Chapter 2

The Accepted Bayes Method of Estimation

2.1 The components of the Bayes approach

We give a brief description of the structure of Bayes estimates with the help of formal mathematical constructions of the Bayes theory. The Bayes scheme of the decision theory includes the following four components.

1) The statistical model, represented by the probability space $(\Omega, \mathscr{L}, P)$. Here Ω is the set of all possible data in some design of experiments Π, $\Omega = \{x\}$. The data x appear to be the data of a random experiment, thus on Ω it is determined some σ-algebra $\mathscr{L}$ of random events; $P \in \mathscr{B}$, where $\mathscr{B}$ is the family of probability measures on Ω, $\mathscr{L}$. In the traditional Bayes approach, the probability measure P is defined by the representation of some parameter θ (vector or scalar), that is, $\mathscr{B} = \{P_\theta\,;\, \theta \in \Theta\}$ is a parameterized family of probability measures.

2) The probability space $(\Theta, \mathscr{E}, H)$ for the parameter θ which is assumed to be random. Here $\mathscr{E}$ is σ-algebra on Θ, H is a probability measure on $(\Theta, \mathscr{E})$. The measure H is called a prior probability measure of the parameter θ. The prior measure H belongs to some given family of probability measures $\mathscr{H}$.

3) The set of such possible decisions D that each element d from D is a measurable function on Ω. In estimation theory the set of decisions D may contain all estimates of the parameter θ or some function $R(\theta)$ measurable on Ω.

4) The loss functions $L(\theta, d)$ (or $L(R(\theta), d)$) determined on $\Theta \times D$. This loss function determines the losses caused by an erroneous estimation, that is, by the replacement of the parameter θ by the decision element d. It is assumed later that the families $\mathscr{B}$ and $\mathscr{H}$ are dominated by some σ-finite measures μ and ζ respectively. If we denote the densities

$$f(x \mid \theta) = P_\theta\{dx\}/\mu\{dx\}, \quad h(\theta) = H\{d\theta\}/\zeta\{d\},$$

which exist in a view of the Radon-Nikodym theorem, the joint density of the probability distribution for the random variables X and θ takes on the form

$$g(x,\theta) = f(x \mid \theta)h(\theta).$$

In accordance with the Bayes theorem, the conditional density for θ given $X = x$ is called the *posterior probability density function* (p.d.f.) of the parameter θ and is written as

$$\hbar(\theta \mid X = x) = \frac{f(x \mid \theta)h(\theta)}{\int_{\Theta} f(x \mid \theta)H\{d\theta\}}, \quad \theta \in \Theta, \tag{2.1}$$

for each $x \in \Omega$ such that

$$f(x) = \int f(x \mid \theta)h(\theta)\zeta\{d\theta\} > 0.$$

If Y is a statistic on $(\Omega, \mathscr{L}, P_\theta)$, then the probability measure P_θ can be obtained by transformation into P_θ^Y. If $f^Y(y \mid \theta) = P_\theta^Y\{dy\}/\mu\{dy\}$ is the density of the probability measure P_θ^Y, then the posterior p.d.f. of the parameter θ given $Y(X) = x$ has the form

$$\hbar^Y(\theta \mid Y = y) = \frac{f^Y(y \mid \theta)h(\theta)}{\int_{\Theta} f^Y(y \mid \theta)h(\theta)\zeta\{d\theta\}}. \tag{2.2}$$

Further, for the sake of simplicity, we will use for the prior and posterior p.d.f. appearing in the Bayes formulas (2.1) and (2.2) the notations $h(\theta)$ and $h(\theta \mid x)$, respectively. Since the denominator in (2.1) and (2.2) is independent of θ and is determined only by the observation x (or by the statistic y), we will determine only the kernel of the prior density using the symbol of proportionality "$\backsim$". So, instead of the expression (2.1), we will write

$$\hbar(\theta \mid x) \backsim f(x \mid \theta)h(\theta) \tag{2.3}$$

taking into account the fact that the normalizing factor of the p.d.f., $\hbar(\theta \mid x)$ has been found from the integral

$$\beta = \left[\int_{\Theta} f(x \mid \theta)h(\theta)\zeta\{d\theta\}\right]^{-1}. \tag{2.4}$$

In the case when the parameter θ takes on the discrete values $\theta_1, \theta_2, \ldots, \theta_k$, the prior distribution is given in the form of prior probabilities $p_j = P\{\theta = \theta_j\}$, $j = 1, \ldots, k$. The expressions (2.3) and (2.4) are also valid for this case, if one represents $h(\theta)$ and $\hbar(\theta)$ by means of a delta-function. In particular,

$$h(\theta) = \sum_{j=1}^{k} p_j \delta(\theta - \theta_j).$$

The Bayes formula (2.3) lets us find the posterior density of the parameter θ in the form

$$\hbar(\theta \mid x) = \sum_{j=1}^{k} \bar{p}_j \delta(\theta - \theta_j).$$

where $\bar{p}_j = P\{\theta = \theta_j \mid x\}$, $j = 1, \ldots, k$, are the posterior probabilities.
For the discrete values of the parameter θ, the Bayes formula is often written in the form

$$p_j = \frac{p_j f(x \mid \theta_j)}{\sum_{i=1}^{k} p_i f(x \mid \theta_i)}, \quad j = 1, 2, \ldots, k. \tag{2.5}$$

