean θ and τ_n^2 is the variance of T_n.

The well-known inequality

$$\frac{z}{1 + z^2} e^{-z^2/2} < \int_z^\infty e^{-u^2/2} < \frac{1}{z} e^{-z^2/2} \tag{1.180}$$

allows us to assert that

$$\liminf_{\gamma\to 0} \liminf_{n\to\infty} \frac{1}{n\gamma^2} \ln P_\theta(|T_n - \theta| > \gamma) = -\frac{1}{2} \liminf_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\tau_n^2}. \tag{1.181}$$

Theorem 1.8.3. *Let the parameter set $\Theta = (a, b) \subset \mathbb{R}^1$ and the density $f(x;\theta)$ satisfy all the conditions of Theorem 1.7.1. Then for any estimator T_n,*

$$\liminf_{\gamma\to 0} \liminf_{n\to\infty} \frac{1}{n\gamma^2} \ln P_\theta(|T_n - \theta| > \gamma) = -\frac{1}{2} \liminf_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\tau_n^2} \geq -\frac{1}{2} I(\theta), \tag{1.182}$$

where $I(\theta)$ is the Fisher information. If, in addition, $-\infty < a < b < \infty$, the function $\frac{\partial^2}{\partial\theta^2} \ln f(x;\theta) < 0$ for all $\theta \in \Theta$ and for almost all $x \in \mathscr{X}$, and $\hat{\theta}_n$ is a maximum likelihood estimator, then

$$-\frac{1}{2} \lim_{\gamma\to 0} \lim_{n\to\infty} n^{-1} \tau_n^{-2}(\theta; \gamma, \hat{\theta}_n) = -\frac{1}{2} I(\theta). \tag{1.183}$$

Thus under specific assumptions, for example, under the conditions of Theorem 1.8.3, maximum likelihood estimators are asymptotically efficient and the phenomenon of superefficiency is not present in relation to Bahadur's definition. It is somewhat unfortunate, however, that now the class of efficient estimators contains also certain undesirable estimators such at Hodges' estimator.

Bahadur's definition does not completely eliminate the evil of superefficiency. This phenomenon many occur, for example, due to the fact that the hypotheses $\theta = t$ and $\theta = t + \gamma$ are very well distinguishable. Consider the following example.

Example 1.8.3. *Assume that $X_1, X_2, \ldots, X_n$ are independent observations on a line with a uniform distribution on the interval $|x - \theta| \leq 1/2$, where parameter $\theta \in \mathbb{R}^1$. The accepted best estimator for θ is*

$$\hat{\theta}_n = \frac{1}{2} (\max X_j + \min X_j) \tag{1.184}$$

and its efficiency is

$$\lim_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\gamma} \ln P_\theta(|\hat{\theta}_n - \theta| > \gamma) = -2. \tag{1.185}$$

In order to prove the last relation it is sufficient to verify that $\hat{\theta}_n - \theta$ possesses a density equal to $n(1 - 2|\alpha|)^{n-1}$ for $|\alpha| \leq 1/2$ and vanishes outside this interval.

Now set

$$\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if at least one observation } X_j \notin [-1/2, 1/2] \\ 0 & \text{if all } X_j \in [-1/2, 1/2] \end{cases} \tag{1.186}$$

Then

$$P_\theta(|\tilde{\theta}_n| > \gamma) = 0 \qquad \theta = 0 \tag{1.187}$$

$$P_\theta(|\tilde{\theta}_n - \theta| > \gamma) \le P_\theta(|\hat{\theta}_n - \theta| > \gamma) + (1 - |\theta|)^n \qquad |\theta| < 1, \theta \neq 0 \tag{1.188}$$

$$P_\theta(|\tilde{\theta}_n - \theta| > \gamma) = P_\theta(|\hat{\theta}_n - \theta| > \gamma) \qquad |\theta| \ge 1. \tag{1.189}$$

Consequently, for $0 < \gamma < 1/2$,

$$\ln P_\theta(|\tilde{\theta}_n - \theta| > \gamma) \le \begin{cases} -\infty & \theta = 0 \\ e^{-2n\gamma} + e^{-n\theta} & \theta \neq 0 \end{cases} \tag{1.190}$$

So that Bahadur's efficiency is

$$\lim_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\gamma} \ln P_\theta(|\tilde{\theta}_n - \theta| > \gamma) \le \begin{cases} -\infty & \theta = 0 \\ -2 & \theta \neq 0 \end{cases} \tag{1.191}$$

and the point $\theta = 0$ *is a point of superefficiency.*

Example 1.8.4. *An analogous example may be constructed also for observations with a finite Fisher information. As above, let* $X_1, X_2, \ldots, X_n$ *be independent observations on the real line with the density* $f(x - \theta)$ *where* $f(x) = 0$ *for* $|x| \ge 1$ *and* $f(x) > 0$ *for* $|x| \le 1$. *The function* $f(x)$ *is assumed to be sufficiently smooth so that*

$$I = \int_{-1}^{1} \frac{|f'(x)|^2}{f(x)} dx < \infty. \tag{1.192}$$

If moreover $(\ln f(x))'' \le 0$ *on* $[-1, 1]$ *then, using an argument similar to the one utilized in the proof of Theorem 1.8.3, one can show that for a maximum likelihood estimator* $\hat{\theta}_n$,

$$\lim_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\gamma^2} \ln P_\theta(|\hat{\theta}_n - \theta| > \gamma) = -\frac{1}{2} I. \tag{1.193}$$

Set

$$\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if at least one observation } X_j \notin [-1, 1] \\ 0 & \text{if all } X_j \in [-1, 1] \end{cases} \tag{1.194}$$

It turns out that

$$\lim_{\gamma\to 0} \lim_{n\to\infty} \frac{1}{n\gamma^2} \ln P_\theta(|\tilde{\theta}_n - \theta| > \gamma) \le \begin{cases} -\infty & \theta = 0 \\ -\frac{1}{2} I & \theta \neq 0 \end{cases} \tag{1.195}$$

and the estimator $\tilde{\theta}_n$ *is superefficient in the Bahadur sense at the point* $\theta = 0$.

1.8.4 Efficiency in C. R. Rao's Sense

Behind the definition suggested by C. R. Rao are quite different considerations: the efficiency of the estimators is determined here by some other properties rather than by their closeness to the estimated parameter. We shall limit ourselves to the case of iid observations $X_1, X_2, \ldots, X_n$ with the common density $f(x; \theta), \theta \in \mathbb{R}^1$, for which there exists finite Fisher information $I(\theta)$. C. R. Rao admits for comparison only those estimators $T_n = T_n(X_1, X_2, \ldots, X_n)$ whose distributions possess densities $\varphi_n(x^n; \theta)$ with respect to measure ν^n, where $\varphi_n(x^n; \theta)$ is an absolutely continuous function of θ. Rao's motivation is as follows: since the likelihood ratio $p_n(X^n; \theta)/p_n(X^n; \theta')$ plays a basic role in statistical problems, for a "nice" estimator T_n the ratio $\varphi_n(X^n; \theta)/\varphi_n(X^n; \theta')$ should therefore be close to the likelihood ratio $p_n(X^n; \theta)/p_n(X^n; \theta')$. This idea is formalized in the following definition: A sequence of estimators $\{T_n\}$ is asymptotically efficient if

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \left| \frac{d}{d\theta} p_n(X^n; \theta) - \frac{d}{d\theta} \varphi_n(X^n; \theta) \right| = 0 \tag{1.196}$$

in probability for all $\theta \in \Theta$.

Since it is difficult to verify (1.196) directly, C. R. Rao proposes as a basic definition the following

Definition 1.8.2. *Statistic T_n is called* asymptotically efficient in Rao's sense *if there exist two nonrandom functions $\alpha(\theta), \beta(\theta)$ such that*

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}p_n(X^n;\theta) - \alpha(\theta) - \beta(\theta)\sqrt{n}(T_n - \theta) \to 0 \tag{1.197}$$

in probability as $n \to \infty$.

It follows directly from Theorems 1.7.1 and 1.7.2 that if the conditions of these theorems are satisfied the maximum likelihood estimators as well as Bayesian estimators are asymptotically efficient in C. R. Rao's sense, provided we choose $\alpha(\theta) = 0, \beta(\theta) = I(\theta)^{-1}$.

It can be shown that under the conditions of Theorem 1.7.1 and some additional restrictions on the distribution of T_n, Fisher information contained in T_n coincides asymptotically with Fisher information contained in the whole sample:

$$\lim_{n\to\infty} \frac{1}{nI(\theta)}\mathbb{E}_\theta \left| \frac{d}{d\theta} \ln \varphi_n(X^n;\theta) \right|^2 = 1. \tag{1.198}$$

It is pleasing to realize that although there is no complete unanimity among statisticians as to the notion of efficiency, the classes of estimators which different groups of statisticians consider to be efficient become identical provided some reasonable conditions are satisfied (such as the conditions of Theorem 1.7.1).

1.9 Two Theorems on the Asymptotic Behavior of Estimators

Consider a family of statistical experiments $(\mathscr{X}^{(\epsilon)}, \mathcal{X}^{(\epsilon)}, P_\theta^{(\epsilon)}, \theta \in \Theta)$ generated by observations X^ϵ be given. Here Θ is an open subset of $\mathbb{R}^k$. Let $p_\epsilon(X^\epsilon;\theta) = \left(\frac{dP_\theta^{(\epsilon)}}{d\nu^\epsilon}\right)(X^\epsilon)$ where ν^ϵ is a measure on $\mathcal{X}^{(\epsilon)}$.

Set

$$Z_{\epsilon,\theta}(u) = Z_\epsilon(u) = \frac{p_\epsilon(X^\epsilon;\theta + \varphi(\epsilon)u)}{p_\epsilon(X^\epsilon;\theta)} \tag{1.199}$$

where $\varphi(\epsilon)$ denotes some matrix nondegenerated normalizing factor; it is also assumed that $|\varphi(\epsilon)| \to 0$ as $\epsilon \to 0$. Thus the function $Z_{\epsilon,\theta}$ is defined on the set $U_\epsilon = (\varphi(\epsilon))^{-1}(\Theta - \theta)$.

Denote by $\mathbf{C}_0(\mathbb{R}^k)$ the normalized space of functions continuous in $\mathbb{R}^k$ which vanish at infinity, with the norm $|\psi| = \sup_y |\psi(y)|$.

Below we shall denote by $\mathbf{G}$ the set of families of functions $\{g_\epsilon(y)\}$ possessing the following properties:

1. For a fixed ϵ, $g_\epsilon(y)$ is monotonically increasing to ∞ as a function of y defined on $[0, \infty)$
2. For any $N > 0$,
$$\lim_{y\to\infty, \epsilon\to 0} y^N e^{-g_\epsilon(y)} = 0. \tag{1.200}$$

For example, if $g_\epsilon(y) = y^\alpha, \alpha > 0$, then $g_\epsilon \in \mathbf{G}$.

Denote by $\hat{\theta}_\epsilon$ the maximum likelihood estimator for θ.

Theorem 1.9.1. *Let the parameter set Θ be an open subset of $\mathbb{R}^k$, functions $Z_{\epsilon,\theta}(u)$ be continuous with probability one possessing the following properties:*

1. *For any compact $K \subset \Theta$, there correspond numbers $a(K) = a$ and $B(K) = B$ and functions $g_\epsilon^K(y) = g_\epsilon(y), \{g_\epsilon\} \in \mathbf{G}$ such that*
   (a) *there exist numbers $\alpha > k, m \geq \alpha$ such that for $\theta \in K$,*
   $$\sup_{|u_1|\leq R, |u_2|\leq R, u_1, u_2 \in U_\epsilon} |u_2 - u_1|^{-\alpha}\mathbb{E}_\theta^{(\epsilon)}|Z_{\epsilon,\theta}^{1/m}(u_2) - Z_{\epsilon,\theta}^{1/m}(u_1)|^m \leq B(1 + R^a) \tag{1.201}$$
   (b) *For all $u \in U_\epsilon, \theta \in K$,*
   $$\mathbb{E}_\theta^{(\epsilon)} Z_{\epsilon,\theta}^{1/2}(u) \leq e^{-g_\epsilon(|u|)}. \tag{1.202}$$
2. *Uniformly in $\theta \in K$ for $\epsilon \to 0$ the marginal (finite dimensional) distributions of the random functions $Z_{\epsilon,\theta}(u)$ converge to marginal distributions of random functions $Z_\theta(u) = Z(u)$ where $Z \in \mathbf{C}_0(\mathbb{R}^k)$.*
3. *The limit functions $Z_\theta(u)$ with probability one attain the maximum at the unique point $\hat{u}(\theta) = u$.*

Then uniformly in $\theta \in K$ the distribution of the random variables $\varphi^{-1}(\epsilon)(\hat{\theta}_\epsilon - \theta)$ converge to the distribution of $\hat{u}(\theta)$ and for any continuous loss function $w \in \mathbf{W}_p$ we have uniformly in $\theta \in K$,

$$\lim_{\epsilon \to 0} \mathbb{E}_\theta^{(\epsilon)} w(\varphi^{-1}(\epsilon)(\hat{\theta}_\epsilon - \theta)) = \mathbb{E} w(\hat{u}(\theta)). \tag{1.203}$$

A similar theorem can be proved for Bayesian estimators as well. Consider the family $\{\tilde{t}_\epsilon\}$ of Bayesian estimators with respect to the loss function $W_\epsilon(u, v) = l(\varphi^{-1}(\epsilon)|u - v|)$ and prior density q. We shall assume that $l \in \mathbf{W}_p$ and that, moreover, for all H sufficiently large and γ sufficiently small

$$\inf_{|u| > H} l(u) - \sup_{|u| \le H^\gamma} l(u) \ge 0. \tag{1.204}$$

Denote by Q the set of continuous positive functions $q : \mathbb{R}^k \to \mathbb{R}^1$ possessing a polynomial majorant.

Theorem 1.9.2. *Let $\{\tilde{t}_\epsilon\}$ be a family of Bayesian estimators with respect to the loss function $l(\varphi^{-1}(\epsilon)|u - v|)$, where $l \in W_p$ and satisfy (1.204), and prior density $q \in Q$. Assume that the normalized likelihood ratio $Z_{\epsilon,\theta}(u)$ possesses the following properties:*

1. *For any compact $K \subset \Theta$ there correspond numbers $a(K) = a, B(K) = B$ and functions $g_\epsilon^K(y) = g_\epsilon(y) \in \mathbf{G}$ such that*

   (a) For some $\alpha > 0$ and all $\theta \in K$

   $$\sup_{|u_1| \le R, |u_2| \le R} |u_2 - u_1|^{-\alpha} \mathbb{E}_\theta^{(\epsilon)} |Z_{\epsilon,\theta}^{1/2}(u_2) - Z_{\epsilon,\theta}^{1/2}(u_1)|^2 \le B(1 + R^a) \tag{1.205}$$

   (b) For all $u \in U_\epsilon, \theta \in K$,

   $$\mathbb{E}_\theta^{(\epsilon)} Z_{\epsilon,\theta}^{1/2}(u) \le e^{-g_\epsilon(|u|)}. \tag{1.206}$$

2. *The marginal distributions of the random functions $Z_{\epsilon,\theta}(u)$ converge uniformly in $\theta \in K$ as $\epsilon \to 0$ to the marginal distributions of random functions $Z_\theta(u) = Z(u)$.*

3. *The random function*

   $$\psi(s) = \int_{\mathbb{R}^k} l(s - u) \frac{Z(u)}{\int_{\mathbb{R}^k} Z(v) dv} du \tag{1.207}$$

   attains its absolute minimum value at the unique point $\tau(\theta) = \tau$.

