MATHEMATICS OF GRAVITATION
PART II, GRAVITATIONAL WAVE DETECTION
BANACH CENTER PUBLICATIONS, VOLUME 41
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 1997

THEORY OF PARAMETER ESTIMATION

RYSZARD ZIELIŃSKI
Institute of Mathematics, Polish Academy of Sciences
Śniadeckich 8, P.O. Box 137, 00-950 Warszawa, Poland
E-mail: rziel@impan.gov.pl

0. Introduction and summary. The analysis of data from the gravitational-wave detectors that are currently under construction in several countries will be a challenging problem. The reason is that gravitational-vawe signals are expected to be extremely weak and often very rare. Therefore it will be of great importance to implement optimal statistical methods to extract all possible information about the signals from the noisy data sets. Careful statistical analysis based on correct application of statistical methods will be essential.

The aim of this series of lectures is to introduce the reader to the contemporary theory of parameter estimation. Principles of main estimation methods are reviewed and the properties of the estimators are discussed. The theory of estimation is considered in a general framework of an appropriate statistical model (Sec. 2). Facing a problem of estimation one can start either with a principle (like "take the value of the parameter which is the nearest to your data"), which is developed in Sec. 3.1 ("Heuristic methods") or with some postulated properties of the estimator (Sec. 3.2, "Optimal estimators"). How much can properties of the estimator chosen change under violations of the theoretical model adopted is discussed in Sec. 4, "Robustness".

1. The problem. Let us begin with two examples.

EXAMPLE 1 [see Hollander and Wolfe (1973)]. The following seven observations represent average measurements of θ, the ratio of the mass of the earth to that of the moon, obtained from seven different spacecrafts.

Mariner 2	81.3001
Mariner 4	81.3015

1991 *Mathematics Subject Classification*: Primary 62F10; Secondary 62-01.
Research supported by KBN grant 2 P303D 021 11.
The paper is in final form and no version of it will be published elsewhere.

Mariner 5	81.3006
Mariner 6	81.3011
Mariner 7	81.2997
Pioneer 6	81.3005
Pioneer 7	81.3021

On the basis of previous Ranger spacecraft findings, the value of θ had been estimated as approximately equal to 81.3035. What is the "true" value of the ratio of the mass of the earth to that of the moon?

EXAMPLE 2 [David and Pearson (1961)]. The following are figures for the length (mm.) of cuckoo's eggs which were found in nests belonging to the hedge-sparrow, the reed-warbler and the wren:

Hedge-sparrow	22.0 23.9 20.9 23.8 25.0 24.0 21.7 23.8 22.8 23.1 23.1 23.5 23.0 23.0	($mean = 23.11$)
Reed-warbler	23.2 22.0 22.2 21.2 21.6 21.6 22.0 22.9 22.8	($mean = 22.14$)
Wren	19.8 22.1 21.5 20.9 22.0 21.0 22.3 21.0 20.3 20.9 22.0 20.0 20.8 21.2 21.0	($mean = 21.12$)

The problem is that the size of the cuckoo's eggs seems to be associated with the size of the nest in which it is laid; it is known that the hedge-sparrow has a larger nest than the reed-warbler and the reed-warbler than the wren. What are the sizes of the cuckoo's eggs laid in the hedge-sparrow, reed-warbler, and wren nests?

2. Statistical model. There are many (perhaps infinitely many) statistical models which can in a more or less adequate way describe a given problem. Let us try the following.

EXAMPLE 1 (cont.). We assume, at least for our purpose now, that the mass of the earth and the mass of the moon are fixed and we denote by θ their ratio. When measuring θ, we obtain a result, say X, which as we believe differs from θ by a (small) random error due to measurement devices and techniques, say ε. In consequence, X is considered, e.g., as a random variable of the form

$$X = \theta + \varepsilon. \tag{1}$$

To conclude "something" about θ having X at our disposal, we must specify the nature of ε. For example, we may believe that the random error ε is (a) as likely positive as negative, or (b) its expectation is equal to zero ($E\varepsilon = 0$), etc.

In case (a) we have the following statistical model of the result X of observation: X takes on its values in R^1 (the set of reals); the probability distribution P, defined on a σ-field $\mathcal{B}^1$ of subsets of R^1, is not completely defined but we assume that it is from a family $\mathcal{P}$ of distributions on the measurable space $(R^1, \mathcal{B}^1)$ such that if $P \in \mathcal{P}$, then $P\{X \leq \theta\} \geq 1/2$ and $P\{X \geq \theta\} \geq 1/2$.