The choice of a *loss function* plays an important role in the theory of Bayes estimating. In most cases the loss function is represented in the following form:

$$L(\theta, d) = C(\theta) W(|d - \theta|), \tag{2.6}$$

where $W(0) = 0$, and $W(t)$ is a monotonically increasing function for $t > 0$; $C(\theta)$ is assumed to be positive and finite. The prior Bayes estimate $\hat{\theta}_H$ is defined as an element from D which minimizes the prior risk [272]

$$G(H, d) = \int_{\Omega} f(x) \mu\{dx\} \int_{\Omega} C(\theta) W(|d(x) - \theta|) H\{d\theta \mid x\} C(\theta) W(|d - \theta|), \tag{2.7}$$

After the X has been observed, the most handy function (from the Bayes point-of-view) for further consideration is not the prior risk (2.7) but the posterior one, having the form

$$\bar{G}(H, d) = \int_{\Theta} C(\theta) W(|d(x) - \theta|) H\{d\theta \mid X\}. \tag{2.8}$$

The Bayes estimate of the parameter θ with respect to the prior distribution H should be the element $\theta_H(X)$ of the set D, minimizing the posterior risk with given X:

$$\int_{\Theta} C(\theta) W(|\hat{\theta}_H(x) - \theta|) H\{d\theta \mid X\} = \inf_{d \in D} \int_{\Theta} C(\theta) W(|d(x) - \theta|) H\{d\theta \mid X\}. \tag{2.9}$$

The analysis of the problem in the above-given setting shows that investigation of the specific solution with the given testing scheme is based on the following three questions:

1) a choice of the family of probability measures $\mathscr{B}$;
2) a choice of the prior distribution H;
3) a choice of the loss function $L(d, \theta)$.

The first question, having a practical significance, is associated with completeness of the statistical model, used by the researcher. The other two are less specific. Some recommendations on a choice of H and $L(d, \theta)$ will be given below.
In applied statistical analysis, the interval estimates are frequently used. The Bayes theory operates with an analogous notion having, however, interpretation which differs from the

classical one. In the simplest case of a scalar parameter θ, a Bayes confidence interval $\underset{\sim}{\theta}$ is introduced by the expression

$$\int_{\underset{\sim}{\theta}} \hbar(\theta \mid x) d\theta = \gamma,$$

where γ is the confidence probability. Since the choice of $\underset{\sim}{\theta}$ can be established in many ways, one adds an additional requirement: the interval $\underset{\sim}{\theta}$ must have the minimal length. In the case of a vector parameter θ, the confidence interval is chosen from the condition

$$\int_{\underline{R} \leqslant R(\theta) \leqslant \overline{R}} \hbar(\theta \mid x) d\theta = \gamma;$$

moreover, the difference $\overline{R} - \underline{R}$ must be the smallest. As seen from the definition above, the classical and the Bayes confidence intervals have different interpretations. In the classical form, the confidence interval, "covering" with a given probability an unknown parameter, is random. In the Bayes approach the parameter is random, while the confidence interval has fixed limits, defined by the prior density and confidence probability γ.

2.2 Classical properties in reference to Bayes estimates

According to classical statistics the quality of statistical estimates may be characterized by: how much these estimates satisfy the requirements of consistency, unbiasedness, effectiveness and sufficiency. As a rule, a classical estimate, approved in each particular case, appears to be a compromise, that is, we give preference to some property to the detriment of the others. As was mentioned, the leading property of the Bayes estimate is its optimality. The classical properties, indicated above, are not adequate to the Bayes methodology. Many authors use them only to keep up the tradition. Here we present some results which modify the classical estimates in the Bayes approach.

2.2.1 *Sufficiency*

The property of sufficiency works smoothly. The Bayes formula (2.1) (or (2.2)) is connected with the Bayes definition of the sufficient statistics. A posterior p.d.f. of the parameter θ, generated by the sufficient statistics $S(X)$, is equivalent (from the Bayes point-of-view) to the posterior p.d.f. of the parameter θ, constructed on the initial observation, that is:

$$\hbar^s(\theta \mid S(X)) = \hbar(\theta \mid X).$$

At the same time, this proves the equivalence of the Bayes and traditional definitions of sufficiency (see, e.g., [272]). Thus, the property of sufficiency in the Bayes estimating theory keeps its classical form.

2.2.2 *Consistency*

With respect to the property of consistency we cannot draw the previous conclusion. Reasoning rigorously, we can ascertain that a traditional analysis of an estimate behavior with the sample size tending to infinity contradicts the essence of the Bayes approach. Indeed, if it is assumed that θ is a random parameter with nondegenerating prior p.d.f., $h(\theta)$, then it is senseless to investigate the asymptotical properties of convergence of the estimate $\hat{\theta}_H$ with respect to $\theta = \theta_0$. By the same reason, it is incorrect when computing the mean value, to compare it with the random variable θ. There are, however, representations that make possible the investigation of the estimate $\hat{\theta}_H$ for large samples. If we assume that a chosen prior distribution is not exact, then one can get the estimate with the help of the Bayes scheme, and, later on, investigate it, digressing from the method used in obtaining the estimate (bearing in mind classical theory). In many cases the Bayes estimates are consistent and converge frequently to the maximum likelihood estimate, MLE.