Then the distribution of the random variables $\varphi^{-1}(\tilde{t}_\epsilon - \theta)$ converges uniformly in $\theta \in K$ to the distribution of τ and for any continuous loss function $w \in \mathbf{W}_p$ we have uniformly in $\theta \in K$,

$$\lim_{\epsilon \to 0} \mathbb{E}_\theta^{(\epsilon)} w(\varphi^{-1}(\epsilon)(\tilde{t}_\epsilon - \theta)) = \mathbb{E} w(\tau(\theta)). \tag{1.208}$$

Remark 1.9.1. *Note that condition 1.(a) is substantially weaker than the analogous condition of Theorem 1.9.1. Condition (3) is automatically satisfied if l is a convex function with a unique minimum, for example, if $l(u) = |u|^p, p \ge 1$.*

1.9.1 Examples

Example 1.9.1. *Consider iid observations X_j with density $f(x; \theta), \theta \in \Theta \subset \mathbb{R}^k$ with respect to some measure ν. Assume the requirements in approximating likelihood ratio by sum of iid random variables are satisfied. In particular, Fisher information $I(\theta)$ exists.*

As we already know, if one chooses $\varphi(n) = 1/\sqrt{n}$ and sets

$$Z_n(u) = \prod_{j=1}^n \frac{f(X_j; \theta + u/\sqrt{n})}{f(X_j; \theta)}, \tag{1.209}$$

then the conditions of Theorem 1.9.2 will be fulfilled. In view of our assumptions,

$$\ln Z_n(u) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \langle \frac{\partial}{\partial \theta} \ln f(X_j; \theta), u \rangle + \frac{1}{2n} \sum_{j=1}^n \langle \frac{\partial^2}{\partial \theta^2} \ln f(X_j; \theta) u, u \rangle + R_n, \quad R_n \to 0. \tag{1.210}$$

In the same manner as in approximating the likelihood ratio by iid random variables, we can prove that

$$\frac{1}{2n}\sum_{j=1}^{n}\langle\frac{\partial^2}{\partial\theta^2}\ln f(X_j;\theta)u,u\rangle\to-\frac{1}{2}\langle I(\theta)u,u\rangle \tag{1.211}$$

with probability one as $n\to\infty$. Therefore the marginal distributions of $Z_n(u)$ converge to the marginal distribution of the random function

$$Z(u)=e^{\langle\xi,u\rangle-\frac{1}{2}\langle I(\theta)u,u\rangle}, \tag{1.212}$$

where ξ is a normal random vector in $\mathbb{R}^k$ with mean zero and correlation matrix $I^{-1}(\theta)$.

Example 1.9.2. *Let $X^{(n)}=(X_1,X_2,\dots,X_n)$ be a sample from a uniform distribution on the interval $[\theta-1/2,\theta+1/2],\theta\in\mathbb{R}^1$. The parameter to be estimated is θ. If we choose $\varphi(n)=n^{-1}$ then the function*

$$Z_n(u)=\prod_{j=1}^{n}\frac{f(X_j-\theta-u/n)}{f(X_j-\theta)} \tag{1.213}$$

satisfy the conditions of Theorem 1.9.2. Here $f(x)$ is the density of a uniform distribution on $[-1/2,1/2]$. The behavior of the marginal distribution of Z_n does not depend on θ. Setting $\theta=0$, we obtain

$$Z_n(u)=\begin{cases}1 & n(\max X_j-1/2)<u<n(\min X_j+1/2)\\ 0 & u\notin[n(\max X_j-1/2),n(\min X_j+1/2)]\end{cases} \tag{1.214}$$

We have shown that the random variables $-n(\max X_j-1/2),n(\min X_j+1/2)$ are asymptotically independent and possess in the limit the exponential distribution with parameter one. Thus condition 2 of Theorem 1.9.2 is also fulfilled and the limiting process

$$Z(u)=\begin{cases}1 & -\tau^-<u<\tau^+\\ 0 & u\notin[-\tau^-,\tau^+]\end{cases} \tag{1.215}$$

here τ^-,τ^+ are iid Exp(1) *random variables. The conditions of Theorem 1.9.1 are obviously not satisfied in this example since both $Z_n(u),Z(u)$ are discontinuous.*

Chapter 2

Local Asymptotic Normality of Families of Distributions

In a number of interesting papers of Hajek, Le Cam and other authors, it was proved that many important properties of statistical estimators follow from the asymptotic normality of the logarithm of the likelihood ratio for neighborhood hypotheses regardless of the relation between the observations which produced the given likelihood function. This chapter is devoted to an investigation of the conditions under which this property is valid for various models and to corollaries of this property.

2.1 Independent Identically Distributed Observations

In this section we shall establish an important property of a family of regular statistical experiments generated by a sequence of iid observations. Let $\mathscr{E}_i = \{\mathscr{X}_i, \mathcal{X}_i, P_\theta, \theta \in \Theta\}$ be an regular experiment corresponding to the i-th observation and let X_i be the i-th observation. The set Θ, as always, will be considered an open subset of $\mathbb{R}^k$.

Let $\mathscr{E}^{(n)} = \mathscr{E}_1 \times \mathscr{E}_2 \times \ldots \times \mathscr{E}_n$ and let

$$\tilde{Z}_{n,\theta}(u) = \prod_{j=1}^{n} \frac{f(X_j; \theta + un^{-1/2})}{f(X_j; \theta)} \tag{2.1}$$

be the normalized likelihood ratio.

The following important theorem is due to Le Cam.

Theorem 2.1.1. *If $\mathscr{E}_i$ are regular experiments in Θ and $\det I(\theta) \neq 0$ for $\theta \in \Theta$, then the normalized likelihood ratio $\tilde{Z}_{n,\theta}(u)$ can be written as*

$$\tilde{Z}_{n,\theta}(u) = e^{\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\langle \frac{\partial \ln f(X_j;\theta)}{\partial \theta}, u \right\rangle - \frac{1}{2}\langle I(\theta)u, u\rangle + \tilde{\psi}_n(u,\theta)}. \tag{2.2}$$

Moreover,

$$P_\theta^n(|\tilde{\psi}_n(u,\theta)| > \epsilon) \to 0 \tag{2.3}$$

as $n \to \infty$ for every $u \in \mathbb{R}^1, \epsilon > 0$ and

$$\mathcal{L}\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{\partial \ln f(X_j;\theta)}{\partial \theta} \middle| P_\theta^n \right) \to \mathcal{N}(0, I(\theta)). \tag{2.4}$$

Remark 2.1.1. *Using the substitution $u = I(\theta)^{-1/2}v$ the assertion of Theorem 2.1.1 can be restated as follows: if the conditions of the theorem are satisfied then*

$$Z_{n,\theta}(v) = \frac{dP^n_{\theta+(nI(\theta))^{-1/2}v}}{dP^n_\theta}(X^n) = e^{\langle v, \Delta_{n,\theta}\rangle - 1/2|v|^2 + \psi_n(v,\theta)}, \tag{2.5}$$

where $P_\theta(|\psi(v,\theta)| > \epsilon) \to 0$ as $n \to \infty$ and

$$\mathcal{L}(\Delta_{n,\theta} | P_\theta^n) \to \mathcal{N}(0, J). \tag{2.6}$$

The requirement of regularity of an experiment was not fully utilized in the proof of Theorem 2.1.1. Actually one can state a condition which is sufficient for the validity of Theorem 2.1.1 without introducing any auxiliary measure ν, but rather by introducing the requirement directly on the family P_θ in the neighborhood of a given point $\theta = t$. We shall now formulate this condition.

Condition 2.1.1 (Condition $\mathscr{A}_t$)**.** *Let $P_{\theta,c}(\cdot)$ and $P_{\theta,s}(\cdot)$ represent the absolutely continuous and the singular components, respectively, of measure $P_\theta(\cdot)$ with respect to $P_t(\cdot)$ (t fixed). Define*

$$\zeta(u) = \left[\frac{dP_{t+u,c}}{dP_t}(X_1)\right]^{1/2} - 1 \tag{2.7}$$

and assume the following conditions are satisfied:

1. *The process $\zeta(u)$ is differentiable in $L_2(P_t)$ for $u = 0$, i.e. there exists a random vector φ such that as $u \to 0$,*

$$\mathbb{E}_t \left(\zeta(u) - \langle \varphi, u\rangle\right)^2 = o(|u|^2). \tag{2.8}$$

2. *As $u \to 0$,*

$$\int dP_{t+u,s} = o(|u|^2). \tag{2.9}$$

It is easy to verify that the proof of Theorem 2.1.1 remains unchanged if condition 2.1.1 is satisfied. In this connection, we should set

$$I(t) = 4\mathbb{E}_t \varphi\varphi^T. \tag{2.10}$$

We also have the following uniform version of Theorem 2.1.1.

Theorem 2.1.2. *If $\mathscr{E}_i$ are regular experiments in Θ and $\det I(\theta) \neq 0, \forall\theta \in \Theta$ then for any compact set $K \subset \Theta$, any sequence $\theta_n \subset K$ and any $u \in \mathbb{R}^k$ the following representation is valid as $n \to \infty$:*

$$Z_{n,\theta_n}(u) = e^{\frac{1}{\sqrt{n}}\left\langle \sum_{j=1}^n \frac{\partial \ln f(X_j,\theta_n)}{\partial\theta}, u\right\rangle - \frac{1}{2}\langle I(\theta_n)u,u\rangle + \psi_n(u,\theta_n)}, \tag{2.11}$$

moreover, for any $u \in \mathbb{R}^k, \epsilon > 0$,

$$P^n_{\theta_n}(|\psi_n(u,\theta_n)| > \epsilon) \to 0, \quad n \to \infty, \tag{2.12}$$

$$\mathcal{L}\left(n^{-1/2}I^{-1/2}(\theta_n)\sum_{j=1}^n \frac{\partial \ln f(X_j,\theta_n)}{\partial\theta}\middle| P^n_{\theta_n}\right) \to \mathcal{N}(0,J), \quad n \to \infty. \tag{2.13}$$

2.2 Local Asymptotic Normality (LAN)

It is important to note that the property of the likelihood ratio proved in Theorem 2.1.1 is valid in a substantially large class of cases than the case of independent observations.

In this connection it is desirable to state the following general definition. Let $\mathscr{E}_\epsilon = (\mathscr{X}^{(\epsilon)}, \mathscr{X}^{(\epsilon)}, P^{(\epsilon)}_\theta, \theta \in \Theta), \Theta \subset \mathbb{R}^k$ be a family of statistical experiments and X^ϵ be the corresponding observation. As usual we shall refer to $\frac{dP^{(\epsilon)}_{\theta_2}}{dP^{(\epsilon)}_{\theta_1}}(X^\epsilon)$ the derivative of the absolutely continuous component of the measure $P^{(\epsilon)}_{\theta_2}$ with respect to measure $P^{(\epsilon)}_{\theta_1}$ at the observation X^ϵ as the likelihood ratio.

Definition 2.2.1 (Local Asymptotic Normality (LAN))**.** *A family $P^{(\epsilon)}_\theta$ is called locally asymptotically normal (LAN) at point $t \in \Theta$ as $\epsilon \to 0$ if for some nondegenerate $k \times k$-matrix $\varphi(\epsilon) = \varphi(\epsilon, t)$ and any $u \in \mathbb{R}^k$, the representation*

$$Z_{\epsilon,t}(u) = \frac{dP^{(\epsilon}_{t+\varphi(\epsilon)u}}{dP^{(\epsilon)}_t}(X^\epsilon) = e^{u^T\Delta_{\epsilon,t} - \frac{1}{2}|u|^2 + \psi_\epsilon(u,t)} \tag{2.14}$$

is valid, where

$$\mathcal{L}(\Delta_{\epsilon,t}|P^{(\epsilon)}_t) \to \mathcal{N}(0,J), \tag{2.15}$$

as $\epsilon \to 0$. Here J is the identity $k \times k$ matrix and moreover for any $u \in \mathbb{R}^k$ we have

$$\psi_\epsilon(u,t) \to 0 \tag{2.16}$$

in $P^{(\epsilon)}_t$-probability as $\epsilon \to 0$.

First, it follows from Theorem 2.1.1 that in the case of iid regular experiments with a nondegenerate matrix $I(t)$, the LAN condition is fulfilled at point $\theta = t$, if we set

$$\epsilon = 1/n, \varphi(\epsilon, t) = (nI(t))^{-1/2}, \Delta_{\epsilon,t} = (nI(t))^{-1/2} \sum_{i=1}^{n} \frac{\partial \ln f(X_j, t)}{\partial t}. \tag{2.17}$$

We now state here a simple theorem which would allow us to check the LAN condition for a one-dimensional parameter set.

Theorem 2.2.1 (Hajek). *Let $\Theta \subset \mathbb{R}^1$ and the density $f(x;\theta)$ of experiments $\mathscr{E}_i$ satisfy the following conditions:*

1. *The function $f(x;\theta)$ is absolutely continuous in θ in some neighborhood of the point $\theta = t$ for all $x \in \mathscr{X}$*
2. *The derivative $\frac{\partial f(x;\theta)}{\partial \theta}$ exists for every θ belonging to this neighborhood for ν-almost all $x \in \mathscr{X}$*
3. *The function $I(\theta)$ is continuous and positive for $\theta = t$.*

Then the family P_θ^n generated by independent experiments $\mathscr{E}_1, \ldots, \mathscr{E}_n$ with density f satisfy the LAN condition for $\theta = t$ and moreover $\epsilon, \varphi, \Delta$ are given by

$$\epsilon = 1/n, \varphi(\epsilon, t) = (nI(t))^{-1/2}, \Delta_{\epsilon,t} = (nI(t))^{-1/2} \sum_{i=1}^{n} \frac{\partial \ln f(X_j, t)}{\partial t}. \tag{2.18}$$

Later a uniform version of the LAN condition will be needed. We now present the corresponding definition.

Definition 2.2.2. *A family $P_\theta^{(\epsilon)}, \theta \in \Theta \subset \mathbb{R}^k$ is called* uniformly asymptotically normal *in $\Theta_1 \subset \Theta$ if for some nondegenerate matrix $\varphi(\epsilon, t)$ and arbitrary sequences $t_n \subset \Theta_1, \epsilon_n \to 0, u_n \to u \in \mathbb{R}^k$ such that $t_n + \varphi(\epsilon_n, t_n) \in \Theta_1$, the representation*

$$Z_{\epsilon_n,t_n}(u) = \frac{dP^{(\epsilon_n)}_{t_n+\varphi(\epsilon_n,t_n)u_n}}{dP^{(\epsilon_n)}_{t_n}}(X^{\epsilon_n}) = e^{\langle \Delta_{\epsilon_n,t_n}, u\rangle - \frac{1}{2}|u|^2 + \psi_{\epsilon_n}(u_n,t_n)} \tag{2.19}$$

is valid; here

$$\mathcal{L}(\Delta_{\epsilon_n,t_n} | P^{(\epsilon_n)}_{t_n}) \to \mathcal{N}(0, J), \quad \epsilon_n \to 0, \tag{2.20}$$

and the sequence $\psi_n(\epsilon_n, t_n)$ converges to zero in $P^{(\epsilon_n)}_{t_n}$-probability.

Theorem 2.1.2 implies that for iid regular experiments witha matrix $I(\theta)$ nondegenerate in Θ the corresponding family of distributions is uniformly asymptotically normal in any compact set $K \subset \Theta$; moreover $\varphi(\epsilon, t)$ and $\Delta_{\epsilon,t}$ may be computed using formula (2.18).

2.3 Independent Nonhomogeneous Observations

Let $X_1, X_2, \ldots, X_n$ be independent observations but we shall assume the densities f_j with respect to the measure ν_j of observations X_j depend on j.

We shall assume that every experiment $\mathscr{E}_i = (\mathscr{X}_i, \mathcal{X}_i, P_{\theta,i}, \theta \in \Theta)$ is regular, and we shall study the problem of what additional conditions should be imposed on the family of experiments so that the measure P_θ^n corresponding to the experiment $\mathscr{E}^{(n)} = \mathscr{E}_1 \times \mathscr{E}_2 \times \ldots \times \mathscr{E}_n$ will satisfy the LAN condition.

Clearly the likelihood ratio in this case becomes

$$\frac{dP^n_{\theta_2}}{dP^n_{\theta_1}} = \prod_{j=1}^{n} \frac{f_j(X_j, \theta_2)}{f_j(X_j, \theta_1)}. \tag{2.21}$$

Assume for the time being that $\Theta \subset \mathbb{R}^1$ and denote by $I_j(\theta)$ the Fisher information amount of the experiment $\mathscr{E}_j$,

$$\Psi^2(n, \theta) = \sum_{j=1}^{n} I_j(\theta). \tag{2.22}$$

Theorem 2.3.1. *Let $\mathscr{E}_i$ be regular experiments, $\Theta \subset \mathbb{R}^1$, $\Psi^2(n,t) > 0$, and for any $k > 0$,*

$$\lim_{n\to\infty} \sup_{|u|<k} \frac{1}{\Psi^2(n,t)} \sum_{j=1}^{n} \int \left(\frac{\partial}{\partial\theta} \sqrt{f_j(x, t + \frac{u}{\Psi(u,t)})} - \frac{\partial}{\partial t} \sqrt{f_j(x,t)} \right)^2 v_j(dx_j) = 0, \tag{2.23}$$

and moreover, let Lindeberg's condition be satisfied: for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\Psi^2(n,t)} \sum_{j=1}^{n} \mathbb{E}_t \left\{ \left| \frac{f_j'(X_j,t)}{f_j(X_j,t)} \right|^2 I\left(\left| \frac{f_j'}{f_j} \right| > \epsilon \Psi(n,t) \right) \right\} = 0. \tag{2.24}$$

Then the family of measures

$$P_\theta^n(A) = \int \ldots_A \int \prod_{j=1}^{n} f_j(x_j,\theta) \nu_j(dx_j) \tag{2.25}$$

satisfies at point $\theta = t$ the LAN condition with

$$\epsilon = 1/n, \varphi(n) = \varphi(n,t) = \Psi^{-1}(n,t), \Delta_{n,t} = \varphi(n,t) \sum_{j=1}^{n} \frac{f_j'(X_j,t)}{f_j(X_j,t)}. \tag{2.26}$$

Analyzing the proof of Theorem 2.3.1 one can easily verify that it remains valid also for the "sequence of series" of independent observations, i.e. triangular array models.