The triplet $(R^1, \mathcal{B}^1, \mathcal{P})$ is considered as the statistical model of the observation X. The unknown θ is a median of the unknown probability distribution of the observation X. Actually, θ can be viewed as a mapping $\theta : \mathcal{P} \to R^1$ such that $\theta(P)$ is a median of P.

Any mapping $\hat{\theta}$ from R^1, the space of observations, to R^1, the space of values of θ such that $\hat{\theta}(X)$ is treated as a "guessed value of θ" is called an *estimator* of θ.

The case (b) is similar: the difference is that "median" should be replaced be "expectation". A comment is however needed: $\mathcal{P}$ in the statistical model $(R^1, \mathcal{B}^1, \mathcal{P})$ is now a family of distributions on the measurable space $(R^1, \mathcal{B}^1)$, for which the expectation exists.

Sometimes the family $\mathcal{P}$ can be specified in a more detailed way, for example as a family of Gaussian distributions, distributions with continuous and strictly increasing cumulative distribution function, etc.

EXAMPLE 2 (cont.). A possible statistical model for the results of measurements presented in Example 2 is as follows. Define θ_1 as a "typical length of cuckoo's eggs laid in the hedge-sparrrow nest". Similarly define θ_2 and θ_3 for reed-warbler and wren, respectively. "Typical length" means that the actual result of observation of the size of the cuckoo's egg laid in the hedge-sparrow nest is of the form $\theta_1 + \varepsilon$, where the random variable ε is modelling the random variability between eggs as well as measurement errors. Assume that $E\varepsilon = 0$. The parameter to be estimated is $(\theta_1, \theta_2, \theta_3)$. (In the original statement of the problem the main question to answer was if $\theta_1 > \theta_2 > \theta_3$ as suggested by observations, but we are not going to follow this direction of testing statistical hypotheses.)

In general, an observation X is considered as a random element: a real valued random variable, a random vector, a random function, etc. The space of values of X, to be denoted by $\mathcal{X}$, is called the sample space. Building a statistical model for a specified observation X consists in specifying a sample space $\mathcal{X}$, a σ-field of observable events $\mathcal{F}$ and a family $\mathcal{P}$ of distribution functions. The triplet $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ is defined to be a *statistical model*.

Now, led $\mathcal{D}$ be a set and let $\theta : \mathcal{P} \to \mathcal{D}$ be a mapping. The value $\theta(P)$, $P \in \mathcal{P}$, will be considered as a parameter of P, and $\mathcal{D}$ as the parameter space. Any (measurable) mapping $\hat{\theta} : \mathcal{X} \to \mathcal{D}$ such that $\hat{\theta}(X)$ is viewed as a "guessed value" of $\theta(P)$ is called an *estimator* of the parameter θ.

If $X = (X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n$ are independent identically distributed (i.i.d.) random variables (r.v.'s) according to a distribution P, then the terms *a sample* or *a random sample* from the *population* P or from the *parent population* P or from the distribution P, are used.

3. Estimation. Given a statistical model $(\mathcal{X}, \mathcal{F}, \mathcal{P})$, a parameter $\theta : \mathcal{P} \to \mathcal{D}$ with the parameter space $\mathcal{D}$, the estimation problem consists in constructing an $\mathcal{F}$-measurable function $\hat{\theta} : \mathcal{X} \to \mathcal{D}$ such that, if X has a distribution $P \in \mathcal{P}$, $\hat{\theta}(X)$ could be reasonably considered as an "estimator" of $\theta(P)$ for all $P \in \mathcal{P}$.

At this level of discussion the statement of the problem is rather hazy: $\hat{\theta}(X)$ *could be reasonably considered as "an estimator" of* $\theta(P)$ is not precise but we feel that it is meant that $\hat{\theta}(X)$ should be close to $\theta(P)$. In the long history of statistical inference the fuzzy concept of such "closeness" was formalized in many different ways which led to the present situation where we have at our disposal many different estimators even in very simple statistical models.

Two main streams of developing the matter are heuristic approaches and theories of optimal estimators.

3.1. *Heuristic approaches*

3.1.1. *Minimum distance estimators.* Suppose that $\mathcal{D} \subset \mathcal{X}$. Let ρ be a metric in $\mathcal{X}$. The minimum distance estimator $\hat{\theta}_{MDE}$ of θ is defined as

$$\hat{\theta}_{MDE}(X) = \operatorname{argmin}_{\theta \in \mathcal{D}} \rho(X, \theta).$$

In typical statistical models $\hat{\theta}_{MDE}$ exists and is unique.