Zacks [272] provides an example when the estimate of the Poisson cumulative distribution function $p(i \mid \lambda) = \mathrm{e}^{-\lambda}\lambda^i/i!$ with the prior $\lambda \backsim \Gamma(1,1)$ (gamma p.d.f.) is consistent. Lindley [146] proves that, if $\hat{\theta}_n$ is the best normal asymptotical estimate, then one can ascertain that Bayesian estimates and MLE are equivalent. The exact form of the prior distribution in this case is not significant, since for the samples with a large size MLE can be replaced with the unknown parameter θ. Bickel and Yahow present [24] the strict proof for the assertion, analogous to the case of the one-parametric exponential family and a loss function having a quadratic form. Asymptotical properties of the Bayes estimators for the discrete cumulative distribution functions were investigated by Freedman [87].

We give the part of Jeffrey's' reasoning's [115] with respect to the properties of the posterior p.d.f. for large-size samples. For a scalar parameter θ, and in accordance with (2.3), let p.d.f.

$$\hbar(\theta \mid x) \backsim h(\theta)\,\ell(\theta \mid x) = h(\theta)\mathrm{e}^{\ln \ell(\theta|x)}, \tag{2.10}$$

where $\ell(\theta \mid x)$ is the likelihood function, being, by essence, a p.d.f. of the observed values $x = (x_1, x_2, \ldots, x_n)$ of a trial and coinciding with $f(x \mid \theta)$.

(The parameter θ is assumed to be an argument of a likelihood function.) It is supposed that $h(\theta)$ and $\ell(x \mid \theta)$ ($\theta \in \Theta$) are nondegenerating and have continuous derivatives; moreover,

$\ell(x \mid \theta)$ has a unique maximum at the point $\hat{\theta}_{\text{m.l.}}$, which is the MLE, generally speaking, $\ln[\ell(\theta \mid x)]$ has the order n, and $h(\theta)$ is independent on the sample size. Thus, it is intuitively clear that for the large-size samples, the likelihood cofactor is dominating in the posterior p.d.f.

Bernstein [169] and Mizes [166] prove the more general statement. The main thrust of this theorem is that, if the prior p.d.f. of the parameter θ is continuous, then, while the number of observations is increasing, the posterior p.d.f. is tending to a limit (which can be found analytically) independent of the prior distribution. Furthermore, since under the more common conditions, the p.d.f. form approaches, with a growth of n, the normal distribution curve centered on the MLE and the posterior p.d.f. for the case of the large-size samples appears to be normal also with the mean value $\hat{\theta}_{\text{m.l.}}$.

The proof of the asymptotical normality of the posterior p.d.f., $\hbar(\theta \mid x)$, can be carried out as follows. Let us expand into a Taylor series the functions $h(\theta)$ and $\ell(\theta \mid x)$ and at the MLE $\hat{\theta}_{\text{m.l.}}$:

$$\begin{aligned} h(\theta) &= h(\hat{\theta}_{\text{m.l.}}) + (\theta - \hat{\theta}_{\text{m.l.}})h'(\hat{\theta}_{\text{m.l.}}) + \frac{1}{2}(\theta - \hat{\theta}_{\text{m.l.}})^2 h''(\hat{\theta}_{\text{m.l.}}) + \cdots \\ &= h(\hat{\theta}_{\text{m.l.}}) \left[1 + \frac{(\theta - \hat{\theta}_{\text{m.l.}})h'(\hat{\theta}_{\text{m.l.}})}{h(\hat{\theta}_{\text{m.l.}})} + \frac{1}{2} \frac{(\theta - \hat{\theta}_{\text{m.l.}})^2 h''((\hat{\theta}_{\text{m.l.}})}{h(\hat{\theta}_{\text{m.l.}})} + \cdots \right] \end{aligned}$$

Denoting by $g(\theta) = \ln \ell(\theta \mid x)$ and taking into account the relation $g'(\hat{\theta}_{\text{m.l.}}) = 0$, we obtain

$$\exp[g(\theta)] \backsim \exp\left[\frac{1}{2}(\theta - \hat{\theta}_{\text{m.l.}})^2 g''(\hat{\theta}_{\text{m.l.}})\right] \times \left[1 + \frac{1}{6}(\theta - \hat{\theta}_{\text{m.l.}})^3 g'''(\hat{\theta}_{\text{m.l.}}) + \cdots\right],$$

where the last equation is obtained by the expansion $e^x = 1 + x + \cdots$.

Multiplication of these expansions gives us

$$\begin{aligned} \hbar(\theta \mid x) \backsim \exp&\left[\frac{1}{2}(\theta - \hat{\theta}_{\text{m.l.}})^2 g''(\hat{\theta}_{\text{m.l.}})\right] \\ &\left[1 + \frac{(\theta - \hat{\theta}_{\text{m.l.}})h'(\hat{\theta}_{\text{m.l.}})}{h(\hat{\theta}_{\text{m.l.}})} + \frac{1}{2}\frac{(\theta - \hat{\theta}_{\text{m.l.}})^2 g''(\hat{\theta}_{\text{m.l.}})}{h(\hat{\theta}_{\text{m.l.}})} + \frac{1}{6}(\theta - \hat{\theta}_{\text{m.l.}})^3 g'''(\hat{\theta}_{\text{m.l.}}) + \cdots\right] \end{aligned} \tag{2.11}$$