Indeed, Lindeberg's theorem as well as the theorem concerning the relative stability of some random variables remain valid when applied to the sequence of the series. It is even more valid as far as the remaining purely analytical calculations are concerned. Thus the following theorem holds.

Theorem 2.3.2. *Let $\mathscr{E}_{jn}, j = 1,2,\ldots,n; n = 1,2,\ldots$ be a sequence of a series of independent regular experiments where f_{jn} is the density of experiment $\mathscr{E}_{jn}$ with respect to a σ-finite measure ν_{jn} and $I_{jn}(\theta)$ is the corresponding information amount. Let*

$$\Psi^2(n,t) = \sum_{j=1}^{n} I_{jn}(t) > 0. \tag{2.27}$$

If, moreover, conditions in Theorem 2.3.1 are satisfied with f_j replaced by f_{jn}, then the family

$$P_\theta^n(A) = \int \ldots_A \int \prod_{j=1}^{n} f_{jn}(x_j,\theta) \nu_{jn}(dx_j) \tag{2.28}$$

satisfies at point $\theta = t$ the LAN condition with $\varphi(n) = \Psi^{-1}(n,t)$,

$$\Delta_{n,t} = \varphi(n) \sum_{j=1}^{n} \frac{f_{jn}'(X_{jn},t)}{f_{jn}(X_{jn},t)}. \tag{2.29}$$

Conditions 2.23 and (2.24) seem to be quite complicated. We shall now present some weaker but more easily verifiable conditions. Condition 2.24 is the Lindeberg condition and it is well known that it follows from Lyapunov's condition: for some $\delta > 0$,

$$\frac{1}{[\Psi(n,t)]^{2+\delta}} \sum_{j=1}^{n} \mathbb{E}_t \left| \frac{f_j'}{f_j}(X_j,t) \right|^{2+\delta} \to 0, \quad n \to \infty. \tag{2.30}$$

If the functions $(f_j^{1/2}(x,\theta))'$ are absolutely continuous in θ for almost all x, it is then easy to devise a simpler sufficient condition for the validity of condition (2.23). Indeed, by the Cauchy–Schwarz inequality,

$$\int \left(\frac{\partial}{\partial\theta} \sqrt{f_j(x,\theta)} - \frac{\partial}{\partial t} \sqrt{f_j(x,t)} \right)^2 \nu_j(dx) = \int \left(\int_t^\theta \frac{\partial^2}{\partial v^2} \sqrt{f_j(x,v)} dv \right)^2 \nu_j(dx) \tag{2.31}$$

$$\leq (\theta - t) \int_t^\theta dv \int \left| \frac{\partial^2}{\partial v^2} \sqrt{f_j(x,v)} \right|^2 \nu_j(dx). \tag{2.32}$$

Consequently, condition (2.23) follows from the condition

$$\lim_{n\to\infty} \frac{1}{\Psi^4(n)} \sup_{|\theta - t| < |u|/\Psi(n,t)} \sum_{j=1}^{n} \int \left| \frac{\partial^2}{\partial\theta^2} \sqrt{f_j(x,\theta)} \right|^2 \nu_j(dx) = 0. \tag{2.33}$$

2.4 Multidimensional Parameter Set

Let us assume once again that the statistical experiment $\mathscr{E}^{(n)}$ is generated by a sequence of independent regular experiments $\mathscr{E}_j$ with density $f_j(x,\theta)$, but $\Theta \subset \mathbb{R}^k$. As above, let $I_j(\theta)$ be the Fisher information matrix of the j-th experiment and assume that the matrix

$$\Psi^2(n,t) = \sum_{j=1}^{n} I_j(t) \tag{2.34}$$

is positive definite. Hence there exists a positive definite matrix

$$\Psi^{-1}(n,t) = \left(\sum_{j=1}^{n} I_j(t) \right)^{-1/2}. \tag{2.35}$$

Theorem 2.4.1. *Let $\Theta \subset \mathbb{R}^k$, the matrix $\Psi^2(n,\theta)$ is positive definite and the following conditions are satisfied:*

1. *for any $k > 0$,*

$$\lim_{n\to\infty} \sup_{|u|<k} \sum_{j=1}^{n} \int \left\langle \frac{f_j^{1/2}(x, t+\Psi^{-1}(n,t)u)}{\partial t} - \frac{\partial f_j^{1/2}(x,t)}{\partial t}, \Psi^{-1}(n,t)u \right\rangle^2 \nu_j(dx) = 0 \tag{2.36}$$

2. *Lindeberg's condition: for any $\epsilon > 0, u \in \mathbb{R}^k$,*

$$\lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}_t \left\{ \left\langle \Psi^{-1}(n,t)u, \frac{\partial \ln f_j(X_j,t)}{\partial t} \right\rangle^2 I\left(\left| \left\langle \Psi^{-1}(n,t)u, \frac{\partial \ln f_j(X_j,t)}{\partial t} \right\rangle \right| > \epsilon \right) \right\} = 0. \tag{2.37}$$

Then the family of measures

$$P_\theta^n(A) = \int \ldots_A \int \prod_{j=1}^{n} f_j(x_j,\theta)\nu_j(dx_j) \tag{2.38}$$

satisfies the LAN condition at $\theta = t$ with

$$\varphi(n) = \left(\sum_{j=1}^{n} I_j(t) \right)^{-1/2} = \Psi^{-1}(n,t), \Delta_{n,t} = \varphi(n) \sum_{j=1}^{n} \frac{f_j'(X_j,t)}{f_j(X_j,t)}. \tag{2.39}$$

Evidently Theorem 2.4.1, subject to corresponding modifications, is valid also for a sequence of series.

Somewhat strengthening the conditions of Theorem 2.4.1 one can assure the uniform asymptotic normality of the corresponding family of distributions.

Theorem 2.4.2. *Let $\mathscr{E}^{(n)} = \mathscr{E}_1 \times \ldots \times \mathscr{E}_n$, where $\mathscr{E}_j$ is a regular experiment with density $f_j(x,\theta), x \in \mathscr{X}_j, \theta \in \Theta \subset \mathbb{R}^k$. Assume that the matrix*

$$\Psi^2(n,\theta) = \sum_{j=1}^{n} I_j(\theta) \tag{2.40}$$

is positive definite for some n uniformly in $\Theta_1 \subset \Theta$ and, moreover, let the following conditions be satisfied:

1. *Random vectors $\eta_j(\theta) = \frac{\partial \ln f_j(X_j,\theta)}{\partial\theta}$ satisfies Lindeberg's condition uniformly in $\theta \in \Theta_1$: for any $\epsilon > 0, u \in \mathbb{R}^k$,*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_1} \sum_{j=1}^{n} \mathbb{E}_\theta^n \left\{ \langle \eta_j(\theta), \Psi^{-1}(n,\theta)u\rangle^2 I(|\langle \eta_j(\theta), \Psi^{-1}(n,\theta)u\rangle| > \epsilon) \right\} = 0. \tag{2.41}$$

2. *For any $k > 0$,*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_1} \sup_{|u|<k} \sum_{j=1}^{n} \int \left[\left\langle \frac{\partial f_j^{1/2}(x, \theta+\Psi^{-1}(n,\theta)u)}{\partial\theta} - \frac{\partial f_j^{1/2}(x,\theta)}{\partial\theta}, \Psi^{-1}(n,\theta)u \right\rangle \right]^2 \nu_j(dx) = 0. \tag{2.42}$$

Then the condition of uniform asymptotic normality is fulfilled for the family P_θ^n in the domain Θ_1.

2.5 Characterizations of Limiting Distributions of Estimators

2.5.1 Estimators of an Unknown Parameter when the LAN Condition is Fulfilled

We shall now begin the study of properties of estimators of an unknown parameter in the case when the LAN condition is fulfilled. Here we shall not confine ourselves to nonrandomized estimators but shall assume, given the value of observation X^ϵ that a statistician can randomly choose an estimator $\tilde{\theta}_\epsilon$ of the parameter θ in accordance with the conditional distribution $\tilde{P}_\epsilon(\tilde{\theta}_\epsilon \in A|X^\epsilon)$ which does not depend on θ. A measure generated by this mode of estimation which corresponds to the value $\theta = t$ will be denoted by $\tilde{P}_t^{(\epsilon)}$ and the expectation with respect to this measure will be denoted by $\tilde{\mathbb{E}}_t^{(\epsilon)}$. If the LAN condition is satisfied, then we could show

$$\sup_{|\xi|<1} \left| \int \xi d\tilde{P}^{(\epsilon)}_{t+\varphi(\epsilon)u} - \int \xi e^{\langle u, \hat{\Delta}_{\epsilon,t}\rangle - \frac{1}{2}|u|^2} d\tilde{P}_t^{(\epsilon)} \right| \to 0. \tag{2.43}$$

We shall try to describe the class of possible limiting distributions of appropriately normalized estimators under the LAN conditions. It was shown in Chapter 1 that under some restrictions the maximum likelihood estimator $\hat{\theta}_n$ in the case of iid observations possesses the following limiting distribution as $n \to \infty$:

$$\mathcal{L}(I(t)^{1/2} n^{1/2}(\hat{\theta}_n - t)|P_t^n) \to \mathcal{N}(0,1). \tag{2.44}$$

Obviously this distribution is not the only one possible. As the examples of superefficient estimators show there exist asymptotically normal estimators which at a given point possess an arbitrarily small variance of the limiting distribution while at the remaining points the variance is one. Below in Theorem 2.8.2 it will be shown that these estimators ought not to be considered since they are "bad" in a certain sense in the neighborhood of a radius of order $n^{-1/2}$ of superefficiency points. Since the exact value of the parameter is unknown to the statistician, it is therefore natural to restrict the study of limiting distributions to the estimators such that a small variation in the "true" value of the parameter yields a small variation in the limiting distribution of the estimator. Such regular estimators are discussed in this section.

However, the limiting distribution is not necessarily normal even in the class of regular estimators. Let, for example, a sample $X_1, X_2, \ldots, X_n$ represent iid observations of $\mathcal{N}(\theta, 1)$. Consider the randomized estimator

$$\theta_n = \bar{X} + n^{-1/2}\eta, \tag{2.45}$$

where η is a random variable independent of $X_1, \ldots, X_n$ with the distribution $G(x)$. It is clear that in this case

$$\mathcal{L}(\sqrt{n}(\theta_n - t)|P_t^n) \to \mathcal{N}(0,1) * G, \tag{2.46}$$

where $*$ denotes convolution. In a very interesting paper by Hajek, it was established that there are no other limiting distributions besides $\mathcal{N}(0,1) * G$ for the regular estimators provided the LAN conditions are satisfied.

2.5.2 Regular Parameter Estimators

Definition 2.5.1 (Regular Estimators)**.** *Let a family $P_\theta^{(\epsilon)}$ satisfy the LAN condition with the normalizing matrix $\varphi(\epsilon)$ at point $\theta = t$. An estimator θ_ϵ (possibly a randomized one) of parameter θ is called* regular *at the point $\theta = t$ if for some proper distribution function $F(x)$ the weak convergence*

$$\mathcal{L}(\varphi^{-1}(\epsilon)(\theta_\epsilon - (t + \varphi(\epsilon)u))|P^{(\epsilon)}_{t+\varphi(\epsilon)u}) \to F \tag{2.47}$$

as $\epsilon \to 0$ for any $u \in \mathbb{R}^k$ is valid; this convergence is uniform in $|u| < c$ for any $c > 0$.

We shall briefly discuss this definition. For $u = 0$ it implies that the random variable $\varphi^{-1}(\epsilon)(\theta_\epsilon - t)$ possesses the proper limiting distribution $F(x)$ as $\epsilon \to 0$, provided the true value of the parameter is t. It is quite natural to require that this convergence be uniform in t. The condition in Definition 2.5.1 represents a weakened version of this requirement since $|\varphi(\epsilon)u| \to 0$ as $\epsilon \to 0$ for any $u \in \mathbb{R}^k$. In particular, it is satisfied at each point $t \in \Delta$ if the relation

$$\mathcal{L}(\varphi^{-1}(\epsilon)(\theta_\epsilon - t)|\tilde{P}_t^{(\epsilon)}) \to F(t, \cdot), \tag{2.48}$$

is valid for some normalizing matrix $\varphi^{-1}(\epsilon)$ and some function $F(t,\cdot)$ continuous in t, uniformly in $t \in \Delta$ as $\epsilon \to 0$.

The question arises: why should only limiting distributions with normalizing matrix $\varphi^{-1}(\epsilon)$ be considered in this definition? Are there other estimators which possess the proper limiting distributions with a "better" normalizing matrix? To formulate a precise result it is necessary to examine the meaning of a normalizing matrix $\Psi(\epsilon)$ which is not better than matrix $\varphi^{-1}(\epsilon)$.

For a one-dimensional parameter set this question does not involve any difficulties: clearly a normalizing factor $\Psi(\epsilon)$ is not better than $\varphi^{-1}(\epsilon)$ if the product $\Psi(\epsilon)\varphi(\epsilon)$ is bounded as $\epsilon \to 0$.

Analogously, in the case $\Theta \subset \mathbb{R}^k, k > 1$, a normalizing matrix $\Psi(\epsilon)$ is called not better than $\varphi^{-1}(\epsilon)$ if for some constant c,

$$\|\Psi(\epsilon)\varphi(\epsilon)\| = \sup_{|x|=1} |\Psi(\epsilon)\varphi(\epsilon)x| \le c. \tag{2.49}$$

This definition is quite natural. Indeed, if for some family of random variables ξ_ϵ the family $x_\epsilon = \varphi^{-1}(\epsilon)\xi_\epsilon$ is compact, then it is evident from (2.49) that $\Psi(\epsilon)\xi_\epsilon = \Psi(\epsilon)\varphi(\epsilon)x_\epsilon$ is also compact, and therefore matrix $\Psi(\epsilon)$ "stretches" the vector ξ_ϵ in the order of magnitude not larger than matrix $\varphi^{-1}(\epsilon)$.

The following lemma shows that regular estimators with normalized matrix $\Psi(\epsilon)$ do not exist if the condition (2.49) is not fulfilled.

Lemma 2.5.1. *Let a family $P_t^{(\epsilon)}$ satisfy at point $\theta = t$ the LAN condition and let relation (2.16) be valid uniformly in $|u| < 1$. Furthermore, let the family of matrices $\Psi(\epsilon)$ be such that*

$$\|\Psi(\epsilon)\varphi(\epsilon)\| \to \infty \quad \epsilon \to 0. \tag{2.50}$$

Then there are no estimators of parameter θ such that for some proper distribution function $F(x)$,

$$\mathcal{L}(\Psi(\epsilon)[\theta_\epsilon - (t + \varphi(\epsilon)u)]|\tilde{P}^{(\epsilon)}_{t+\varphi(\epsilon)u}) \to F \tag{2.51}$$

as $\epsilon \to 0$ uniformly in $|u| < 1$.

Theorem 2.5.1. *Let the family $P_\theta^{(\epsilon)}$ satisfy the LAN condition for $\theta = t$ and let θ_ϵ be a family of estimators (possibly randomized) of the parameter θ which is regular at $\theta = t$.*

Then

1. *the limiting distribution $F(x)$ of the random vector $\zeta_\epsilon = \varphi^{-1}(\epsilon)(\theta_\epsilon - t)$ is a composition of $\mathcal{N}(0, J)$ and some other distribution $G(x) : F = \mathcal{N}(0, J) * G$;*
2. *$G(x)$ is the limiting distribution law of the difference $\zeta_\epsilon - \Delta_{\epsilon,t}$ as $\epsilon \to 0$.*

A refinement of Theorem 2.5.1 is the following.