This simple, "natural", and intuitively appealing idea has a long history. For the case where $\mathcal{X} = R^n$, Gauss (1821) [see Lehmann (1986)] suggested taking for ρ the Euclidean metric in R^n, which led to the celebrated Least Square Estimators (LSE). Laplace (1820) [see Lehmann (1986)] proposed the L_1 norm which has given us Minimum Absolute Deviation (MAD) estimators. In the statistical model of repeated measuring of an unknown quantity θ,

$$X_i = \theta + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

$LSE(\theta)$ is equal to the arithmetic mean $\sum_{i=1}^{n} X_i/n$ of observations, and $MAD(\theta)$ is equal to a median of X_1,$X_2, \ldots, X_n$ (take the permutation $X_{1:n}$,$X_{2:n}, \ldots, X_{n:n}$ of X_1, $X_2, \ldots, X_n$ such that $X_{1:n} \leq X_{2:n} \leq \ldots X_{n:n}$; if n is odd, the median of X_1,$X_2, \ldots, X_n$ is $X_{(n+1)/2:n}$; otherwise the median is any value from the interval $\left(X_{n/2:n}, X_{n/2+1:n}\right)$).

Example 3 (nonlinear regression). In signal identification problem, when a signal of the form

$$X(t) = A\cos(\omega t) + noise, \quad t > 0 \tag{3}$$

with unknown (A, ω) is observed at $t_1, t_2, \ldots, t_n$, the LSE of (A, ω) is that of minimizing

$$\sum_{i=1}^{n} \left(X(t_i) - A\cos(\omega t_i)\right)^2. \tag{4}$$

Example 4 (linear models). Suppose that the observations X are of the form

$$X(t) = \theta^T t + noise \tag{5}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is an unknown (k-dimensional) parameter, $t = (t_1, t_2, \ldots, t_k)$ and t_1,$t_2, \ldots, t_k$ are factors ("independent variables") influencing the observation, quantitative (e.g. t in Example 3) or qualitative (e.g. t_1 = hedge-sparrow, t_2 = reed-warbler, and t_3 = wren in Example 2). All "vectors" are treated here as one-column matrices and $(\cdot)^T$ denotes transposition. Suppose that the observation is repeated, say n times, each time at different levels of factors. As a result we obtain the observations

$$X_i = \theta_1 t_{i,1} + \theta_2 t_{i,2} + \ldots + \theta_i t_{i,k} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

If $n \geq k$ and the rang of the matrix $\mathbf{D} = (t_{i,j})_{i=1,2,\ldots,n;j=1,2,\ldots,k}$, known as the matrix of experimental design, or simply *design matrix*, is equal to k, then $LSE(\theta)$ is $\hat{\theta} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}X$, where $X^T = (X_1, X_2, \ldots, X_k)$. Examples 1 and 2 are examples of linear models.

3.1.2. *Maximum likelihood estimators*

EXAMPLE 5. Let X be the number of occurrences of a specified random event A in n independent experiments ("Bernoulli scheme"). If the probability of A occurring in each separate trial is equal to (an unknown) θ, then we are dealing with the statistical model with the set $\mathcal{X} = \{0, 1, \ldots, n\}$ as the sample space, $2^{\mathcal{X}}$ (the set of all subsets of $\mathcal{X}$) as the (σ-)field of observable events, and the family of binomial distributions

$$P_\theta\{X = x\} = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n; \quad \theta \in [0, 1] \tag{7}$$

as the family of all possible distributions. The problem is to estimate θ on the basis of the observation X.

Observe that $P_\theta\{X = x\}$ is considered here as a function of x under a fixed value of the parameter θ. Write

$$lik(\theta; x) = P_\theta\{X = x\} \tag{8}$$

and, for a given observation $X = x$, consider $lik(\theta; x)$ as a function on the parameter space $[0, 1]$. Were $\theta = \theta_1$ for some fixed θ_1, the probability of getting the result $X = x$ would be equal to $lik(\theta_1; x)$; were $\theta = \theta_2$ for some fixed θ_2, the probability of getting the result $X = x$ would be equal to $lik(\theta_2; x)$. Then, if $lik(\theta_1; x) > lik(\theta_2; x)$, "it is more likely that our observation $X = x$ has been obtained under θ_1 than θ_2".

Function (8) is called the likelihood function (more precisely: $lik(\theta; x)$ is the likelihood of θ when the result of observation was $X = x$). The value of $\theta = \theta(X)$ that maximizes (8) is called the Maximum Likelihood Estimator (MLE) of the parameter θ.