The dominating factor has the form of a normal p.d.f. with the mean value equal to MLE $\hat{\theta}_{\text{m.l.}}$ and the variance

$$[-g''(\hat{\theta}_{\text{m.l.}})]^{-1} = \left[-\frac{d^2 \ln \ell(\theta \mid x)}{d\theta^2}\right]^{-1}_{\theta = \hat{\theta}_{\text{m.l.}}}$$

Thus, if we use only the dominating cofactor, the approximation of the posterior p.d.f. of θ for large sizes of the sample n takes the form

$$\hbar(\theta \mid x) = \frac{1}{\sqrt{2\pi}} |g''(\hat{\theta}_{\text{m.l.}})|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\theta - \hat{\theta}_{\text{m.l.}})^2 |g''(\hat{\theta}_{\text{m.l.}})|\right] \tag{2.12}$$

Since $|g''(\hat{\theta}_{\text{m.l.}})|$ usually is a function of n, then with the growth of n, the posterior p.d.f. takes on the more pointed form. Jeffrey's [115] points out the fact that this approximation gives errors of order $n^{-1/2}$. Koks and Hinkly [125] generalize these results to the case of a vector parameter θ.

To justify (2.12), one may only use the fact that the likelihood must concentrate with increasing order about its maximum. Hence, these conclusions can be used more broadly than the case of independent random variables with the same distributions. Dawid [54] and Walker [262] carry out a scrupulous investigation of a set of regularity conditions under which the posterior distribution with the probability equal to unity is asymptotically normal. These conditions are almost alike as regularity conditions which are necessary for the asymptotic normality of the MLE. The research on the application of the expansion (2.11) for statistical outcomes may be found in works by Lindley [143] and Johnson [117].

The relation between the consistency of Bayes estimates and MLE is studied in the work by Strasser [242]. The point is that there is an example by Schwartz [231] in which the MLE is nonconsistent, but the Bayes estimate, under the same conditions, possesses the property of consistency. Strasser [242] complemented the regularity conditions up to the strict consistency of the MLE [272] by the conditions for the prior measure H so that the consistency of the maximal likelihood estimator implies the consistency of a Bayes estimator. The investigation of the relations between the consistency of a Bayes estimator and MLE is carried on by Le Cam [136].

If it is assumed that the parameter θ is fixed but unknown, then the consistency of a Bayes estimator can be investigated more naturally and nearer to the classical interpretation. These questions are considered by De Groot [63], Berk [19], Mizes [166] and by other authors. They may be interpreted as follows: if $x_1, x_2, \ldots, x_n$ is a sample from the distribution with the unknown parameter θ and if the value of θ is actually equal to θ_0, then, as $n \to \infty$, $\hbar(\theta \mid x)$ will concentrate more strongly about the value θ_0. The estimate of the parameter constructed on such a posterior distribution may be named, to all appearances, as consistent.

We give a brief explanation for this phenomenon. Let the parameter θ take only the finite set of values $\theta_1, \theta_2, \ldots, \theta_k$. Suppose that $P\{\theta = \theta_i\} = p_i$ for $i = 1, 2, \ldots, k$, and for each given value $\theta = \theta_i$, the random variables $x_1, x_2, \ldots, x_n$ generate a sample from the distribution with the p.d.f., f_i. It is assumed also that all f_i are different in such a sense that, if Ω is a sample space corresponding to the single observation, then

$$\int_{\Omega} |f_i(x) - f_j(x)| d\mu(x) > 0, \quad \forall\, i \neq j.$$

Let, for the observed values $x_1, x_2, \ldots, x_n$, $\bar{p}_i$ denote the posterior probability of the event $\theta = \theta_i$ for which, due to the Bayes theorem, we have

$$\bar{p}_i = \frac{p_i \prod_{j=1}^{n} f_i(x_j)}{\sum_{r=1}^{k} \Pr \prod_{j=1}^{n} f_r(x_j)}, \quad i = 1, 2, \ldots, k.$$

Suppose now that $x_1, x_2, \ldots, x_n$ is the sample from the distribution with the p.d.f., f_t, where t is some of the values $1, 2, \ldots, k$. As was shown in [63], the following limit relations are valid with the probability equal to one:

$$\lim_{n\to\infty} \bar{p}_t(x) = 1, \quad \lim_{n\to\infty} \bar{p}_i(x) = 0, \quad \forall\, i \neq t.$$

We give another example with a continuous distribution of the parameter θ which will be a mean value of the Gaussian random variable with a given measure of exactness r (the variance is equal to r^{-2}). Suppose that a prior distribution θ is Gaussian with a mean value μ and exactness measure τ ($-\infty < \mu < \infty$, $\tau > 0$). It is easy to check that the posterior distribution θ is Gaussian with the mean value

$$\mu' = \frac{\tau\mu + nr\hat{\mu}}{\tau + nr}, \text{ where } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

with the exactness measure $\tau + nr$. Let us rewrite the expression for μ' in the form

$$\mu' = \frac{\tau\mu}{\tau + nr} + \hat{\mu}\,\frac{nr}{\tau + nr}.$$

Assume now that the sample $x_1, x_2, \ldots, x_n$ is factually taken from the Gaussian distribution with the mean value θ_0. In accordance with the law of large numbers (see, for example, Kolmogorov's second theorem [208]), $\hat{\mu} = \theta_0$ with the probability 1. At the same time, it follows from the formula for μ', $\mu' \to \hat{\mu}$ as $n \to \infty$. In this sense, μ' is consistent. Furthermore, since the posterior variance of the parameter θ tends to zero, the posterior distribution of θ converges to the degenerated distribution, concentrated at the point θ_0.