Theorem 2.5.2. *Let the conditions of Theorem 2.5.1 be fulfilled. Then the random variables $\zeta_\epsilon - \Delta_{\epsilon,t}$ and $\Delta_{\epsilon,t}$ are asymptotically independent in the sense that the following weak convergence*

$$P_t^{(\epsilon)}(\zeta_\epsilon - \Delta_{\epsilon,t} < x, \Delta_{\epsilon,t} < y) \to G(x)\Phi(y) \tag{2.52}$$

is valid as $\epsilon \to 0$, here $\Phi(y)$ is the distribution function of the normal law $\mathcal{N}(0, J)$.

2.6 Asymptotic Efficiency under LAN Conditions

Various definitions of asymptotic efficiency were discussed in Chapter 1. Here, we shall prove some theorems interrelating these definitions using results proved in this chapter.

We know that in the case of iid observations, asymptotic efficiency in the Fisher sense reduces to the requirement of asymptotic normality of estimators with parameters $0, I^{-1}(\theta)$. The following definition which relates to a more general situation is in complete agreement with the classical one.

Definition 2.6.1 (Asymptotic Efficiency in Fisher's sense)**.** *Let a family of measures $P_\theta^{(\epsilon)}$ satisfy the LAN condition with the normalizing matrix $\varphi(\epsilon)$ at the point $\theta = t$. A family of estimators $\tilde{\theta}_\epsilon$ is called* asymptotically efficient in Fisher's sense *at the point $\theta = t$ if*

$$\mathcal{L}(\varphi^{-1}(\epsilon)(\tilde{\theta}_\epsilon - t)|P_t^\epsilon) \to \mathcal{N}(0, J) \tag{2.53}$$

as $\epsilon \to 0$.

J. Wolfowitz proposes a different definition of efficiency of statistical estimators. His reasoning is roughly as follows. Asymptotic efficiency in Fisher's sense is natural if we confine ourselves to estimators whose distribution uniformly converges to the limiting normal distribution with zero mean. However, there are not logical foundations for such a restriction because by enlarging the class of estimators one may possibly obtain better estimators in a certain sense. Of course, one cannot omit the requirement of uniform convergence due to the existence of superefficient estimators, although it may be reasonable to omit the requirement of asymptotic normality.

However, how can one compare two family of estimators $\theta_\epsilon^{(1)}, \theta_\epsilon^{(2)}$ where one is asymptotically normal but the other is not? Wolfowitz suggests comparing the quality of estimators by the degree of their "concentration" about the true value of the parameter. More precisely, in the case of a one-dimensional parameter space Θ, he proposes to consider as the better one, the family for which the $P_\theta^{(\epsilon)}$ probability that the estimator takes on a value in the interval $[\theta - a(\epsilon), \theta + a(\epsilon)]$ is the largest. However, two questions arise in this connection. First, how should one select the family $a(\epsilon)$? For overly small $a(\epsilon)$ all the estimators would be equally bad since the probability $P_\theta^{(\epsilon)}(\theta_\epsilon^{(i)} \in [\theta - a(\epsilon), \theta + a(\epsilon)])$ will be close to zero, while for $a(\epsilon)$ to large the proposed criterion will not be sensitive enough, since for too many families of estimators this probability will be close to one. If, however, a family of distributions $P_\theta^{(\epsilon)}$ satisfies the LAN condition it is then natural to put $a(\epsilon) = \lambda\varphi(\epsilon)$ which leads to Definition 2.6.2. The second question is what is the multidimensional analog of this method of comparing estimators? Kaufman suggested replacing symmetric intervals in the case $\Theta \subset \mathbb{R}^k$ by symmetric convex sets. We thus arrive at the following definition.

Definition 2.6.2 (Asymptotic Efficiency in Wolfowitz's sense)**.** *Let $\Theta \subset \mathbb{R}^k$ and the family of measures $P_\theta^{(\epsilon)}$ satisfy the LAN condition with the normalizing matrix $\varphi(\epsilon)$ at the point $\theta = t$. A family of estimators $\tilde{\theta}_\epsilon$ will be called* asymptotically efficient in Wolfowitz's sense *at the point $\theta = t$, if for any regular family T_ϵ and any centrally symmetric convex set $A \subset \mathbb{R}^k$ the relation*

$$\lim_{\epsilon\to 0} P_t^{(\epsilon)}(\varphi^{-1}(\epsilon)(\tilde{\theta}_\epsilon - t) \in A) \geq \limsup_{\epsilon\to 0} P_t^{(\epsilon)}(\varphi^{-1}(\epsilon)(T_\epsilon - t) \in A) \tag{2.54}$$

is valid.

Note that in this definition it is not required formally that the family $\tilde{\theta}_\epsilon$ be regular. However, this definition can hardly be considered natural if there exist no regular estimators which are efficient in Wolfowitz's sense. It will be shown below that both maximum likelihood and Bayesian estimators under quite general conditions are efficient in the Wolfowitz's sense as well as are regular ones. Here we shall present a sufficient condition for efficiency in Wolfowitz's sense.

Theorem 2.6.1. *If a family of estimators $\tilde{\theta}_\epsilon$ is asymptotically efficient in the sense of Definition 2.6.1, then it is also asymptotically efficient in the sense of Definition 2.6.2.*

Theorem 2.5.1 and results in the last section also allow us to obtain asymptotic bounds from below for the risk of regular estimators.

Theorem 2.6.2. *Let the family T_ϵ be regular at the point $t \in \mathbb{R}^k$ and $w(x) \geq 0, x \in \mathbb{R}^k$ be a continuous function satisfying*

1. *$w(x) = w(-x)$*
2. *the set $\{x : w(x) < c\}$ is convex in $\mathbb{R}^k$ for any $c > 0$. Then*

$$\liminf_{\epsilon\to 0} \mathbb{E}_t^{(\epsilon)} w[\varphi^{-1}(\epsilon)(T_\epsilon - t)] \geq \mathbb{E}w(\xi), \tag{2.55}$$

where $\mathcal{L}(\xi) = \mathcal{N}(0, J)$.

Corollary 2.6.1. *If the family T_ϵ is regular at the point $\theta = t$, then the matrix inequality*

$$\liminf_{\epsilon\to 0} [\varphi^{-1}(\epsilon)\mathbb{E}_t^{(\epsilon)}(T_\epsilon - t)(T_\epsilon - t)^T \varphi^{-t}(\epsilon)^T] \geq J \tag{2.56}$$

is valid.

Thus a regular estimator cannot have a covariance matrix which is "better" than the limiting covariance matrix of an asymptotically efficient estimator in Fisher's sense.

We now return to the definition of the asymptotic efficiency in Rao's sense.

Recall that a family of estimators θ_ϵ is asymptotically efficient in Rao's sense at the point $\theta = t$ if for some matrix $B(t)$ which does not depend on the observations the relation

$$\varphi^{-1}(\epsilon, t)(T_\epsilon - t) - B(t)\Delta_{\epsilon,t} \to 0 \tag{2.57}$$

in $P_t^{(\epsilon)}$-probability is valid as $\epsilon \to 0$.

If estimator T_ϵ is regular then Theorem 2.5.2 implies that the difference $\varphi^{-1}(\epsilon, t)(T_\epsilon - t) - \Delta_{\epsilon,t}$ is asymptotically independent of $\Delta_{\epsilon,t}$ and therefore in the case of regular estimators relation (2.57) can be fulfilled only if $B(t) = J$. It follows from here and from the second assertion of Theorem 2.5.1 that for regular estimators the asymptotic efficiency in Rao's sense coincides with the asymptotic efficiency in Wolfowitz's sense.

It follows from the above that according to the LAN condition it is natural to normalize the estimator by means of the factor $\varphi^{-1}(\epsilon,t)$. The corresponding loss function is $w_\epsilon(T_\epsilon - t) = w(\varphi^{-1}(\epsilon,t)(T_\epsilon - t))$.

Later it will be shown in Theorem 2.7.1 that for any function $w \in \mathbf{W}$ and any estimator $\tilde{\theta}_\epsilon$ under the LAN conditions the relation

$$\lim_{\delta\to 0}\liminf_{\epsilon\to 0}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}w(\varphi^{-1}(\epsilon,t)(\tilde{\theta}_\epsilon-\theta))\geq \mathbb{E}w(\xi),\quad \mathcal{L}(\xi)=\mathcal{N}(0,J) \tag{2.58}$$

is valid.

From (2.58) and (1.162) we obtain that the estimator T is asymptotically efficient for the loss function $w(\varphi^{-1}(\epsilon,t)x)$ at the point $\theta = t$ provided that

$$\lim_{\delta\to 0}\liminf_{\epsilon\to 0}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}w(\varphi^{-1}(\epsilon,t)(\tilde{\theta}_\epsilon-\theta))=\mathbb{E}w(\xi). \tag{2.59}$$

Below, we shall refer, under the LAN conditions, to an estimator T_ϵ which satisfies (2.59) as an *asymptotically efficient estimator for the loss function $w(\varphi^{-1}(\epsilon,t)x)$ at the point $\theta = t$.*

2.7 Asymptotically Minimax Risk Bound

In chapter 1, when we discuss about asymptotic efficiency, we established a theorem which yields an asymptotically minimax bound for the risks of arbitrary statistical estimators provided the asymptotic properties of Bayesian estimators are known. The LAN condition allows us to strengthen this result. The following interesting theorem is due to Hajek.

Theorem 2.7.1. *Let the family $P_\theta^{(\epsilon)}$ satisfy the LAN condition at the point $\theta = t$ with the normalizing matrix $\varphi(\epsilon)$ and let $\mathrm{tr}\varphi(\epsilon)\varphi(\epsilon)^T \to 0$ as $\epsilon \to 0$. Then for any family of estimators T_ϵ, any loss function $w \in \mathbf{W}_{\epsilon,2}$, and any $\delta > 0$ the inequality*

$$\liminf_{\epsilon\to 0}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))]\geq\frac{1}{(2\pi)^{k/2}}\int_{\mathbb{R}^k}w(x)e^{-|x|^2/2}dx=\mathbb{E}w(\xi) \tag{2.60}$$

is valid. Here $\mathcal{L}(\xi) = \mathcal{N}(0, J)$.

If, moreover, $\Theta \subset \mathbb{R}^1$ and $w(x) \in \mathbf{W}^1_{\epsilon,2}$, then the equality

$$\lim_{\delta\to 0}\lim_{\epsilon\to 0}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))]]=\mathbb{E}w(\xi) \tag{2.61}$$

is possible if and only if the difference $\varphi^{-1}(\epsilon)(T_\epsilon - t) - \Delta_{\epsilon,t} \to 0$ in $P_t^{(\epsilon)}$-probability as $\epsilon \to 0$.

Remark 2.7.1. *Since the first assertion of Theorem 2.7.1 is valid for any family of estimators T_ϵ, it can be written in the form*

$$\liminf_{\epsilon\to 0}\inf_{T_\epsilon}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}w[\varphi^{-1}(\epsilon)(T_\epsilon-\theta)]\geq\mathbb{E}w(\xi). \tag{2.62}$$

Thus Theorem 2.7.1 yields an asymptotic minimax bound from below for a wide class of loss functions. Below we shall see that in many important particular cases this bound is exact.

Remark 2.7.2. *Denote by K_b a cube in $\mathbb{R}^k$ whose vertices possess coordinates $\pm b$. If we drop the condition $\mathrm{tr}\varphi(\epsilon)\varphi(\epsilon)^T \to 0$, we could replace the basic inequality of Theorem 2.7.1 by the inequality*

$$\lim_{b\to\infty}\liminf_{\epsilon\to 0}\sup_{\theta:\varphi^{-1}(\epsilon)(\theta-t)\in K_b}\mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))]\geq\mathbb{E}w(\xi) \tag{2.63}$$

Moreover, it follows from the proof that for any $b > 0$, the inequality

$$\liminf_{\epsilon\to 0}\sup_{\theta:\varphi^{-1}(\epsilon)(\theta-t)\in K_b}\mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))]\geq\frac{1}{(2\pi)^{k/2}}\int_{K_{\sqrt{b}}}w(y)e^{-|y|^2/2}dy(1-b^{-1/2})^k \tag{2.64}$$

is valid.

Analogously one can somewhat strengthen the second assertion of the theorem also and replace it by the assertion that the equality

$$\lim_{b\to\infty}\lim_{\epsilon\to 0}\sup_{\theta:\varphi^{-1}(\epsilon)|\theta-t|\leq b}\mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))]=\mathbb{E}w(\xi) \tag{2.65}$$

is possible if and only if the difference $\varphi^{-1}(\epsilon)(T_\epsilon - t) - \Delta_{\epsilon,t} \to 0$ in $P_t^{(\epsilon)}$-probability as $\epsilon \to 0$.

Remark 2.7.3. *Inequality (2.64) presents a nontrivial bound from below only if $b > 1$. We could show the following coarse but nontrivial bound from below for any $b > 0$:*

$$\liminf_{\epsilon\to 0} \sup_{\theta:\varphi^{-1}(\epsilon)(\theta-t)\in K_b} \mathbb{E}_\theta^{(\epsilon)}[w(\varphi^{-1}(\epsilon)(T_\epsilon-\theta))] \geq 2^{-k}\frac{1}{(2\pi)^{k/2}}\int_{K_{b/2}} w(y)e^{-|y|^2/2}dy. \tag{2.66}$$

Remark 2.7.4. *It follows from the proof that the first assertion of Theorem 2.7.1 remains valid if in the left hand side of the basic inequality w is replaced by w_ϵ where $w_\epsilon \in \mathbf{W}_{l,2}$ which is the family of functions convergent to $w(x)$ for almost all x as $\epsilon \to 0$.*

2.8 Some Corollaries. Superefficient Estimators

Comparing Theorems 2.7.1 and 2.6.2, one arrives at the following conclusion: the asymptotic bound from below on the risks of regular estimators derived in section 2.6 is also the minimax asymptotic bound on the risks of arbitrary estimators. For example, setting $w(x) = I(x \in A^c)$ where A is a convex set in $\mathbb{R}^k$ symmetric with respect to the origin, we shall obtain from Theorem 2.7.1 the following assertion (see Theorem 2.6.1)

Theorem 2.8.1. *If the family $P_t^{(\epsilon)}$ satisfies the LAN condition at the point $\theta = t$ with the normalizing matrix $\varphi(\epsilon)$, then for any convex centrally-symmetric set A and any $\delta > 0$, the inequality*

$$\limsup_{\epsilon\to 0}\sup_{T_\epsilon}\inf_{|\theta-t|<\delta} P_\theta^{(\epsilon)}(\varphi^{-1}(\epsilon)(T_\epsilon-\theta)\in A) \leq \frac{1}{(2\pi)^{k/2}}\int_A e^{-|x|^2/2}dx \tag{2.67}$$

is valid.

Setting $w(x) = \langle h, x\rangle^2 = h^T x x^T h$ we shall obtain from Theorem 2.7.1 that under the LAN conditions the matrix inequality

$$\liminf_{\epsilon\to 0}\varphi^{-1}(\epsilon)\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}(T_\epsilon-\theta)(T_\epsilon-\theta)^T\varphi^{-1}(\epsilon)^T \geq J \tag{2.68}$$

is valid for any $\delta > 0$.

In accordance with the definition at the end of section 2.6, the estimator θ_ϵ is asymptotically efficient for the loss function $w(\varphi^{-1}(\epsilon, t)x)$ at point t provided

$$\lim_{\delta\to 0}\liminf_{\epsilon\to 0}\sup_{|\theta-t|<\delta}\mathbb{E}_\theta^{(\epsilon)}w(\varphi^{-1}(\epsilon,t)(\theta_\epsilon-\theta)) = \mathbb{E}w(\xi). \tag{2.69}$$

In agreement with chapter one we shall call the estimator T_ϵ a *superefficient estimator for the loss function* $w(\varphi^{-1}(\epsilon,\theta)x)$ in Θ provided for $\theta \in \Theta$

$$\lim_{\epsilon\to 0}\mathbb{E}_\theta^{(\epsilon)}w(\varphi^{-1}(\epsilon,\theta)(T_\epsilon-\theta)) \leq \mathbb{E}w(\xi), \quad \mathcal{L}(\xi) = \mathcal{N}(0, J) \tag{2.70}$$

and for at least one point $t \in \Theta$ the strict inequality

$$\lim_{\epsilon\to 0}\mathbb{E}_\theta^{(\epsilon)}w(\varphi^{-1}(\epsilon,\theta)(T_\epsilon-\theta)) < \mathbb{E}w(\xi) \tag{2.71}$$

is valid.