In general: considering a statistical model $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ we assume that there exists a σ-finite measure on the measurable space $(\mathcal{X}, \mathcal{F})$ (for example the Lebesgue measure) such that every distribution $P \in \mathcal{P}$ has a density (Radon-Nikodym derivative) with respect to that measure and the density is of the form $p(x; \theta)$. In the statistical model under consideration "everything" except θ is assumed to be known and the problem is to estimate θ. The likelihood function is defined as

$$lik(\theta; x) = p(x; \theta) \tag{9}$$

where x is the known result of observation. $MLE(\theta)$ is the θ which maximizes $lik(\theta; x)$.

EXAMPLE 6. If $X = (X_1, X_2, \ldots, X_n)$ and $X_1, X_2, \ldots, X_n$ are i.i.d. r.v.'s distributed according to a normal distribution with density function $(\sigma\sqrt{2\pi})^{-1}\exp\{-(x-\mu)^2/2\sigma^2\}$, where μ and σ are unknown parameters, then

$$lik(\mu, \sigma; x) = (\sigma\sqrt{2\pi})^{-n}\exp\{-\sum_{i=1}^{n}(x_i - \mu)^2/2\sigma^2\} \tag{10}$$

and $MLE(\mu, \sigma)$ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2. \tag{11}$$

It may happen that MLE does not exist or that MLE is not unique. Another kind of difficulty may arise when one tries to find the global maximum of the likelihood function.

3.1.3. *Statistical functionals.* Let $X = (X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n$ are i.i.d. r.v.'s distributed according to a distribution P with cumulative distribution function F:

$$F(x) = P\{X \le x\}. \tag{12}$$

In many common statistical models the parameter θ of interest can be represented in the form

$$\theta = \int h(x)dF(x). \tag{13}$$

Here and below the integration is meant over the whole real line. Let $F_n(x)$ be the empirical distribution function:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i). \tag{14}$$

It seems to be "natural" to take

$$\hat{\theta} = \int h(x)dF_n(x) \tag{15}$$

as an estimator of θ.

The approach was initiated by K. Pearson approximately one hundred years ago, in its original formulation for $h(x)$ of the form x^m, $m = 1, 2, \ldots$ That was known as the method of moments. For example, for the expected value θ we have $h(x) = x$ and the estimator is

$$\int x dF_n(x) = \sum_{i=1}^{n} X_i/n \tag{16}$$

which is the sample mean. Similarly the sample variance is an estimator of the variance, etc.

The following version of the method of moments has been developed. Suppose that the distribution P depends on a real ("one-dimensional") parameter θ. Then $E_\theta X$ is a function of θ. The expectation $E_\theta X$ is estimated by the sample mean. Then an estimator for θ may be obtained as a solution (with respect to θ) of the equation

$$E_\theta(X) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{17}$$

A somewhat more general case is presented in the following example.

EXAMPLE 7. If X has gamma distribution with the density

$$f_{\alpha,\lambda}(x) = \frac{1}{\lambda^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\lambda}, \quad x \ge 0, \alpha \ge 0, \lambda > 0, \tag{18}$$

then the expected value and the variance of X are

$$E_{\alpha,\lambda}X = \alpha\lambda \quad \text{and} \quad Var_{\alpha,\lambda}X = \alpha\lambda^2. \tag{19}$$

Substituting to (19) the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ and the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ for the expectation $E_{\alpha,\lambda}X$ and the variance $Var_{\alpha,\lambda}X$, respectively, and

solving the resulting equations we obtain

$$\hat{\lambda} = \frac{S^2}{\bar{X}} \quad \text{and} \quad \hat{\alpha} = \frac{\bar{X}^2}{S^2} \tag{20}$$

as estimators of the parameters λ and α, respectively.

A warning is needed: $\int h(x)dF_n(x)$ always exists though $\int h(x)dF(x)$ may not be defined.

3.2. *Optimal estimators*

3.2.1. *Statement of the problem.* Let $(\mathcal{X}, \mathcal{F}, \mathcal{P})$ be a given statistical model with a parameter θ and the parameter space $\mathcal{D} = \theta(\mathcal{P})$. Given an observation X, we are looking for an estimator $\hat{\theta} : \mathcal{X} \to \mathcal{D}$.