2.2.3 *Unbiasedness*

Here, evidently, we meet a situation analogous to those in the above investigation of consistency. Since the parameter θ appears to be random, it is absurd to attempt to find the estimate $\hat{\theta}$ in the form $E[\hat{\theta}]$. Due to this circumstance, many authors, including Ferguson [80], proceed in the following way: at first they obtain the Bayes estimate in some definite form, thereafter they "forget" about the parameter randomness and investigate unbiasedness in the usual interpretation of this term.

Sage and Melsa attempt to give the Bayes definition of unbiasedness. In particular, they give two definitions:

1) a Bayes estimate $\hat{\theta}(x)$ is called *unbiased*, if

$$E[\hat{\theta}] = E[\theta(x)]; \text{ and}$$

2) a Bayes estimate $\hat{\theta}(x)$ is called *conditionally unbiased*, if

$$E[\hat{\theta}(x) \mid \theta] = E[\theta].$$

The second definition is more essential, since, in the first place, it is closer (by sense) to the classical one, and in the second place, is clearer than the first definition. Since $E[\hat{\theta}]$ is a mean value with respect to the prior measure, the equality $E[\hat{\theta}(x)] = E[\theta]$ corresponds to that ideal scheme when the posterior distribution coincides with the prior one (it is, however, not necessary). According to this, the Bayes estimate cannot be unbiased. Bland [25] gives the example of the statistical model which proves the incorrectness of the first definition.

The second definition of unbiasedness, mentioned above, is used by Hartigan in his work [102]. The estimate $\hat{\theta}(x)$, obeying the condition $E[\hat{\theta}(x) \mid \theta] = E[\theta]$.

The possibility of using the first definition is not even discussed. Moreover, he introduces the definition of the exact Bayes estimate which must satisfy the condition $P\{\hat{\theta}(x) \neq \theta\} = 0$. Hartigan proves that an unbiased Bayes estimator is exact.

2.2.4 *Effectiveness*

We say that one Bayes estimate is more effective than the other bearing in mind the following reason. If the posterior probability is chosen as a measure of effectiveness, its value, for identical observations, will be defined by the prior distribution (on the whole by the prior variance of the parameter). Consequently, the comparison criterion is not objective, since the prior distribution may be chosen, in some sense, arbitrarily. If we use the same prior distribution, then in view of the uniqueness of the Bayes solution, there are no two Bayes estimates for the same observation.

2.3 Forms of loss functions

As was mentioned in § 2.1, the loss function $L(d,\theta)$, where usually $d = \hat{\theta}$, is frequently represented in the form

$$L(d,\theta) = C(\theta)W(|d-\theta|). \tag{2.13}$$

Here $W(0) = 0$ and $W(t)$ is a monotonically increasing function; $C(\theta)$ is assumed to be positive and finite. As mentioned by Le Cam [136], the loss function of the form (2.13) was

proposed by Laplace who understood that the exact expression for the function W cannot be found analytically.

The choice of the loss function is an important question in the theory of Bayes statistical estimating. Accuracy of such a choice stipulates the estimate quality. Frequently authors use a squared-error loss function, for which $W(|d-\theta|) = (\theta - d)^2$. This function gives a good approximation for any loss function of the form $C(\theta)W(|d-\theta|)$ having a smooth character in the neighborhood of the origin. Another function which is the convex loss function satisfying the condition $W(|d-\theta|) = (d-\theta)^k$, $k \geqslant 1$.

It is well known that the Bayes estimator of the function $R(\theta)$ for the squared-error loss function has a form of the posterior mean value

$$\hat{R}^* = \int_{\Theta} R(\theta)\hbar(\theta \mid x)d\theta.$$

For $k = 1$, the median of the posterior distribution appears to be the Bayes estimator. The theory of Bayes estimating for convex loss functions was developed by De Groot and Rao [64]. Rukhin [214] proves a theorem for which the estimate of the parameter θ is the same for any convex loss function, if only the posterior density is unimodal and symmetric with respect to θ. The loss functions represented above are unbounded. This circumstance may be a reason for misunderstanding. In particular, Girshik and Savidge [88] give an example in which the Bayes estimate, minimizing the posterior risk, has an infinite prior risk.

In a number of works there are also other loss functions whose properties are connected with the properties of the statistical models and peculiarities of the phenomena investigated by researchers. The direct generalization of the squared-error loss function will be the relative squared-error loss function given by:

$$L_{S_1}(\hat{\theta}, \theta) = \left(\frac{\theta - \hat{\theta}}{\theta}\right)^2 \tag{2.14}$$

and its modification

$$L_{S_2}(\hat{\theta}, \theta) = \left(\frac{\theta^\beta - \hat{\theta}^\beta}{\theta^\beta}\right)^2, \quad \beta > 0, \tag{2.15}$$

Which are broadly used when the investigations are directed precisely toward the relative errors. As shown by Higgins and Tsokos [108], the Bayes estimator of $R(\theta)$ under some loss function, minimizing the posterior risk with the loss function (2.15), is computed in

the following manner:

$$\hat{R}^*_\beta = \left[\frac{\int_\theta \frac{\hbar(\theta \mid x)}{[R(\theta)]^\beta}\, d\theta}{\int_\theta \frac{\hbar(\theta \mid x)}{[R(\theta)]^{2\beta}}\, d\theta} \right]^{1/\beta}.$$

Harris [99] proposes to use for the investigation of probability of a nonrenewal system being operated without breakdowns with a loss function given by:

$$L_H(\hat{\theta}, \theta) = \left| \frac{1}{1-\hat{\theta}} - \frac{1}{1-\theta} \right|^k. \tag{2.16}$$

Thereafter he states: "If the reliability of a system is 0.99, then it fails, in the average, only once in 100 trials; if, at that time the system reliability is 0.999, then it fails only once in 1,000 trials, that is, this is ten times better. Therefore, the loss function must depend on how well we can estimate the quantity $(1-\theta)^{-1}$."