Stein's example shows that in the case when the dimension of the parameter space $k \geq 3$, there exist asymptotically efficient estimators for the loss function $w = |x|^2$ which are superefficient at one point of the set $\mathbb{R}^k$. Then the theorem shows that such a result is possible for $k = 1$ for *no* loss function belonging to a sufficiently wide class of functions.

Theorem 2.8.2. *If family $P_\theta^{(\epsilon)}$ satisfies the LAN condition at the point $\theta = t$, $\Theta \subset \mathbb{R}^1$ and T_ϵ is an asymptotically efficient estimator at the point $\theta = t$ for some loss function $w_0 \in \mathbf{W}'$, then T_ϵ cannot be superefficient at this point for any loss function $w \in \mathbf{W}$.*

Theorem 2.7.1 allows us to investigate properties of superefficient estimators also in the case of multidimensional parameter set. We have the following theorem.

Theorem 2.8.3. *Let T_ϵ be an estimator of the vector $\theta \in \Theta \subset \mathbb{R}^k$ in the parametric family $P_\theta^{(\epsilon)}$ satisfying the LAN condition at $\theta = t$. Assume that every component of the vector T_ϵ is an asymptotically efficient (at the point $\theta = t$) estimator of the corresponding component of the vector θ for some loss function $w_0 \in \mathbf{W}'$. Then the estimator T_ϵ can be superefficient at point $\theta = t$ for* no *loss function $w(x) \in \mathbf{W}, x \in \mathbb{R}^k$.*

Corollary 2.8.1. *The components of Stein's vector estimator are not asymptotically efficient estimators of the components of the mean value vector in a multivariate normal distribution provided $k \geq 3$.*

Chapter 3

Some Applications to Nonparametric Estimation

Nonparametric estimation is a large branch of mathematical statistics dealing with problems of estimating functionals of elements of some functional spaces in situations when these are not determined by specifying a finite number of parameters. In this chapter we shall show by means of several examples how the ideas of parametric estimation presented in previous chapters can be applied to problems of this kind.

3.1 A Minimax Bound on Risks

The problem of parametric estimation can be considered as a particular case of the following more general statistical problem, which we shall, for the time being, consider only for iid observations. Thus let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables with values from a measurable space $(\mathscr{X}, \mathcal{X})$ and let $F(\cdot)$ be their common (unknown to the observer) distribution which belongs to some (known) class of distributions $\mathbf{F}$ on $\mathcal{X}$. Let $\Phi(F)$ be a real functional. The problem consists of estimating the functional $\Phi(F)$ based on observations $X_1, X_2, \ldots, X_n$. We shall basically be concerned with the bounds on the precision of estimation and with asymptotically best estimators.

In particular, if the class $\mathbf{F}$ is one-parametric, $\mathbf{F} = \{F(\cdot;\theta), \theta \in \Theta\}$ and $\Phi(F(\cdot;\theta)) = \theta$, then we arrive at the usual problem of estimating one-parametric distributions. Retaining the same set $\mathbf{F}$ and considering different functionals Φ, we arrive at the problem of estimating a function of a parameter θ. However, more interesting examples can be obtained by considering other classes of distributions $\mathbf{F}$.

Example 3.1.1. *Let* $\mathbf{F}$ *be a subset of the family of distributions such that the integral*

$$\mathbb{E}_F|\varphi(X)| = \int |\varphi(x)|F(dx), \quad \varphi : \mathscr{X} \to \mathbb{R}^1 \tag{3.1}$$

is finite and $\Psi(F) = \int \varphi(x)F(dx)$*. Evidently, in this case the following statistic is a rather "good" estimator of functional* $\Phi(F)$*:*

$$\hat{\varphi}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i). \tag{3.2}$$

Will this estimator be the best in a certain sense? An answer to this question depends, of course, on how "extended" the family $\mathbf{F}$ *is. For example, if* $\mathbf{F}$ *is a parametric set, then this estimator is not in general the best in any asymptotic sense (consider the uniform location model).*

Nevertheless, it follows from the theorems proved below that for a sufficiently "substantial" set $\mathbf{F}$ *this empirical mean estimator can not be asymptotically improved for all* $F \in \mathbf{F}$*.*

Example 3.1.2. *Consider one of the possible generalizations of the preceding example. Let* $\mathbf{F}$ *be a subset of a class of distributions for which the functionals*

$$\int |\varphi_i(x)|F(dx), \quad i = 1, 2, \ldots, r \quad \varphi_i : \mathscr{X} \to \mathbb{R}^1, \tag{3.3}$$

are finite and $\varphi_0 : \mathbb{R}^r \to \mathbb{R}^1$ *is a sufficiently smooth function. Consider the functional*

$$\Phi(F) = \varphi_0\left(\int \varphi(x)F(dx)\right), \tag{3.4}$$

in which $\varphi(x)$ is a vector in $\mathbb{R}^r$ with coordinates $\varphi_1(x), \ldots, \varphi_r(x)$. Methods for constructing asymptotically efficient (and minimax) estimators for smooth parametric families $\mathbf{F}$ were considered above. But how can one construct an asymptotically best (in a certain sense) estimator of the functional $\Phi(F)$ if the function F is known only approximately with precision up to a neighborhood of some fixed function F_0 in the corresponding functional space? It will be shown that such an estimator is, for example, a function of the arithmetic mean

$$\Phi_n = \varphi_0(\hat{\varphi}_n) \tag{3.5}$$

Example 3.1.3. *Let $\mathbf{F}$ be a contraction of the set of distributions on $\mathbb{R}^1$ for which the median t is uniquely determined as the solution of the equation $F(t) = 1/2$ and $\Phi(F) = \text{med } F$.*

A natural nonparametric estimator of the value of med F is the sample median, which is the $[n/2]$-th order statistic. It follows from the arguments presented in this and the succeeding sections that this estimation is asymptotically minimax if the class $\mathbf{F}$ is sufficiently "massive" in the neighborhood of the given distribution F_0.

In this present section, under some restrictions on the regularity of $\mathbf{F}$ and Φ, a minimax bound from below on the quality of nonparametric estimators will be derived. The idea of the arguments presented below was heuristically stated by Stein and was developed in detail in a number of papers by Levit. This is an extremely simple idea.

Let F be a distribution in $\mathbf{F}$: denote $\Phi(F)$ by t. Consider a smooth parametric family $\varphi = \{F_h(x)\} \in \mathbf{F}$ which "passes" through the "point" F at $h = t(F_t = F)$ and such that the value of the parameter h on this family coincides with the value of the functional Φ in some neighborhood of $h = t$, i.e. $\Phi(F_h) = h$. The smoothness of the family φ will be, for the time being, interpreted in the sense that there exists Fisher information quantity $I(F, \varphi)$ for this family and that the LAN condition with normalization $(I(F, \varphi)n)^{-1/2}$ is satisfied.

Now it is easy to obtain a certain minimax bound on the risks for the problem of estimating the functional $\Phi(F)$ with a loss function $w \in \mathbf{W}$. Indeed, for any estimator Φ_n of the functional Φ, for some $\delta > 0$ the inequalities

$$\sup_{F \in \mathbf{F}} \mathbb{E}_F w(\sqrt{n}(\Phi_n - \Phi(F))) \geq \sup_{\{F_h\}} \mathbb{E}_{F_h} w(\sqrt{n}(\Phi_n - \Phi(F_h))) \geq \sup_{|h-t|<\delta} \mathbb{E}_h w(\sqrt{n}(\Phi_n - h)) \tag{3.6}$$

are self-evident.

In view of Theorem 2.7.1, we have for any $\delta > 0$,

$$\liminf_{n\to\infty} \inf_{\Phi_n} \sup_{|h-t|<\delta} \mathbb{E}_h w(\sqrt{n}(\Phi_n - h)) \geq \frac{1}{\sqrt{2\pi}} \int w(xI^{-1/2}(F, \varphi)) e^{-x^2/2} dx \tag{3.7}$$

and we have obtained the derived bound from below on the risks. This bound depends on the choice of the smooth family $\varphi = \{F_h\} \subset \mathbf{F}$ with the properties indicated above. Clearly, the smaller $I(F; \varphi)$ is this bound is.

Remark 3.1.1. *The most stringent requirement on the family φ is the requirement that $\Phi(F_h) = h$. We shall show that the inequality*

$$\liminf_{n\to\infty} \inf_{\Phi_n} \sup_{F \in \mathbf{F}} w(\sqrt{n}(\Phi_n - \Phi(F))) \geq \frac{1}{\sqrt{2\pi}} \int w(xI^{-1/2}(F, \varphi)) e^{-x^2/2} dx \tag{3.8}$$

remains valid if this requirement is replaced by a somewhat weaker one:

$$\Phi(F_h) = h + o(h - t) \tag{3.9}$$

as $h \to t$.

For this purpose, it is evidently sufficient to check that

$$\liminf_{n\to\infty} \sup_{|h-t|<\delta} \mathbb{E}_h w(\sqrt{n}(\Phi_n - h - o(h - t))) \geq \frac{1}{\sqrt{2\pi}} \int w(xI^{-1/2}(F, \varphi)) e^{-x^2/2} dx. \tag{3.10}$$

The proof of this refined version of Theorem 2.7.1 does not differ from the proof of Theorem 2.7.1 if one observers that function $w \in \mathbf{W}$ is continuous a.e. in $\mathbb{R}^1$.

As a result of this discussion the following definition would seem natural.

Consider all possible families $\varphi = \{F_h\} \in \mathbf{F}$ parametrized by a real parameter h with the properties

1. $F_t = F$
2. $\Phi(F_h) = h + o(h - t)$ as $h \to t$

3. The random variable

$$\eta_u = \left[\frac{dF^c_{t+u}}{dF_t}(X)\right]^{1/2} \tag{3.11}$$

possess a derivative with respect to u in the mean square sense at $u = 0$, so that for some random variable $\dot\eta_0$

$$\lim_{u\to 0}\frac{1}{u^2}\mathbb{E}_F\left[\left(\frac{dF^c_{t+u}}{dF_t}(X)\right)^{1/2} - 1 - u\dot\eta_0\right]^2 = 0 \tag{3.12}$$

and, moreover

$$\lim_{u\to 0}\frac{1}{u^2}\int(\sqrt{dF_{t+u}(x)} - \sqrt{dF_t(x)})^2 = \mathbb{E}_F\dot\eta_0^2 < \infty. \tag{3.13}$$

Let $I(F;\varphi) = 4\mathbb{E}_F\dot\eta_0^2$.

Definition 3.1.1 (Information Quantity in Nonparametric Estimation Problems)**.** *The quantity*

$$\inf_{\varphi} I(F_0,\varphi) = I(F_0) \tag{3.14}$$

where φ belongs to $\mathbf{F}$ and satisfies conditions above is called the information quantity *in the estimation problem $\Phi(F), F\in\mathbf{F}$, ad the point $F = F_0$. If there are no parametric families $\varphi\in\mathbf{F}$ which satisfies conditions above we shall then set $I(F_0) = \infty$.*

It follows from this definition that in the case $I(F_0) < \infty$ there exists a sequence of parametric families $\varphi_N = \{F^N_h\}, N = 1, 2, \dots$ satisfying conditions above such that

$$I(F_0,\varphi_N)\to I(F_0),\quad N\to\infty. \tag{3.15}$$

We have the following theorem.

Theorem 3.1.1. *If $I(\mathbf{F}) = \inf_{F\in\mathbf{F}} I(F) > 0$, then for the problem of estimating the functional $\Phi(F), F\in\mathbf{F}$, the following minimax bound from below on the risks for any loss function $w\in\mathbf{W}$ is valid:*

$$\liminf_{n\to\infty}\inf_{\Phi_n}\sup_{F\in\mathbf{F}}\mathbb{E}_F w(\sqrt{n}(\Phi_n - \Phi(F))) \ge \sup_{F_0\in\mathbf{F}}\frac{1}{\sqrt{2\pi}}\int w(xI(F_0)^{-1/2})e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int w(xI(\mathbf{F})^{-1/2})e^{-x^2/2}dx. \tag{3.16}$$

The quantity $I(F_0)$ appearing in this theorem evidently depends also on the set $\mathbf{F}$. Moreover the quantity $I(F_0)$ in general is not determined by the local structure of the set $\mathbf{F}$ in the neighborhood of F_0.

It we want to localize the assertion of Theorem 3.1.1, i.e., to obtain a nonparametric bound on the risks from below analogous to Hajek's bound in the neighborhood of a given distribution (see Theorem 2.7.1), then it is necessary first of all to decide which neighborhoods of the given distribution, i.e., which topologies on the set of distributions will be considered.

In the parametric case the specification of a topology by means of an Euclidean metric on the set Θ is sufficiently natural. Unfortunately, in a nonparametric situation there is, in general, no such natural method of defining a topology although some requirements related to this topology follow from the actual problem.

First, the topology should be such that the estimating functional $\Phi(F)$ be continuous in this topology since otherwise a consistent estimation is impossible. Other requirements on this topology follow from the properties of nonparametric information quantity. Next, if we want that quantity $I(F_0)$ to be defined by the local structure of the set $\mathbf{F}$ in the neighborhood of a given distribution, it is then necessary to require that any family F_h satisfying conditions above will be continuous in the topology under consideration. In this connection we introduce the following definition.

Definition 3.1.2 (Coordinated Topology)**.** *Topology R is called* coordinated *with the estimation problem $(\Phi(F), F\in\mathbf{F})$ if*

1. *functional $\Phi(F)$ is continuous on $\mathbf{F}$ in this topology*

2. *any family $\varphi = \{F_h\}$ satisfying the three natural conditions above is continuous for $h = t = \Phi(F)$ in this topology.*

For not too degenerate estimation problems, the choice of topologies R coordinated with the estimation problem is sufficiently large. We shall not dwell on this, but remark that for any estimation problem the second requirement in Definition 3.1.2 is fulfilled for the topology defined by Hellinger's distance ρ_0 and for any weaker topology. Indeed for $|h - t| < \delta$,

$$\rho_0(F_h, F_t) = \int(\sqrt{dF_h} - \sqrt{dF_t})^2 \le c(h-t)^2 \tag{3.17}$$

in view of the third conditions in the three natural conditions, and therefore any family satisfying the third condition in the three natural conditions is continuous at point $h = t$ in this topology. The first condition in Definition 3.1.2 is also generally not too restrictive.

If we confine ourselves to neighborhoods $U(F_0) \subset \mathbf{F}$ of a fixed distribution F_0 in topologies coordinated with the estimation problem $\Phi(F), F \in \mathbf{F}$, then the arguments leading to Theorem 3.1.1 can be localized, since in this case in the left hand side of (3.6) instead of considering the upper bound over $F \in \mathbf{F}$ we can consider the upper bound over $F \in U(F_0)$. We thus arrive at the local version of Theorem 3.1.1.

Theorem 3.1.2. *Assume $I(F_0) > 0$ and let $U_N(F_0)$ be an arbitrary sequence of neighborhoods of distribution F_0 in the topology coordinated with the estimation problem $\Phi(F), F \in \mathbf{F}$, such that $U_N(F_0) \downarrow F_0$ as $N \to \infty$. Then for any loss function $w \in \mathbf{W}$ the following asymptotically minimax bound on risk is valid:*

$$\lim_{N\to\infty} \liminf_{n\to\infty} \inf_{\Phi_n} \sup_{F\in U_N(F_0)} \sup_{F\in U_N(F_0)} \mathbb{E}_F w(\sqrt{n}(\Phi_n - \Phi(F))) \geq \frac{1}{\sqrt{2\pi}} \int w(xI^{-1/2}(F_0))e^{-x^2/2}dx. \tag{3.18}$$

In accordance with the point of view of this book, an estimator $\tilde{\Phi}_n$ such that for $F_0 \in \mathbf{F}$ and any sequence $U_N(F_0)$ of neighborhoods converging to F_0 in the topology R the relation

$$\lim_{N\to\infty} \liminf_{n\to\infty} \sup_{F\in U_N(F_0)} \sup_{F\in U_N(F_0)} \mathbb{E}_F w(\sqrt{n}(\tilde{\Phi}_n - \Phi(F))) = \frac{1}{\sqrt{2\pi}} \int w(xI^{-1/2}(F_0))e^{-x^2/2}dx. \tag{3.19}$$

is valid will be called a $(\mathbf{F}, R, w)$-asymptotically efficient nonparametric estimator of the functional $\Phi(F)$ at point F_0.