From now on we shall use a notation which is more suitable for what we are going to speak about. A statistical model will be denoted by $(\mathcal{X}, \mathcal{F}, \{P_\theta : \theta \in \Theta\})$, where $\mathcal{X}$ is a sample space, $\mathcal{F}$ a σ-field of observable random events, and $\{P_\theta : \theta \in \Theta\}$ is a family of distributions on the measurable space $(\mathcal{X}, \mathcal{F})$ indexed by a parameter $\theta \in \Theta$. We assume that the parameter θ is identifiable, which means that $\theta_1 = \theta_2$ iff $P_{\theta_1} = P_{\theta_2}$. To clearly present the main ideas without too many technicalities, we shall consider the problem of estimating the value $g(\theta)$, where $g : \Theta \to R^1$ is a given real-valued function. An estimator of g will be traditionally denoted by $\hat{g}$.

To reasonably speak about optimal estimators we have to define an order (at least a partial order) in the space of estimators; this of course amounts to defining an optimality criterion. We shall do this as follows.

Let $L : \Theta \times R^1 \to R^1_+$ be a given function with the following interpretation: if X comes from a distribution P_θ and the estimator takes on the value $\hat{g}(X)$, then our "loss" is equal to $L(\theta, \hat{g}(X))$. Under such interpretation it is natural to restrict ourselves to functions L satisfying the following conditions: 1) $L(\theta, g(\theta')) = 0$ iff $g(\theta) = g(\theta')$, and 2) $L \geq 0$. Roughly speaking, the optimal estimator is an estimator which minimizes losses.

A new concept in our considerations has emerged: the loss function. The "loss" may be considered as a distance between the value of the estimator and what is to be estimated. Or, in a more general statement of the problem, one can imagine that as a result of estimation we take an "action", say $a(X)$, from a space $\mathcal{A}$ of actions, and define L as a function on $\Theta \times \mathcal{A}$. This is an approach developed in the statistical decision theory (see e.g. Ferguson (1967), Lehmann (1986)). We confine ourselves to the case $\mathcal{A} = R^1$.

The loss function, as it is, is not suitable for ordering estimators first of all due to the fact that it is a random variable. This is easy to overcome: take the expected value of L and consider the quantity

$$R(\theta, \hat{g}) = E_\theta L(\theta, \hat{g}(X)) = \int L(\theta, \hat{g}(x))P_\theta(dx) \tag{21}$$

For a fixed estimator $\hat{g}$, the function $R(\cdot, \hat{g}) : \Theta \to R^1$ is called the risk function of the estimator $\hat{g}$. This gives us a partial ordering in the space of possible estimators: an estimator $\hat{g}_1$ is said to be as good as an estimator $\hat{g}_2$ if

$$R(\theta, \hat{g}_1) \leq R(\theta, \hat{g}_2) \quad \text{for all } \theta \in \Theta; \tag{22}$$

$\hat{g}_1$ is said to be better than $\hat{g}_2$ if it is as good as $\hat{g}_2$ and

$$R(\theta, \hat{g}_1) < R(\theta, \hat{g}_2) \quad \text{for some } \theta \in \Theta. \tag{23}$$

Now it is clear what an optimal estimator (the best estimator) is. A somewhat weaker concept is that of admissibility: an estimator $\hat{g}$ is said to be admissible if there exists no estimator better than $\hat{g}$.

Suppose we are interested in the best (under the above partial ordering) estimator in a class $\mathcal{G}$ of estimators under a fixed loss function L. It appears that if $\mathcal{G}$ is too large, then the optimal estimator does not exist. Suppose, for example, that $\mathcal{G}$ contains constants. Take $g_0 = g(\theta_0)$ as an estimator. Then $R(\theta_0, g_0) = 0$. It follows that for the optimal estimator $\hat{g}_{opt}$ one should have $R(\theta, g_{opt}) = 0$ for all $\theta \in \Theta$, which is impossible by the very definition of L.

There are two ways to deal with the problem. The first possibility is to restrict the class of estimators under consideration. For example it seems to make sense to eliminate constants as estimators. This we shall discuss in Sec. 3.2.2 and 3.2.3. Another way is to define a linear ordering (not partial ordering only) in the set of estimators. The problem is discussed in Sec. 3.2.4. The basic concept in developing the theory of optimal estimators is that of sufficiency.