Higgins and Tsokos [108] propose to use the loss function of the form

$$L_e(\hat{\theta}, \theta) = \frac{f_1 \, \mathrm{e}^{-f_2(\hat{\theta}-\theta)} + f_2 \, \mathrm{e}^{-f_1(\hat{\theta}-\theta)} f_1 + f_2}{-1}, \quad f_1 > 0, \;\; f_2 > 0, \tag{2.17}$$

which enforces the losses, if the estimate is substantially different from the parameter. It is interesting that for small $\theta - \hat{\theta}$,

$$L_e(\hat{\theta}, \theta) = \frac{f_1 f_2}{2}(\theta - \hat{\theta})^2 + \mathrm{O}\big((\theta - \hat{\theta})^3\big).$$

The authors compare Bayes estimator of probability of failures, the mean time prior to failures with the reliability function for squared-error loss function with those having the form (2.15), (2.16), (2.17), and a linear loss function of the following form:

$$L_p(\hat{\theta}, \theta) = \begin{cases} p|\theta - \hat{\theta}|, & \hat{\theta} \leqslant \theta, \\ (1-p)|\theta - \hat{\theta}|, & \hat{\theta} > \theta, \end{cases} \tag{2.18}$$

Which generalizes the function $L(\hat{\theta}, \theta) = |\theta - \hat{\theta}|$ mentioned above on the case for unequal significance for exceeding and underestimating of the estimate $\hat{\theta}$ with respect to the parameter θ. The conclusions given by Tsokos and Higgins may be interpreted as follows:

1) The quadratic loss function is less stable in comparison with the others we have considered. If the quadratic loss function is used for approximating, then the obtained approximation of the Bayes estimate is unsatisfactory.
2) The Bayes estimator is very sensitive with respect to the choice of loss function.
3) The choice of loss function should be based, not on the mathematical conveniences, but on the practical significance.

The loss function which uses the fact that exceeding the parameter is worse than decreasing (this is intrinsic, for example, for the reliability measure) is written by Cornfield [47] in the form

$$L(\hat{\theta},\theta)=\begin{cases} K_1\left(\dfrac{\hat{\theta}}{\theta}-1\right)^2, & \hat{\theta}\leqslant\theta, \\ K_1\left(\dfrac{\hat{\theta}}{\theta}-1\right)^2+K_2\left(\dfrac{\hat{\theta}}{\theta}-1\right)^2, & \hat{\theta}>\theta. \end{cases} \tag{2.19}$$

In order to ensure for the loss function the different significance for the positive and negative errors, Zelner [274] introduces the so called linearly-exponential loss function

$$L_{\mathrm{EX}}(\hat{\theta}-\theta)=b\left[\mathrm{e}^{a(\hat{\theta}-\theta)}-a(\hat{\theta}-\theta)-1\right] \tag{2.20}$$

This function is asymmetric and is nearly symmetric for small a and can be well approximated by the quadratic functions. We give an example of the estimator which is obtained from the loss function (2.20). If X is a Gaussian random variable with the given mean value θ and given variance σ^2, and the prior distribution density θ satisfies the condition $h(\theta)\backsim$ const., then

$$\hat{\theta}^*=\bar{x}=\frac{a\sigma^2}{2n},$$

where $\bar{x}$ is a sample mean value. It is not difficult to verify that for small a and/or for large sample sizes n, the estimator $\hat{\theta}^*$ is near the MLE. The loss function (2.20) solves actually almost the same problem as that in (2.19). But the last loss function is not so handy in calculations because we cannot find with it the desired estimates in close analytical form; instead we have to use special numerical methods to obtain the desired approximation.

El-Sayyad [73] uses, in addition to the loss functions given above, the following loss function:

$$L_{\alpha\beta}(\hat{\theta},\theta)=\theta^{\alpha}(\hat{\theta}^{\beta}-\theta^{\beta})^2, \tag{2.21}$$

and

$$L_{\ln}(\hat{\theta},\theta)=(\ln\hat{\theta}-\ln\theta)^2, \tag{2.22}$$

Smith [235] determines the class of bounded loss functions A given by the conditions: the loss unction is symmetric with respect to $|\hat{\theta}-\theta|$, decreases with respect to $|\hat{\theta}-\theta|$ and satisfies the conditions

$$\sup_{\hat{\theta},\theta}L(\hat{\theta},\theta)=1,\quad \inf_{\hat{\theta},\theta}L(\hat{\theta},\theta)=0.$$

Smith scrupulously investigates the Bayes estimates of the so-called step loss function

$$L_b(\hat{\theta},\theta) = L_b(\hat{\theta}-\theta) = \begin{cases} 0 & \text{if } |\hat{\theta}-\theta| < b, \\ 1 & \text{if } |\hat{\theta}-\theta| \geqslant b, \end{cases} \tag{2.23}$$

Estimators were found for many parametric families. These estimators differ substantially from the Bayes estimators with the squared-error loss function.