In connection with these definitions and Theorems 3.1.1 and 3.1.2 several questions are:

1. How to compute $I(F)$ for a given problem of nonparametric estimation
2. Is the bound (3.18) attainable, i.e., are there asymptotically efficient nonparametric estimators for a given estimation problem?
3. If the answer to the second question is positive, for which estimators is the bound (3.18) attainable in specific cases?

Answers to these questions are closely interconnected. Indeed the inequality $I(F) \leq I(F, \varphi)$ follows from Definition 3.1.1.

On the other hand, if for some positive functional $A(F)$ continuous in the topology R, a family of estimators $\tilde{\Phi}_n$ is found such that for some loss function $w \in \mathbf{W}$ and some domain $U \subset \mathbf{F}$,

$$\limsup_{n\to\infty} \sup_{F\in U} \left| \mathbb{E}_F w(\sqrt{n}(\tilde{\Phi}_n - \Phi(F))) - \frac{1}{\sqrt{2\pi}} \int w(xA^{-1/2}(F))e^{-x^2/2}dx \right| = 0, \tag{3.20}$$

then it follows from (3.18) and the monotonicity of w that $A(F_0) \leq I(F_0)$.

Thus if it is possible to construct a sequence of parametric families $\varphi_r = \{F_h^r\}, r = 1, 2, \ldots$ such that the corresponding information quantities $I(F_0, \varphi_r)$ converge to $A(F_0)$ as $r \to \infty$, and a sequence of estimators $\tilde{\Phi}_n$ satisfying relation (3.20), then $I(F_0) = A(F_0)$ and $\tilde{\Phi}_n$ is an $(\mathbf{F}, R, w)$-asymptotically efficient in U nonparametric estimator. We shall adhere to this outline of investigation of properties of nonparametric estimators in the next sections for the examples considered above.

3.2 Bounds on Risks for Some Smooth Functionals

Definition 3.2.1 (Differentiability in von Mises' sense)**.** *Let $\mathbf{F}$ be a set of distributions on $(\mathscr{X}, \mathcal{X})$ where for any $F_1 \in \mathbf{F}, F_2 \in \mathbf{F}$, $h \in (0, 1)$ the distribution $(1-h)F_1 + hF_2 \in \mathbf{F}$. A functional $\Phi(F)$, $F \in \mathbf{F}$ is* differentiable in von Mises' sense *in $\mathbf{F}$ if for any distributions $F_1 \in \mathbf{F}, F_2 \in \mathbf{F}$, and for some functional $l(F, y), F \in \mathbf{F}, y \in \mathscr{X}$, the equality*

$$\Phi(F_1 + h(F_2 - F_2)) = \Phi(F_1) + h \int l(F_1, y)\,(F_2(dy) - F_1(dy)) + o(h) \tag{3.21}$$

is valid as $h \to 0$.

For differentiable functionals $\Phi(F)$ one can find a class of parametric families satisfying the three natural conditions. This class is convenient because the problem of minimization of the corresponding information quantity $I(F, \varphi)$ is easily solved in this class. Evidently, in general, minimization with respect to this class may not lead to $I(F)$, the nonparametric information quantity. However, in almost all known examples, the bound on the quality of estimation thus obtained is asymptotically the best.

Thus suppose we solve the estimation problem of a differentiable, in von Mises' sense, functional $\Phi(F), F \in \mathbf{F}$. Consider a parametric family of distributions $\{F_h\}$ defined by the equality

$$F_h(dx) = F(dx)[1 + (h - t)\psi(x)], t = \Phi(F). \tag{3.22}$$

Clearly, the conditions

$$\int \psi(x)F(dx) = 0, \quad |\psi(x)| < N \tag{3.23}$$

are sufficient for $F_h(dx) = F(dx)[1 + (h - t)\psi(x)]$ to define a probability measure for $|h - t| < \delta = \delta(N)$. Assume also that $F_h \in \mathbf{F}$ for $|h - t| < \delta(N)$ for any ψ satisfying (3.23) with $N > 0$.

The first natural condition in the first section is automatically fulfilled for the family (3.22). Setting

$$F_1(\Gamma) = F(\Gamma), \quad F_2(\Gamma) = F(\Gamma) + \frac{1}{N}\int_\Gamma \psi(x)F(dx), \tag{3.24}$$

we obtain from (3.21) that

$$\Phi(F_h) = \Phi(F_1 + (h - t)N(F_2 - F_1)) = \Phi(F) + (h - t)\int l(F, y)\psi(y)F(dy) + o(h - t). \tag{3.25}$$

This implies that the second natural condition in the first section is fulfilled for the family (3.22) provided

$$\int l(F, y)\psi(y)F(dy) = 1. \tag{3.26}$$

Furthermore, since under condition (3.23) we have for the family (3.22)

$$\left[\frac{dF_h}{dF}(x)\right]^{1/2} = [1 + (h - t)\psi(x)]^{1/2} = 1 + 1/2(h - t)\psi(x) + o(h - t), \tag{3.27}$$

the third natural condition is also fulfilled and, moreover,

$$\dot{\eta}_0 = \frac{1}{2}\psi(x), \quad I(F, \varphi) = \int \psi^2(x)F(dx). \tag{3.28}$$

Equations (3.23) and (3.26) yield

$$\int [l(F, y) - \mathbb{E}_F l(F, X)]\psi(y)F(dy) = 1. \tag{3.29}$$

From here and the Cauchy-Schwarz inequality, we obtain the following bound from below on the information quantities of parametric families of the form (3.22):

$$I(F, \varphi) = \int \psi^2(x)F(dx) \geq \left[\int [l(F, y) - \mathbb{E}_F l(F, X)]^2 F(dy)\right]^{-1} = \sigma^{-2}(l, F). \tag{3.30}$$

If the functional $l(F, x)$ is bounded in x, then setting

$$\psi(x) = \psi_0(x) = (l(F, x) - \mathbb{E}_F l(F, X))\sigma^{-2}(l, F), \tag{3.31}$$

we arrive at a parametric family for which the lower bound (3.30) is attained.

Assume now that functional $l(F, y)$ is unbounded but square integrable with respect to the measure F. We show that in this case there exists a sequence of parametric families of the form (3.22) whose information quantities are arbitrarily close to $\sigma^{-2}(l, F)$. Let $l^{(N)}(F, \cdot)$ be a sequence of bounded functions converging to $l(F, \cdot)$ in $L_2(F)$ as $N \to \infty$. Then as $N \to \infty$, the following relations are obviously valid:

$$\mathbb{E}_F l^{(N)}(F, X) = \int l^{(N)}(F, x)F(dx) \to \mathbb{E}_F l(F, X) \tag{3.32}$$

$$\sigma_N^2(l, F) = \int (l(F, x) - \mathbb{E}_F l(F, X))(l^{(N)}(F, x) - \mathbb{E}_F l^{(N)}(F, X))F(dx) \to \sigma^2(l, F). \tag{3.33}$$

Clearly the parametric family (3.22) in which $\psi(x)$ is of the form

$$\psi^{(N)}(x) = (l^{(N)}(F, x) - \mathbb{E}_F l^{(N)}(F, X))\sigma_N^{-2}(l, F) \tag{3.34}$$

satisfies (3.23) to (3.26), moreover (3.33) easily yields that

$$I(F,\psi^{(N)}) \to \sigma^{-2}(l,F). \tag{3.35}$$

as $N \to \infty$.

Relation (3.35) and Definition 3.1.1 imply the inequality

$$I(F) \le \sigma^{-2}(l,F), \tag{3.36}$$

provided the parametric family (3.22) with $\psi^{(N)}(x)$ in place of $\psi(x)$ belongs to $\mathbf{F}$ for $N>0, |h-t|<\delta(N)$. This inequality, together with Theorem 3.1.2, yields the following assertion:

Theorem 3.2.1. *If the functional $\Phi(F)$ is differentiable in the sense of (3.21) and $l(F,\cdot)\in L_2(F)$, then family (3.22) with $\psi=\psi^{(M)}, M>0$, belongs to $\mathbf{F}$ for all $|h-\Phi(F_0)|<\delta(M), \delta(M)>0, w\in\mathbf{W}$, and the sequence of neighborhoods $U_N(F_0)$ and topology R satisfy the conditions of Theorem 3.1.2, then*

$$\lim_{N\to\infty}\liminf_{n\to\infty}\inf_{\Phi_n}\sup_{F\in U_N(F_0)}\sup_{F\in U_N(F_0)}\mathbb{E}_F w(\sqrt{n}(\Phi_n-\Phi(F))) \ge \frac{1}{\sqrt{2\pi}}\int w(x\sigma(l,F_0))e^{-x^2/2}dx. \tag{3.37}$$

where $\sigma^2(l,F)=\int[l(F,y)-\mathbb{E}_F l(F,X)]^2F(dy)$.

The functional $\sigma^2(l,F_0)$ as well as the bound (3.37) can be computed in many specific cases without much difficulty.

Example 3.2.1. *Consider the functional $\Phi(F)$ given in Example 3.1.1 on the set $\mathbf{F}_2$ of distributions F such that $\int|\varphi(x)|^2F(dx)<\infty$. In this case, due to the linearity of Φ, we have for any $F_1,F_2\in\mathbf{F}_2$,*

$$\Phi(F_1+h(F_2-F_1)) = \Phi(F_1)+h\int\varphi(x)[F_2(dx)-F_1(dx)]. \tag{3.38}$$

Therefore $l(F,x)=\varphi(x)$,

$$\sigma^2(l,F)=\int[\varphi(x)-\mathbb{E}_F\varphi(X)]^2F(dx). \tag{3.39}$$

Example 3.2.2. *Let*

$$\Phi(F)=\int\cdots\int\varphi(x_1,x_2,\dots,x_m)F(dx_1)\cdots F(dx_m), \tag{3.40}$$

where $\varphi(x_1,x_2,\dots,x_m)$ is a symmetric function of $x_1,x_2,\dots,x_m$ and $\mathbf{F}$ is the set of distributions such that

$$\int\cdots\int|\varphi(x_1,x_2,\dots,x_m)|^2F(dx_1)\cdots F(dx_m)<\infty. \tag{3.41}$$

In this case it is easy to verify that

$$\Phi(F_1+h(F_2-F_1)) = \Phi(F_1)+mh\int\cdots\int\varphi(y,x_2,\dots,x_m)\times F_1(dx_2)\cdots F_1(dx_m)[F_2(dy)-F_1(dy)]+o(h). \tag{3.42}$$

Therefore,

$$l(F,x)=m\int\cdots\int\varphi(y,x_2,\dots,x_m)\times F(dx_2)\cdots F(dx_m) \tag{3.43}$$

$$\sigma^2(l,F)=m^2\int\left[\int\cdots\int\varphi(x,x_2,\dots,x_m)F(dx_2)\cdots F(dx_m)-\Phi(F)\right]^2F(dx). \tag{3.44}$$

Example 3.2.3. *Consider the functional given in Example 3.1.2 on the set $\mathbf{F}_2$ of Example 3.1.1. Recall that in this case $\varphi:\mathscr{X}\to\mathbb{R}^r$. Assume that the function φ_0 is continuously differentiable. Under these assumptions we obtain from Taylor's formula*

$$\Phi(F_1+h(F_2-F_1))-\Phi(F_1) = \varphi_0\left[\int\varphi(x)F_1(dx)+h\int\varphi(x)(F_2(dx)-F_1(dx))\right]-\varphi_0\left(\int\varphi(x)F_1(dx)\right) \tag{3.45}$$

$$= h\left\langle\nabla\varphi_0\left(\int\varphi(x)F_1(dx)\right),\int\varphi(y)(F_2(dy)-F_1(dy))\right\rangle+o(h). \tag{3.46}$$

Thus,

$$l(F,y)=\langle\nabla\varphi_0(\mathbb{E}_F\varphi(X)),\varphi(y)\rangle \tag{3.47}$$

and therefore

$$\sigma^2(l,F)=\int\langle\varphi(y)-\mathbb{E}_F\varphi(X)),\nabla\varphi_0(\mathbb{E}_F\varphi(X))\rangle^2F(dy). \tag{3.48}$$

Example 3.2.4. *The functional med F considered in Example 3.1.3 under the conditions stipulated there can be defined as the root $t = \Phi(F)$ of the equation*

$$\int \operatorname{sign}(x-t)F(dx) = 0. \tag{3.49}$$

Consider the more general functional $\Phi(F)$ which represents the root $t = \Phi(F)$ of the equation

$$\mathbb{E}_F\varphi(X,t) = \int \varphi(x,t)F(dx) = 0, \tag{3.50}$$

where $\varphi : \mathscr{X} \times \mathbb{R}^1 \to \mathbb{R}^1$ and the function φ and set $\mathbf{F}$ are such that

1. *for any distribution $F \in \mathbf{F}$, (3.50) possesses the unique solution $\Phi(F)$*
2. *the function $\lambda_F(h) = \mathbb{E}_F\varphi(X,h)$ is continuous in h for $F \in \mathbf{F}$, differentiable for $h = t = \Phi(F), F \in \mathbf{F}$, and the derivative $\lambda'_F(F) \neq 0$.*
3. *for $F \in \mathbf{F}$, the relation $\mathbb{E}_F|\varphi(X,\Phi(F))|^2 < \infty$ is satisfied*
4. *for any $F_1, F_2 \in \mathbf{F}$, the relation $\Phi(F_1 + h(F_2 - F_1)) \to \Phi(F_1)$ as $h \to 0$ is valid.*

From the equalities

$$\int \varphi(x,\Phi(F_1))F_1(dx) = 0 \tag{3.51}$$

$$\int \varphi(x,\Phi(F_1 + h(F_2 - F_1)))[F_1(dx) + h(F_2(dx) - F_1(dx))] = 0 \tag{3.52}$$

and the conditions 1,2,4 we easily obtain that

$$\lambda_{F_1}[\Phi(F_1 + h(F_2 - F_1))] - \lambda_{F_1}(\Phi(F_1)) = (\lambda'_{F_1}(\Phi(F_1)) + o(h))(\Phi(F_1 + h(F_2 - F_1)) - \Phi(F_1)) \tag{3.53}$$

$$= -h\int \varphi[x,\Phi(F_1 + h(F_2 - F_1))](F_2(dx) - F_1(dx)). \tag{3.54}$$

Here (3.53) follows from the Taylor formula for $\lambda_{F_1}(h)$ at point $h = \Phi(F_1)$, and (3.54) follows from the fact that $\lambda_{F_1}(\Phi(F_1)) = 0$ and (3.52).

Utilizing once again conditions 2 and 4, we arrive at the equalities

$$l(F,y) = -\frac{\varphi(y,\Phi(F))}{\lambda'_F(\Phi(F))} \tag{3.55}$$

$$\sigma^2(l,F) = \int \varphi^2(y,\Phi(F))F(dy)[\lambda'_F(\Phi(F))]^{-2}. \tag{3.56}$$

Condition 4 can be replaced by the following: the function $\varphi(x,h)$ varies monotonically in h. If this function is strictly monotone in h, then condition 1 can be omitted. In particular, for a functional of the type considered in Example 3.1.3, conditions 1-4 are fulfilled if $\mathbf{F}$ is the set of distributions possessing positive density $f(x)$ with respect to Lebesgue measure on the real line. Moreover,

$$\lambda'_F(med\ F) = 2f(med\ F), \quad \sigma^2(l,F) = [4f^2(med\ F)]^{-1}. \tag{3.57}$$

In the conclusion of this section we shall consider two examples in which the lower bound on risks is obtained not from Theorem 3.2.1 but as a direct choice of a suitable parametric family. The first example may be viewed as a generalization of this example.