3.2.2. *Sufficient statistics*

EXAMPLE 8. Let $X_1, X_2, \ldots, X_n$ be a sample from the two-point distribution

$$P_\theta\{X_1 = 1\} = \theta = 1 - P_\theta\{X_1 = 0\}. \tag{26}$$

Then

$$P_\theta\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \theta^s(1-\theta)^{n-s} \tag{27}$$

where $s = x_1 + x_2 + \ldots + x_n$. Define a new random variable: $S = X_1 + X_2 + \ldots + X_n$. The distribution of S is given by the formula

$$P_\theta\{S = s\} = \binom{n}{s}\theta^s(1-\theta)^{n-s}, \quad s = 0, 1, \ldots, n. \tag{28}$$

The conditional distribution of the sample $(X_1, X_2, \ldots, X_n)$ if $S = s$ is

$$P_\theta\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, S = s\} = \begin{cases} 1/\binom{n}{s} & \text{if } \sum_{i=1}^n x_i = s, \\ 0 & \text{otherwise.} \end{cases}$$

The point is that the conditional distribution of the sample, under the condition that $S = s$, does not depend on θ. An obvious interpretation is: if we want to conclude something about the unknown value of the parameter θ and we know S, then all other information about the sample is irrelevant for the conclusion. The statistic S is *sufficient*.

In general: in a given statistical model $(\mathcal{X}, \mathcal{F}, \{P_\theta : \theta \in \Theta\})$ of the observation X, a statistic $T = T(X)$ is said to be *sufficient* if the conditional distribution of X given $T = t$ is independent of θ for all t.

Typically there are many sufficient statistics in a given statistical model. A statistic S is said to be a *minimal sufficient statistic* if for every sufficient statistic T there exists a function f such that $S = f(T)$.

The significance of sufficiency follows from a theorem which states that if T is a sufficient statistic then for every estimator $\hat{g}(X)$ there exists an estimator $\hat{\gamma}(T)$ based on T which is as good as $\hat{g}(X)$.

3.2.3. *Minimum variance unbiased estimators.* Suppose that we are interested in estimating $g(\theta)$ for a fixed function $g : \Theta \to R^1$. Take $L(\theta, \hat{g}(X)) = (\hat{g}(X) - g(\theta))^2$ as the loss. Then the risk of the estimator $\hat{g}$

$$R(\theta, \hat{g}) = \int (\hat{g}(x) - g(\theta))^2 P_\theta(dx)$$

is called the *mean square error* (MSE) of the estimator. As above, in the class of all estimators, there exists no estimator uniformly minimizing the risk.

A possible criterion for restriction of the class of estimators to be considered is based on the concept of the *bias*.

Given a loss function L, an estimator $\hat{g}(X)$ is said to be *an unbiased estimator* of $g(\theta)$ if for each $\theta \in \Theta$

$$E_\theta L(\theta, \hat{g}(X)) \leq E_\theta L(\theta', \hat{g}(X)) \quad \text{for all } \theta' \in \Theta. \tag{24}$$

An equivalent (except some rather special situations) condition in the case of the MSE is

$$E_\theta \hat{g}(X) = g(\theta) \quad \text{for all } \theta \in \Theta. \tag{25}$$

When considering the loss as a measure of a distance between the estimator and what is to be estimated, condition (24) states that an unbiased estimator is that which in mean is closer to the true value of the parameter than to any other of its values. In the case of MSE as risk, this amounts to that the expected value of an unbiased estimator is equal to what is to be estimated.

It could happen that the class of unbiased estimators is empty. Sometimes it contains exactly one element (it is perhaps a good place to say that in statistics, or more generally in probability theory, *exactly one* means that if there exist other elements of a given property, then they differ from the given element on a set whose probability is equal to zero). Typically, however, especially in statistical models with repeated observations, it is not the case.

Observe that if $\hat{g}$ is an unbiased estimator then its MSE is equal to its variance. Hence "the minimum variance unbiased estimator" is an unbiased estimator with (uniformly) minimal risk function. In standard statistical models, under some additional technical conditions ("completeness") such estimator can be explicitly constructed; it is an estimator based on the minimal sufficient statistic. A beautiful presentation of the theory and an abundance of examples can be found in Lehmann (1986).

3.2.4. *Best equivariant estimators.* Suppose we are interested in estimating the mean $\theta = EX$ of a random variable X. Let $X_1, X_2, \ldots, X_n$ be a sample from the underlying distribution and let $T(X_1, X_2, \ldots, X_n)$ be an estimator. Now consider a new random variable $X + c$, where $c \in R^1$ is a constant. It is obvious that $E(X + c) = EX + c$ so that it is natural to state the condition for the estimator

$$T(X_1 + x, X_2 + c, \ldots, X_n + c) = T(X_1, X_2, \ldots, X_n) + c \quad (\forall c \in R^1). \tag{29}$$