2.4 The choice of a prior distribution

The choice of prior distribution in applied problems of Bayes estimating is one of the most important questions. At the same time, the solution of this problem doesn't touch the essence of the Bayes approach. The existence of a prior distribution is postulated. All further arguments are based on this postulation. Some authors, however, investigate the question of choice of a prior distribution being in the framework of the Bayes approach. We propose the following three recommendations for the choice of a prior distribution. They are, correspondingly, based on: 1) the conjugacy principle; 2) the absence of information; and 3) the information criterion. We shall discuss individually each of these recommendations.

2.4.1 *Conjugated prior distributions*

Each prior distribution, due to the Bayes theorem, can be used together with any likelihood function. It is convenient, however, to choose a prior distribution of a special form giving the simple estimators. For a given distribution $f(x \mid \theta)$ we may find such families of prior p.d.f. that a posterior p.d.f. will be the elements of the same family. Such a family is called *closed with respect to the choice or conjugated with respect to* $f(x \mid \theta)$. It is said sometimes: "naturally-conjugated family of prior distributions". Most of the authors state that this approach is dictated by the convenience of theoretical arguments and practical conclusions. Haifa and Shleifer [202] attempt to give a more convincing justification of a conjugated prior distribution. We discuss this question in detail.

Assume that sample distributions are independent and allowing the sufficient statistics $\mathbf{y}$ of fixed dimension with the domain $\Omega_{\mathbf{y}}$. A family $\mathscr{H}$ of all prior distributions is constructed in the following way: each element $\mathscr{H}$ is associated with the element $\Omega_{\mathbf{y}}$. If, prior to a trial, for θ is chosen the element from $\mathscr{H}$ corresponding to $\mathbf{y}' \in \Omega_e$ and a sample gives the sufficient statistics $\mathbf{y}$, then the posterior distribution also belongs to $\mathscr{H}$ and assigns to some element $\mathbf{y}'' \in \Omega_e$. For the definition of $\mathbf{y}''$ with the help of $\mathbf{y}$ and $\mathbf{y}'$ a binary

operation $\mathbf{y}'' = \mathbf{y}' * \mathbf{y}$ is introduced. We consider below the formalization of a conjugated prior distribution given in [202].

It is supposed that for arbitrary samples $x = (x_1, x_2, \ldots, x_n)$ with each fixed n there is a sufficient statistic

$$\mathbf{y}_n = (x_1, x_2, \ldots, x_n) = \mathbf{y} = (y_1, y_2, \ldots, y_s),$$

where y_j is a real number and the dimension of the vector $\mathbf{y}$ is independent of n.

For any given n and arbitrary sample $(x_1, x_2, \ldots, x_n)$, there exists a function k and s-dimensional vector $\mathbf{y} = (y_1, y_2, \ldots, y_s)$ consisting of real numbers, that the likelihood function satisfies the relation

$$\ell_n(\theta \mid x_1, x_2, \ldots, x_n) \backsim k(\theta \mid \mathbf{y}).$$

The function $k(\theta \mid \mathbf{y})$ is called a *likelihood kernel*. We will touch upon an important property of the kernel $k(\theta \mid \mathbf{y})$.

Theorem 2.1. *Let* $\mathbf{y}^{(1)} = y_p(x_1, x_2, \ldots, x_p)$ *and* $\mathbf{y}^{(2)} = y_{n-p}(x_{p+1}, \ldots, x_n)$. *Then we can find such a binary operation* $*$ *that satisfies the relation*

$$\mathbf{y}^{(1)} * \mathbf{y}^{(2)} = \mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_s^*)$$

and possesses the following properties:

$$\ell_n(\theta \mid x_1, x_2, \ldots, x_n) \backsim k(\theta \mid \mathbf{y}^*),$$

and

$$k(\theta \mid \mathbf{y}^*) \backsim k\big(\theta \mid \mathbf{y}^{(1)}\big)\, k\big(\theta \mid \mathbf{y}^{(2)}\big).$$

As it follows from the theorem, $\mathbf{y}^*$ can be found only by using $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, without $(x_1, x_2, \ldots, x_n)$.

The posterior p.d.f., is constructed with the help of the kernel function $k(\theta \mid \mathbf{y})$ in the following way:

$$h(\theta \mid \mathbf{y}) = N(\mathbf{y})\, k(\theta \mid \mathbf{y}), \tag{2.24}$$

where $\mathbf{y}$ is some statistic, $N(\mathbf{y})$ is a function which needs to be defined.

In order for the function $h(\theta \mid \mathbf{y})$, defined on Θ by the relation (2.24), to be a p.d.f., it is necessary and sufficient that this function be nonnegative everywhere, and the integral from this function over Θ will be equal to unity. Since $k(\theta \mid \mathbf{y})$ is a kernel function of a joint p.d.f. of observations, defined on Θ for all $\mathbf{y} \in \Omega_{\mathbf{y}}$, is necessarily nonnegative for all

$(\mathbf{y},\theta)$ from $\Omega_{\mathbf{y}} \times \Theta$. Consequently, if there exists the integral of $k(\theta \mid \mathbf{y})$ over Θ, then $N(\mathbf{y})$ is determined by the relation

$$[N(\mathbf{y})]^{-1} = \int_{\Theta} k(\theta \mid \mathbf{y}) d\theta$$

and $h(\theta \mid \mathbf{y})$, represented by the expression (2.24), will be a p.d.f.