Example 3.2.5. *Let the functions $\varphi_1(x,h), \dots, \varphi_r(x,h) : \mathbb{R}^1 \to \mathbb{R}^1 \to \mathbb{R}^1$ and the family of distributions $\mathbf{F}$ on the real line be such that the following conditions are satisfied:*

1. *For any $F \in \mathbf{F}$, the equations*

$$\int \varphi_i(x,t)F(dx) = 0, \quad i = 1,2,\dots,r, \tag{3.58}$$

have a common solution $t = \Phi(F)$ and, moreover, this solution is unique;

2. *The functions $\lambda^i_F(h) = \mathbb{E}_F\varphi_i(X,h)$ are continuous in h for $F \in \mathbf{F}$ and the derivatives $\frac{d\lambda^i_F}{dh}$ exist at $h = \Phi(F)$*

3. *The functions $\varphi_i(x,h) \in L_2(F)$ for all $F \in \mathbf{F}$ and are continuous with respect to h in $L_2(F)$ at $h = t = \Phi(F)$*

4. *The determinant $|B_F(\Phi(F))| \neq 0$, where $B_F(h) = \|b_F^{ij}(h)\|, b_F^{ij}(h) = \int \varphi_i(x,h)\varphi_j(x,h)F(dx)$. Note that the determinant B_F is evidently the Gram determinant and thus this condition is the condition of linear independence of the functions $\varphi_i(x,\Phi(F))$ in $L_2(F)$ for $F \in \mathbf{F}$.*

To avoid some technical difficulties we shall assume, in addition, that the functions φ_i are bounded for $x \in \mathbb{R}^1, h \in \mathbb{R}^1$, and we shall seek for a parametric family $F_h \in \mathbf{F}$ in the form

$$F_h(dx) = F(dx)[1 + (h-t)\psi(x,h)] \tag{3.59}$$

$$\psi(x,h) = \sum_{j=1}^{r} \gamma_i(h)[\varphi_j(x,h) - \mathbb{E}_F\varphi_j(X,h)]. \tag{3.60}$$

The function $\psi(x,h)$ for any h satisfies the condition (3.23) and hence F_h is a distribution. The equalities

$$\int \varphi_i(x,h)[1 + (h-t)\psi(x,h)]F(dx) = 0, \quad i = 1,2,\ldots,r \tag{3.61}$$

follows from the condition $\Phi(F_h) = h$. From here and from (3.58), noting the choice of $\psi(x,h)$, we arrive at a system of equations for the coefficients $\gamma_1(h),\ldots,\gamma_r(h)$ (which so far are not determined):

$$\sum_{j=1}^{r}[b_F^{ij}(h) - \mathbb{E}_F\varphi_i(X,h)\mathbb{E}_F\varphi_j(X,h)]\gamma_j(h) = -\frac{\lambda_F^i(h) - \lambda_F^i(t)}{h-t}. \tag{3.62}$$

It follows from the four conditions in this example that for $|h-t| < \delta$, the system (3.62) possesses a unique solution which converges as $h \to t$ to the unique solution of the system

$$\sum_{j=1}^{r} b_F^{ij}(t)\gamma_j(t) = -\frac{d\lambda_F^i(t)}{dt}, \quad i = 1,2,\ldots,r. \tag{3.63}$$

Let the function $\psi(x,h)$ and the family of distributions $F_h, |h-t| < \delta$ be chosen in accordance with (3.60) with functions $\gamma_j(h)$ defined by (3.63) and assume that this family F_h belongs to $\mathbf{F}$.

We obtain from (3.60)

$$\left[\frac{dF_h}{dF}(x)\right]^{1/2} = 1 + \frac{1}{2}(h-t)\psi(x,h) + o(h-t). \tag{3.64}$$

From here and the third condition in this example it follows that the first condition given in section 1 is fulfilled for the family (3.60) as well as the equality for the information quantity

$$I(t,F) = \int \psi^2(x,t)F(dx). \tag{3.65}$$

Next we obtain

$$\int \psi^2(x,t)F(dx) = \int \left|\sum_{j=1}^{r} \gamma_j(t)\varphi_j(x,t)\right|^2 F(dx) = \sum_{i,j=1}^{r} \gamma_i(t)\gamma_j(t)b_F^{ij}(t) = I(t,F) \tag{3.66}$$

where $\gamma_1(t),\ldots,\gamma_r(t)$ is the solution of the system of equations (3.63).

The quantity $I(t,F)$ may be interpreted geometrically, provided the "true" distribution $F(x)$ belongs to some parametric family G_h with a known (up to parameter h) density $g(x,h)$ with respect to some σ-finite measure ν on $\mathcal{X}$ and furthermore $G_t(x) = F(x), \Phi(G_h) = h$.

Then relation (3.58) can be re-written in the form

$$\int \varphi_j(x,h)g(x,h)\nu(dx) = 0, \quad i = 1,2,\ldots,r. \tag{3.67}$$

Assume that these identities may be differentiated with respect to h at $h = t$ under the integral sign and let $J(x,t) = \frac{g_t'(x,t)}{g(x,t)}$. Then we obtain from (3.67)

$$-\gamma_i'(t) = \int \varphi_i(x,t)J(x,t)F(dx) = \langle \varphi_i(\cdot,t), J(\cdot,t)\rangle_F \tag{3.68}$$

From here and from (3.63), we obtain that $\sum \gamma_i \varphi_i$ is the projection in $L_2(F)$ of the vector J on the subspace A generated by the vectors $\varphi_1(x,t), \ldots, \varphi_r(x,t)$ and hence

$$I(t,F) = \| \sum \gamma_j \varphi_j \|_F^2 \tag{3.69}$$

is the square of the length of the projection. Evidently $I(t,F) \le \|J\|_F^2 = I_g(t)$, where $I_g(t)$ is the information quantity of the density $g(x,t)$ and the equality is valid if and only if $J(x,t) \in A$. Thus $I(t,F)$ can be somewhat loosely interpreted as a part of Fisher information on the parameter t of density g which is contained in relation (3.67).

If the family (3.60) with functions $\gamma_i(h)$ defined by (3.62) belongs to $\mathbf{F}$ for $|u-t| < \delta$, then $I(t,F) \ge I(F)$ and, by means of Theorem 3.1.2, we obtain the lower bound on the risk of estimators $\Phi(F), F \in \mathbf{F}$. Only slightly more complicated is the argument in the case of unbounded functions $\varphi_i(x,h)$. In this case it is necessary to approximate $\psi(x,h)$ by means of a sequence of bounded functions $\psi_N(x,h)$ converging to $\psi(x,h)$ in $L_2(F)$ and to postulate that the family obtained from (3.60) by replacing ψ with ψ_N belongs to $\mathbf{F}$ for $|u-t| > \delta(N)$.

Example 3.2.6. *Consider the problem of estimating a location parameter in the so-called* two-sample problem *which can be described as follows. Let $\tilde{\mathbf{F}}$ be the set of all distributions on the real line possessing an absolutely continuous density $f(x)$ with respect to the Lebesgue measure and, moreover*

$$I_f = \int \frac{(f'(x))^2}{f(x)} dx < \infty. \tag{3.70}$$

Let a two-dimensional sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a general population with distribution functions $F(x), F(y-t), F \in \tilde{F}$ be given. It is required to estimate t. Thus in this example the set $\mathbf{F}$ is the totality of all two-dimensional distributions of the form $F(x)F(y-t), t \in \mathbb{R}^1, F \in \tilde{\mathbf{F}}$.

In a somewhat different manner, this problem can be described as follows. It is required to estimate the functional $t = \Phi(F,G)$ defined as the unique common root of the equations

$$\int \int [\varphi(x+t) - \varphi(y)] dF(x) dG(y) = 0 \tag{3.71}$$

for an arbitrary bounded $\mathcal{X}$ measurable function φ. In such a formalization this problem may be viewed as an infinite dimensional generalization of the preceding example. Let $J(x) = f'(x)/f(x)$ and consider the parametric family of distributions in $\mathbb{R}^2$ with density

$$f_h(x) f_h(y-t) = [f(x)f(x+h-t)f(y-h)f(y-t)]^{1/2} \left[\int (f(z)f(z+h-t))^{1/2} dz \right]^{-2}. \tag{3.72}$$

Clearly this family belongs to $\mathbf{F}$ and the first two conditions of section 1 are satisfied for this family. Moreover, the condition $I_f < \infty$ assures the validity of the relations

$$\int [f(x)f(x+h-t)]^{1/2} dx = 1 + O((h-t)^2) \tag{3.73}$$

$$\left[\frac{f(x+h-t)}{f(x)} \right]^{1/2} = 1 + \frac{1}{2}(h-t)J(x) + o(h-t) \tag{3.74}$$

as $h \to t$.

This implies that the third natural condition of section 1 and the equalities

$$\psi(x,y) = \frac{1}{2}(J(x) - J(y-t)), \quad I(t,f) = \frac{I_f}{2} \tag{3.75}$$

are valid.

The topology R in this problem may be chosen quite naturally: a sequence of distributions $F_n((x) \times F_n(y-h_n)$ converges to the distribution $F(x)F(y-h)$ provided $I_{f_n} \to I_f$ and $h_n \to h$.

From Theorem 3.1.2, we obtain the following minimax bound on the risks with $w \in \mathbf{W}$ in the two-sample problem:

$$\lim_{N\to\infty} \liminf_{n\to\infty} \inf_{\Phi_n} \sup_{F \in U_N(F_0)} \mathbb{E}_F w(\sqrt{n}(\Phi_n - \Phi(F))) \ge \frac{1}{\sqrt{2\pi}} \int w(x 2^{1/2} I_{f_0}^{-1/2}) e^{-x^2/2} dx. \tag{3.76}$$

The example considered is also of interest due to the fact that in this example when computing $I(F)$ one cannot restrict oneself to linear parametric families of the type (3.22) since linear parametric families are not contained in $\mathbf{F}$.

3.3 Examples of Asymptotically Efficient Estimators

As it is known, for parametric families of distributions there exist general methods of construction of asymptotically efficient estimators in various senses.

Unfortunately, in a nonparametric situation there are as yet no general methods of this kind. For the first five examples considered in section 2 in a corresponding class of distributions $\bar{\mathbf{F}} \subset \mathbf{F}$ uniformly asymptotically efficient in this class estimators were constructed for a wide class of loss functions w. However, the methods of construction differ from one example to another. Their description as well as the proof of asymptotic efficiency would occupy a large amount of space and would presumably be useless for other classes of these problems.

Therefore we shall confine ourselves here to the construction of asymptotically efficient estimators in two simplest classes: Example 3.2.1 and Example 3.2.3.

We start with the investigation of properties of the arithmetic mean, where

$$\Phi(F) = \int \varphi(x) F(dx). \tag{3.77}$$

In the class of distributions $\mathbf{F}_2$ in Example 3.2.1, this estimator is not even uniformly consistent. Therefore it is necessary to consider a "narrower" class of distributions.

Denote by $\mathbf{F}_2^\alpha \subset \mathbf{F}_2$ the class of distributions such that for some chosen function $\alpha = \alpha(N)$ which tends to zero as $N \to \infty$, the inequality

$$\int_{|\varphi|>N} |\varphi(x)|^2 F(dx) \le \alpha(N) \tag{3.78}$$

is fulfilled. In this class the Lindeberg condition for random variables $\varphi(X_1), \dots, \varphi(X_n)$ is obviously fulfilled uniformly for F and thus the weak convergence

$$\mathcal{L}(\sqrt{n}[\hat{\varphi}_n - \Phi(F)]|F) \to \mathcal{N}(0, \sigma^2(F)), \tag{3.79}$$

where $\sigma^2(F) = \int |\varphi(x) - \mathbb{E}_F \varphi(X)|^2 F(dx)$ is also uniform in $F \in \mathbf{F}_2^\alpha$ for any function $\alpha(N) \to 0$.

Furthermore, the equality

$$\mathbb{E}_F \zeta_n^2 = \sigma^2(F) \tag{3.80}$$

is clearly valid for the sequence $\zeta_n = \sqrt{n}(\hat{\varphi}_n - \Phi(F))$. From these relations, the uniform in $F \in \mathbf{F}_2^\alpha$ integrability of the sequence ζ_n^2 follows.

From (3.79) and the uniform integrability, for any function $w \in \mathbf{W}$ such that

$$w(x) \le c(|x|^2 + 1), \tag{3.81}$$

the relation

$$\lim_{n\to\infty} \sup_{F\in\mathbf{F}_2^\alpha} \left| \mathbb{E}_F w(\sqrt{n}(\hat{\varphi}_n - \Phi(F))) - \frac{1}{\sqrt{2\pi}} \int w(x\sigma(F)) e^{-x^2/2} dx \right| = 0 \tag{3.82}$$

follows.

If we introduce in the set $\mathbf{F}_2^\alpha$ any topology R-coordinated with the estimation problem of functional in Example 3.2.1–in which the functional $\sigma^2(F)$ is continuous then, as it follows from the arguments at the end of section 1 and (3.82), $\hat{\varphi}_n$ is $(\mathbf{F}_2, R, w)$-asymptotically efficient uniformly in $\mathbf{F}_2^\alpha$ estimator of the functional for any loss function $w \in \mathbf{W}$ satisfying the condition (3.81) and $I(F) = \sigma^{-2}(F)$.

Note that for faster growing loss functions the estimator $\hat{\varphi}_n$ will not, in general, by asymptotically efficient. However, one can construct a corresponding truncation of the estimator $\hat{\varphi}_n$ which is asymptotically efficient for any $w \in \mathbf{W}$.

We now turn to a study of an estimator of the functional $\Phi(F) = \varphi_0(\int \varphi(x) F(dx))$ in Example 3.1.2. A lower bound in this case was obtained in Theorem 3.2.1 where $\sigma(l, F)$ was computed in Example 3.2.3. We shall now prove that in a corresponding class of distribution functions this bound cannot be improved and that the estimator

$$\varphi_0(\hat{\varphi}_n) \tag{3.83}$$

is asymptotically efficient.

Assume now that function $\varphi_0 : \mathbb{R}^r \to \mathbb{R}^1$ possesses bounded derivatives of the first order in all its arguments satisfying the Lipschitz condition. The set of distributions $\mathbf{F}_2^\alpha$ is defined here as in the first example of this section, but the function $\varphi : \mathscr{X} \to \mathbb{R}^1$ is now replaced by the function $\varphi : \mathscr{X} \to \mathbb{R}^r$.

Let Φ_n be an estimator of the form

$$\Phi_n = \varphi_0(\hat{\varphi}_n). \tag{3.84}$$

Then expanding the function φ_0 in Taylor's formula and taking into account the boundedness and the Lipschitz condition for $\nabla\varphi_0$, for some constant $c > 0$ which is common for all $F \in \mathbf{F}_2^\alpha$ we shall obtain the inequality via the mean value theorem:

$$|\Phi_n - \Phi(F) - \langle\hat{\varphi}_n - \mathbb{E}_F\varphi(X), \nabla\varphi_0(\mathbb{E}_F\varphi(X))\rangle| \le c|\hat{\varphi}_n - \mathbb{E}_F\varphi(X)|^2[1 + |\hat{\varphi}_n - \mathbb{E}_F\varphi(X)|]^{-1}. \tag{3.85}$$

Let

$$\zeta_n = \langle\hat{\varphi}_n - \mathbb{E}_F\varphi(X), \nabla\varphi_0(\mathbb{E}_F\varphi(X))\rangle. \tag{3.86}$$

As in the preceding example we easily obtain that ζ_n is–uniformly in $\mathbf{F}_2^\alpha$–asymptotically normal with parameters $(0, n^{-1}\tilde{\sigma}^2(F))$, where

$$\tilde{\sigma}^2(F) = \int \langle\varphi(y) - \mathbb{E}_F\varphi(X), \nabla\varphi_0(\mathbb{E}_F\varphi(X))\rangle^2 F(dy). \tag{3.87}$$

Moreover, $\mathbb{E}_F\zeta_n^2 = \tilde{\sigma}^2(F)/n$. This implies uniform in $\mathbf{F}_2^\alpha$ integrability of the random variables $n\zeta_n^2$ by Scheffe's lemma. Analogously we verify the uniform integrability of $n|\hat{\varphi} - \mathbb{E}_F\varphi(X)|^2$.

The last assertion and (3.85) allow us to obtain for any function $w \in \mathbf{W}$ satisfying condition (3.81) the relation

$$\lim_{n\to\infty} \sup_{F\in\mathbf{F}_2^\alpha} \left| \mathbb{E}_F w(\sqrt{n}(\Phi_n - \Phi(F))) - \frac{1}{\sqrt{2\pi}} \int w(x\tilde{\sigma}(F)) e^{-x^2/2} dx \right| = 0. \tag{3.88}$$

As above, (3.88) implies uniform in $\mathbf{F}_2^\alpha$ asymptotic efficiency of estimator $\Phi_n = \varphi_0(\hat{\varphi}_n)$ in the corresponding topology for the above indicated class of loss functions as well as the equality $I(F) = \tilde{\sigma}^{-2}(F)$.