If $L(\theta, T)$ is the loss when estimating θ by T, then is natural to expect that

$$L(\theta + c, T(X) + c) = L(\theta, T) \quad (\forall c \in R^1). \tag{30}$$

In general, let $(\mathcal{X}, \mathcal{F}, \{P_\theta : \theta \in \Theta\})$ be a statistical model of an observation X under consideration and consider the problem of estimating $h(\theta)$ for a given transformation $h : \Theta \to \mathcal{H}$. Let $\mathcal{G}$ be a group of $1:1$ transformations g of $\mathcal{X}$ onto itself. Let gX be the random variable which takes on the value gx when $X = x$ and let $P_{\theta'}$ be the distribution of gX if P_θ is the distribution of X. We assume that $\theta' \in \Theta$ so that the transformation g of $\mathcal{X}$ onto itself generates a transformation $\theta' = \bar{g}\theta$ of Θ into itself. We assume that $\bar{g}$ is a transformation of Θ onto itself and that $\bar{\mathcal{G}} = \{\bar{g} : g \in \mathcal{G}\}$ is a group of transformations. If we are interested in estimating $h(\theta)$ then a natural additional requirement is that for every fixed $\bar{g}$

$$h(\bar{g}\theta_1) = h(\bar{g}\theta_2) \quad \text{whenever} \quad h(\theta_1) = h(\theta_2). \tag{31}$$

This induces a transformation $\tilde{g} : \mathcal{H} \to \mathcal{H}$ such that

$$h(\bar{g}\theta) = \tilde{g}h(\theta) \quad (\forall \theta \in \Theta). \tag{32}$$

Now it is obvious that if $T(X)$ estimates $h(\theta)$ then $T(gX)$ should estimate $\tilde{h}(\theta)$. An estimator T is said to be *equivariant* if

$$T(gX) = \tilde{g}T(X) \quad (\forall g \in \mathcal{G}). \tag{33}$$

It appears that if T is an equivariant estimator then its risk function satisfies

$$R(\bar{g}\theta, T) = R(\theta, T) \quad (\forall \theta \in \Theta) \tag{34}$$

and if $\bar{\mathcal{G}}$ is transitive, then the risk function is constant. If this is the case, the best equivariant estimator is obtained by minimizing that constant.

Though not always the problem may be so dramatically simplified, in many situations typical for applied statistics the best invariant estimators can be explicitly constructed.

3.2.5. *Bayes and minimax estimators.* Ordering estimators $T \in \mathcal{T}$ by their risk functions $R(\theta, T)$ causes some problems generated by the fact that the risk function enables one to define a partial ordering only. Under some restrictions on $\mathcal{T}$ (unbiasedness, invariance) it is sometimes possible to effectively construct uniformly minimum risk estimators in a given class as we saw in Sec. 3.2.3 and 3.2.4. Other approaches consist in defining (linear) orders in the set $\mathcal{T}$ of estimators under consideration: the best estimator is then clearly defined though its explicit construction may be difficult.

Consider a statistical model $(\mathcal{X}, \mathcal{F}, \{P_\theta : \theta \in \Theta\})$ of an observation X. Let $R : \Theta \times \mathcal{T} \to R^1$ be a fixed risk function.

The Bayes approach consists in introducing a probability measure, say π, on the measurable space $(\Theta, \mathcal{F}_\Theta)$, where $\mathcal{F}_\Theta$ is a σ-field of subsets of Θ, averaging the risk of a given estimator $T \in \mathcal{T}$ with respect to that probability:

$$r(T) = \int R(\theta, T)\pi(d\theta) \tag{35}$$

and finding T^* such that

$$r(T^*) \leq r(T) \quad (\forall T \in \mathcal{T}) \tag{36}$$

The distribution π is called *a prior distribution*, T^* is called the *Bayes estimator* (under given loss function and prior distribution), and $r(T)$ is called the *Bayesian risk* of the estimator T.

The Bayes approach raises some controversies. The *prior* distribution is sometimes understood as a description of the state of mind of the statistician facing an estimation problem: he believes that *a priori*, i.e. before any observation is performed, some values of the unknown parameter θ are more plausible than others and he expresses this belief by putting more probability mass π to those more plausible values. Sometimes the prior distribution π summarizes all past experience of the statistician. The interpretation plays a crucial role e.g. in statistical quality control and reliability theory as well as in actuarial sciences. The Bayes approach may be also considered as a pure mathematical tool like scalarization in multi-criterion optimization problems. The fact is that except some rather special situations the Bayes estimators are admissible.

If P_θ is the distribution of the observation X, $lik(\theta; x)$ is the likelihood function of θ when $X = x$ was observed, and π is a *prior* distribution of the unknown parameter θ, then

$$\pi_x(\theta \in A) = \int_A lik(\theta; x)\pi(d\theta) \tag{37}$$

defines the *posterior* distribution of θ. In many practical problems, construction of the Bayes estimator is really very simple. For example, under quadratic loss function (MSE estimators!) the Bayes estimator of a parameter is the expected value of the parameter with respect to the posterior distribution, and under the absolute deviation loss (MAD estimators!) it is a median of the posterior distribution. The Bayes estimator may be viewed as a result of mixing our prior knowledge about the unknown parameter θ with what we learned from the observation of X; it is of no wonder that the influence of the former is diminishing when the number of our observations is growing.

The *minimax* approach is as follows. Given a class $\mathcal{T}$ of estimators and a risk function $R : \Theta \times \mathcal{T} \in R^1$, consider the "worst" result for a fixed estimator T: $\sup_{\theta\in\Theta} R(\theta, T)$. The estimator T^* which minimizes that quantity is said to be *minimax*:

$$\sup_{\theta\in\Theta} R(\theta, T^*) \le \sup_{\theta\in\Theta} R(\theta, T) \quad (\forall T \in \mathcal{T}). \tag{37}$$

The problem is that explicit construction of a minimax estimator in most problems is not easy and in fact each problem has to be treated separately.

4. Robustness. A statistical model adopted for a given real-life problem of estimation (see Ex. 1 and Ex. 2) should be 1) adequate and 2) mathematically tractable. Typically the latter prevails: otherwise we might not be able to construct an estimator. The theory of robustness enables one to work in nice and mathematically tractable models, and to take into account some inadequacies of the given theoretical model. For example we may assume in Examples 1 and 2 that the noise is Gaussian but we should ask how our estimator would behave if the noise is "not exactly Gaussian". Or, in theoretical considerations we would like to assume that repeated observations are independent random variables but we should answer the question what happens to the estimator constructed

under such assumptions if the random errors are not independent (for example, when performing a new measurement the researcher may keep in mind the results of previous measurements).

There are many different approaches to the problem. Perhaps the most general one is presented in Zieliński (1983).

Let $(\mathcal{X}, \mathcal{F}, \mathcal{P}_0 = \{P_\theta : \theta \in \Theta\})$ be a statistical model under consideration. Let $\mathcal{P}$ be the family of all probability distributions on $(\mathcal{X}, \mathcal{F})$, so that $\mathcal{P}_0 \subset \mathcal{P}$. The nonadequacy (or violation) of the model may be described by a mapping $\beta : \mathcal{P}_0 \to 2^{\mathcal{P}}$ with the following interpretation: instead of a given model distribution $P_\theta \in \mathcal{P}_0$, the observation has an unknown distribution from the set $\beta(P_\theta)$. Suppose that we have constructed an estimator T and we are interested in its property ρ, e.g. its bias, variance, risk, etc. The property may be considered as a mapping from $\bigcup_{\theta\in\Theta} \beta(P_\theta)$ into a metric space (the real line, R^m, or a more general space). Fix $\theta \in \Theta$. If the "true" distribution P of the observation X runs over $\beta(P_\theta)$, then $\rho(T, P)$ runs over some set and the diameter of that set, say $r(\theta, T)$, is a measure of stability of the estimator T under violation of the original model at the point $\theta \in \Theta$. For a given estimator T, the function $r(\cdot, T) : \Theta \to R^1$ characterizes stability of T in the model $(\mathcal{X}, \mathcal{F}, \mathcal{P}_0 = \{P_\theta : \theta \in \Theta\})$ under the violation β. Estimator T_1 is more stable (more robust) than T_2 if $r(\theta, T_1) \leq r(\theta, T_2)$ for all $\theta \in \Theta$ and $r(\theta, T_1) < r(\theta, T_2)$ for a $\theta \in \Theta$.

For some models and their violations, the uniformly most robust estimators have been effectively constructed.

References

F. N. David, E. S. Pearson, (1961), *Elementary Statistical Exercises*, Cambridge University Press.

T. S. Ferguson, (1967), *Mathematical Statistics. A Decision Theoretic Approach*, Academic Press.

M. Hollander, D. A. Wolfe, (1973), *Nonparametric Statistics Methods*, Wiley.

E. L. Lehmann, (1986), *Theory of Point Estimation*, Wiley.

R. Zieliński, (1983), Robust Statistical Procedures: a General Approach. In: *Stability Problems for Stochastic Models*, Lecture Notes in Mathematics 982, Springer Verlag.