In the conclusion of this section we shall describe without proofs asymptotically efficient nonparametric estimators in some other cases.

In Example 3.2.2, U-estimators, i.e., estimators of the form

$$U_n = \frac{1}{c_n^m} \sum_{S_n} \varphi(X_{\alpha_1}, \dots, X_{\alpha_m}), \tag{3.89}$$

where

$$S_n = \{(\alpha_1, \dots, \alpha_m) : 1 \le \alpha_1 < \alpha_2 < \dots < \alpha_m \le n\} \tag{3.90}$$

are asymptotically efficient for loss functions satisfying condition (3.81). For a wider class of loss functions certain truncations of U-estimators will be asymptotically efficient. In Example 3.2.4 and 3.1.2 under quite general assumptions, Huber M-estimators and their truncations will be asymptotically efficient. These are defined as the solution of the equation

$$\int \varphi(x, t) dF_n(x) = 0 \tag{3.91}$$

where F_n is the empirical distribution function. In Example 3.2.5 the information amount under natural restrictions on $\mathbf{F}$ coincides with $I(t, F)$ and asymptotically efficient nonparametric estimators may be constructed recursively. For Example 3.2.6, estimators which are asymptotically normal with parameters $(0, \frac{1}{2}I_f)$ are constructed in a number of papers.

3.4 Estimation of Unknown Density

We have seen that in the case of estimation of, say, a one-dimensional parameter of a distribution in the regular case, there exist usually $\sqrt{n}$-consistent estimators provided the consistent ones exist. In a nonregular case, one can construct estimators which converge even faster. In the nonparametric case the situation is quite different. Here there are many interesting problems for which the nonparametric information quantity introduced vanishes on the whole set of distributions $\mathbf{F}$ under consideration but a consistent estimator (with a slower rate of convergence than $\sqrt{n}$) is nevertheless possible.

Note that in parametric problems the equality $I(\theta) \equiv 0$ on a whole interval implies the nonexistence of consistent estimators of the parameter on this interval, since the density does not depend on the parameter θ.

This type of problem contains the problems of estimating a probability density at some fixed point on the whole real line, derivatives of a density, mode of distribution based on independent observations, spectral density based on observations of a stationary process, and others.

In this section we shall consider only one example of this type of problem–namely, estimation of a probability density at a point based on observations in $\mathbb{R}^1$.

Let $X_1, X_2, \dots, X_n \in \mathbb{R}^1$ be a sample from a population with unknown density $f(x)$. If $f(x)$ depends on a finite number of parameters and is a known function of x and of these parameters, we again arrive at a problem of parametric estimation. If, however, the only thing that is known is that $f(x)$ belongs to a sufficiently large class $\mathbf{F}$ of functions then the problem of estimating $f(x)$ becomes infinitely dimensional, i.e. nonparametric.

We proceed from the empirical distribution function $F_n(x) = \nu_n(x)/n$ where $\nu_n(x)$ is the number of observations smaller than x. $F_n(x)$ is a well known estimator for the distribution function $F(x)$. Setting

$$\chi(x) = I(x \geq 0), \tag{3.92}$$

we have the representation

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \chi(x - X_k). \tag{3.93}$$

As it is known, the function $F_n(x)$ is close to the actual distribution function

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{3.94}$$

provided that n is sufficiently large.

Therefore one would expect that its derivative is close to $f(x) = F'(x)$. However,

$$F_n'(x) = \frac{1}{n}\sum_{k=1}^{n} \delta(x - X_k), \tag{3.95}$$

where $\delta(x)$ is the Dirac δ-function, which is not a function in the sense of classical analysis. It would therefore be natural to "smooth" $F_n(x)$ and use as an estimator of the density the derivative of such a smooth function. We thus arrive at estimators satisfying the condition

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} V\left(\frac{x - X_i}{h_n}\right), \tag{3.96}$$

where $V(x)$ is absolutely integrable and satisfies the condition

$$\int_{-\infty}^{\infty} V(x)dx = 1, \tag{3.97}$$

while the sequence h_n is such that

$$h_n \to 0, \quad nh_n \to \infty. \tag{3.98}$$

The class of estimators (3.96) was first introduced by Rosenblatt and Parzen. They are called the *Parzen–Rosenblatt estimators*. Obviously the convergence $f_n(x) \to f(x)$ in a certain sense is valid only under some restrictions on $f(x)$. If, for example, $f(x)$ possesses points of discontinuity then the convergence is uniform for no choice of h_n and $V(x)$. If it is known beforehand that f belongs to a certain class of continuous functions, then one can find in class (3.96) estimators which converge to $f(x)$ with a given rate.

We shall discuss this point in more detail. Let it be known that $f(x), x \in \mathbb{R}^1$, belongs to the class of functions satisfying the Lipschitz condition with constant L:

$$|f(x_2) - f(x_1)| \leq L(|x_2 - x_1|). \tag{3.99}$$

Denote by $\Sigma(1, L)$ the set of all such functions. Let $f_n(x)$ be an estimator of the form (3.96). We shall bound the quantity

$$D_n(x) = \mathbb{E}_f(f_n(x) - f(x))^2 = \mathbb{E}_f(f_n(x) - \mathbb{E}_f f_n(x))^2 + (\mathbb{E}_f f_n(x) - f(x))^2. \tag{3.100}$$

First we shall consider the bias term. Clearly,

$$\mathbb{E}_f f_n(x) - f(x) = \frac{1}{h_n}\int V\left(\frac{x-y}{h_n}\right)[f(y) - f(x)]dy = \int V(z)[f(x - h_n z) - f(x)]dz. \tag{3.101}$$

If the function $|zV(z)|$ is integrable, then we obtain from the last relation the bound

$$|\mathbb{E}_f f_n(x) - f(x)| \leq Lh_n \int |zV(z)|dz, \tag{3.102}$$

which is valid for $f \in \Sigma(1, L)$. In the same manner,

$$\mathbb{E}_f(f_n(x) - \mathbb{E}_f f_n(x))^2 = \frac{1}{nh_n^2}\left\{\int V^2\left(\frac{x-y}{h_n}\right) f(y)dy - \left[\mathbb{E}_f V\left(\frac{x - X_1}{h_n}\right)\right]^2\right\} \leq \frac{1}{nh_n}\int V^2(z) f(x - h_n z)dz. \tag{3.103}$$

If V^2 is integrable, then for some constant c common for all $f \in \Sigma(1, L)$ the inequality

$$\mathbb{E}_f(f_n(x) - \mathbb{E}_f f_n(x))^2 \leq \frac{c}{nh_n} \tag{3.104}$$

is valid. Evidently the best bound (in order of magnitude) for $D_n(x)$ is obtained if we set $h_n = n^{-1/3}$.

For this choice of h_n, we have $D_n(x) \leq c_1 n^{-2/3}$, where as it is easy to verify, the constant c_1 does not depend on x and $f \in \Sigma(1, L)$. We thus obtain the following result:

If $h_n = n^{-1/3}$ and the functions $|xV(x)|, V^2(x)$ are integrable, then for an estimator $f_n(x)$ of an unknown density $f(x) \in \Sigma(1, L)$ constructed in accordance with (3.96) the inequality

$$\sup_{f \in \Sigma(1,L)} \sup_{x \in \mathbb{R}^1} \mathbb{E}_f(f_n(x) - f(x))^2 \leq cn^{-2/3} \tag{3.105}$$

is valid for all n.

The result can be generalized in various directions. In particular, loss functions which are not quadratic may be considered as well as classes of functions f other than $\Sigma(1, L)$. It is not difficult to verify that if $V(x)$ decreases rapidly as $|x| \to \infty$, then for any integer $k > 0$,

$$\sup_n \sup_{f \in \Sigma(1,L)} \sup_{x \in \mathbb{R}^1} \mathbb{E}_f |(f_n(x) - f(x))n^{1/3}|^{2k} < \infty. \tag{3.106}$$

This fact evidently implies that for any loss function $w(x)$ whose growth is at most polynomial as $|x| \to \infty$ and for any estimator of the form (3.96) with $h_n = n^{-1/3}$ and with a finite, say, function $V(x)$ satisfying $\int_{-\infty}^{\infty} V(x)dx = 1$, the relation

$$\sup_n \sup_{f \in \Sigma(1,L)} \sup_{x \in \mathbb{R}^1} \mathbb{E}_f w((f_n(x) - f(x))n^{1/3}) < \infty \tag{3.107}$$

is valid.

Consider now yet another generalization for other families of functions f. We shall see in particular that for families of f satisfying more stringent smoothness conditions one can find among estimators of the form (3.96) estimators which converge to $f(x)$ even faster and the attainable rate of convergence depends substantially on the smoothness of f.

Denote by $\Sigma(\beta, L), \beta = k + \alpha, 0 < \alpha \leq 1, k \geq 0$ the class of functions possessing k-th order derivatives and such that for $x_i \in \mathbb{R}^1$,

$$|f^{(k)}(x_2) - f^{(k)}(x_1)| \leq L|x_2 - x_1|^\alpha \tag{3.108}$$

and $\Sigma(beta) = \bigcup_{L>0} \Sigma(\beta, L)$. Thus $\Sigma(\beta), \beta = k + \alpha$ is the class of functions with k-th order derivatives satisfying Holder's condition with exponent α.

In order not to specify each time the conditions for convergence of the corresponding integrals, we shall confine ourselves below to the study of procedures of type (3.96) with bounded functions $V(x)$.

Theorem 3.4.1. *For an estimator of the form (3.96) with $h_n = n^{-1/(2\beta+1)}$, $\beta = k + \alpha$ and a bounded function $V(x)$ satisfying condition $\int_{-\infty}^{\infty} V(x)dx = 1$ and conditions*

$$\int_{-\infty}^{\infty} x^j V(x)dx = 0, \quad j = 1, 2, \ldots, k, \tag{3.109}$$

the inequality

$$\sup_n \sup_{f \in \Sigma(\beta,L)} \sup_{x \in \mathbb{R}^1} \mathbb{E}_f[(f_n(x) - f(x))n^{\beta/(2\beta+1)}]^2 < \infty \tag{3.110}$$

is valid for any $L > 0$.

We shall now show that this result can be extended to a substantially wider class of loss functions.

Theorem 3.4.2. *Assume the conditions of Theorem 3.4.1 are fulfilled and let $w(x) \in \mathbf{W}_{e,2}$. Then for any $L > 0$ the relation*

$$\sup_n \sup_{f \in \Sigma(\beta,L)} \sup_{x \in \mathbb{R}^1} \mathbb{E}_f w(n^{\beta/(2\beta+1)}(f_n(x) - f(x))) < \infty \tag{3.111}$$

is valid.

3.5 Minimax Bounds on Estimators for Density

We have shown that in the case when the information $f \in \Sigma(\beta, L)$ is available, there exist estimators for the density which converge to the density with the rate $n^{-\beta/(2\beta+1)}$. Are there, however, even more rapidly convergent estimators? This problem was first considered by Cencov. For one class of measures of deviations of f_n from f–as it was shown by Cencov–the answer is negative if one considers minimax bounds. The result presented below is due to Farrel. We shall not only establish the existence of a minimax bound from below of order $n^{-\beta/(2\beta+1)}$ but also indicate some qualitative bounds.

Denote by $\mathbf{F}_n$ the class of all possible estimators for a density based on observations $X_1, X_2, \ldots, X_n$. Let $w(x)$ be an arbitrary symmetric monotone (for $x > 0$) function such that $w(0) = 0, w(x) \not\equiv 0$. As above, we shall denote the class of these functions by $\mathbf{W}$.

Theorem 3.5.1. *For any $L > 0, x_0 \in \mathbb{R}^1, k \geq 0, \alpha > 0$ and $w \in \mathbf{W}$, the inequality*

$$\liminf_{n\to\infty} \inf_{\tilde{f}_n(x)\in\mathbf{F}_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f w((\tilde{f}_n(x_0) - f(x_0))n^{\beta/(2\beta+1)}) > 0 \tag{3.112}$$

is valid. Here $\beta = k + \alpha$.

Proof. Let $f_0(x)$ be an arbitrary density belonging to $\Sigma(\beta, L/2)$ not vanishing for all $x \in \mathbb{R}^1$ and let $g(x) \in \Sigma(\beta, L/2)$ be finite and satisfy $\int g(x)dx = 0, g(0) \neq 0$.

Elementary verification shows that the function

$$\varphi_n(x, \theta) = f_0(x) + \frac{\theta}{n^{\beta\delta}} g((x - x_0)n^{\delta}\kappa) \tag{3.113}$$

for any $|\theta| < \kappa^{-\beta}, \delta > 0$, belongs to the set $\Sigma(\beta, L)$. Moreover, for all $n \geq n_0$ this function is a probability density.

Consider now a sample $X_1, X_2, \ldots, X_n$ from a population with density $\varphi_n(x, \theta), |\theta| < \kappa^{-\beta}$. Denote by P_θ^n the family of measures generated by this sample. We could show that for $\delta = \frac{1}{2\beta+1}$ the LAN condition is satisfied and we have the Fisher information

$$I_0 = \kappa^{-1} \int g^2(y)dy/f_0(x_0). \tag{3.114}$$

To complete the proof, the following lemma is required, which is a direct consequence of Remark 2.7.3 right after Theorem 2.7.1.

Lemma 3.5.1. *For any estimator T_n of the parameter $|\theta| > \kappa^{-\beta}$ in the parametric family (3.113) and any even and monotone for $x > 0$ loss function w_0, for any $c \leq I_0^{1/2}\kappa^{-\beta}$ the inequality*

$$\liminf_{n\to\infty} \sup_{|u|<c/I_0^{1/2}} \mathbb{E}_u^n w_0((T_n - u)I_0^{1/2}) \geq \frac{1}{\sqrt{2\pi}} 2^{-1} \int_{|y|<c/2} w_0(y)e^{-y^2/2}dy. \tag{3.115}$$

is valid.

Since $\varphi_n \in \Sigma(\beta, L)$, using the notation $\gamma = \frac{\beta}{2\beta+1}$ we obtain for any estimator $\tilde{f}_n(x)$ of the density f and any constant $c \leq I_0^{1/2}\kappa^{-\beta}$ the inequality

$$\sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f w((\tilde{f}_n(x) - f(x_0))n^{\gamma}) \geq \sup_{|\theta|I_0^{1/2}<c} \mathbb{E}_\theta^n w((\tilde{f}_n(x_0) - \varphi_n(x_0, \theta))n^{\gamma}). \tag{3.116}$$

By means of the density estimator $\tilde{f}_n$ one can, in particular, construct an estimator of the parameter θ in the parametric family φ_n using the formula

$$\hat{\theta}_n = [\tilde{f}_n(x_0) - f_0(x_0)]g^{-1}(0)n^{\gamma}. \tag{3.117}$$

The following equality is self-evident:

$$(\hat{\theta}_n - \theta)g(0) = [\tilde{f}_n(x_0) - \varphi_n(x_0, \theta)]n^{\gamma}. \tag{3.118}$$

Setting $w_0(x) = w(xg(0)I_0^{-1/2})$, one obtains the inequality

$$\liminf_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f w((\tilde{f}_n(x_0) - f(x_0))n^{\gamma}) \geq \frac{1}{2\sqrt{2\pi}} \int_{|x|<I_0^{1/2}\kappa^{-\beta}/2} w(xg(0)I_0^{-1/2})e^{-x^2/2}dx, \tag{3.119}$$

which is valid for any estimator $\tilde{f}_n(x_0)$.

Since κ is arbitrarily small the proof is complete. □

Remark 3.5.1. *Note that (3.119) gives a bound from below for the minimax risk if one maximizes the right hand side. It is of interest to obtain the exact achievable bound as it was done for the parametric case. This problem has not been solved yet.*

Remark 3.5.2. *It is not too difficult to strengthen the assertion of Theorem 3.5.1, rendering it to be in a certain sense uniform in x_0. More precisely, the inequality of Theorem 3.4.1 can be replaced by the inequality*

$$\liminf_n \inf_{\tilde{f}_n \in \mathbf{F}_n} \inf_{x_0 \in [a,b]} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_f w((\tilde{f}_n(x) - f(x_0)) n^{\beta/(2\beta+1)}) > 0, \tag{3.120}$$

which is valid for all $-\infty < a < b < \infty$.

Bibliography

[1] I. A. Ibragimov, R. Z. Has' minskii, and S. Kotz, *Statistical estimation: asymptotic theory.* Springer-Verlag New York, 1981, vol. 2.