Suppose now that $\mathbf{y}$ is a sufficient statistic, determined with the help of the observed sample $(x_1, x_2, \ldots, x_n)$, and $h(\theta)$ is a prior p.d.f. In accordance with the Bayes theorem for the posterior distribution density we have

$$\hbar(\theta \mathbf{y}) \backsim h(\theta) k(\theta \mid \mathbf{y})$$

If now $h(\theta)$ is a p.d.f., conjugated to the kernel k with the parameter $\mathbf{y}' \in \Omega_{\mathbf{y}}$, that is, $h(\theta) \backsim k(\theta \mid \mathbf{y}')$, then, in accordance with the Bayes theorem,

$$h(\theta \mid \mathbf{y}) \backsim k(\theta \mid \mathbf{y}') k(\theta \mid \mathbf{y}) \backsim k(\theta \mid \mathbf{y}' * \mathbf{y}).$$

Thus: 1) the kernel of a prior p.d.f. is combined with the kernel of a sample in a manner similar to the combination of two sample kernels; 2) both a prior and posterior p.d.f. are induced by the same likelihood kernel, but their generating statistics are different. These conclusions may be interpreted as a follows: the prior distribution is a result of processing some nonexistent data (or data which exist but are lost) for the same statistical model as a likelihood function.

Let us consider the following example. A Bernoulli process with the parameter $\theta = p$ induces independent random variables $(x_1, x_2, \ldots, x_n)$ with the same probabilities $p^x(1-p)^{1-x}$, where $x = 0, 1$. If n is the number of observed values and $r = \sum x_i$, then the likelihood of a sample is written as

$$\ell_n(p \mid x_1, x_2, \ldots, x_n) \backsim p^r (1-p)^{n-r}.$$

In addition to this, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2) = (r, n)$ is a sufficient statistic whose dimension is equal to 2, independently of n. A prior p.d.f. conjugated with the likelihood kernel and induced by the statistics $y' = (r', n')$ is a density of the beta distribution

$$h(p) = \frac{p^{r'}(1-p)^{n'-r'}}{B(r'+1, n'-r'+1)}, \quad 0 \leqslant p \leqslant 1,$$

where

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

In view of the Bayes theorem

$$\hbar(p \mid \mathbf{y}) \backsim h(p) p^r (1-p)^{n-r} \backsim p^{r'+r}(1-p)^{n+n'-(r+r')},$$

that is, the kernels of the prior and posterior p.d.f. coincide, and the beta distribution with the parameters $r'' = r + r'$ and $n'' = n + n'$ appears to also be posterior.

A family of conjugated posterior d.d. may be enlarged by the extension of the domain $\Omega_{\mathbf{y}}$ up to and including all values for which $k(\theta \mid \mathbf{y})$ is nonnegative for all θ, and the integral of $k(\theta \mid \mathbf{y})$ over the domain Θ is convergent.

In the example we have considered, the parameters r and n take the values of positive integers. At the same time, an integral of $k(\theta \mid \mathbf{y})$ over $\Theta[0,1]$ converges for all real $r > -1$ and $n > -1$ to the complete beta function. Therefore, we can obtain, assuming that the parameter $\mathbf{y} = (r,n)$ may take an arbitrary value from the domain determined in such a way, the family of densities which is substantially broader.

A complicated report on naturally conjugated p.d.f. is given in the monographs [91] and [202]. Dawid and Guttman [55] investigate the question of obtaining conjugated distribution in foreshortening of singularities of models. In particular, it is shown that simple forms of conjugated distributions are an implication of a group structure of models.

2.4.2 *Jeffrey's introduces prior distributions representing a "scantiness of knowledge"*

These distributions are the subject of consideration in the work by Zelner [272] and also are investigated by other authors. This assumption is an implication of the desire not to leave the Bayes approach in the cases when an investigator doesn't have enough knowledge about the properties of the model parameters or knows nothing at all. Jeffrey's [115] proposes two rules for the choice of a prior distribution, which, in his opinion, "embrace the most widespread situations", when we don't have the information on the parameter:

1) if the parameter exists in the finite segment $[a,b]$ or in the interval $(-\infty,+\infty)$ then its prior probability should be supposed to be uniformly distributed;
2) if the parameter takes the value in the interval $[0,\infty)$, then the probability of its logarithm should be supposed to be uniformly distributed.

Consider the first rule. If the parameter interval is finite, then for obtaining the posterior distribution we may use the standard Bayes procedure.

In so doing, a prior distribution is not conjugated with the kernel of a likelihood function. If the interval for the parameter θ is infinite, we deal with improper prior p.d.f. The rule of Jeffrey's for the representation of the fact of ignorance of the parameter value should be

interpreted in this case as

$$h(\theta)\,d\theta \backsim d\theta, \quad -\infty < \theta < \infty, \tag{2.25}$$

that is, $h(\theta) \backsim \text{const}$. Thus, we have

$$\int_{-\infty}^{\infty} h(\theta)\,d\theta = \infty.$$

Jeffrey's proposes to use for the representation of a probability of a certain event instead of 1. Exactly this fact, in his opinion, allows us to obtain a formal representation of ignorance. For any two intervals (a,b) and (c,d) the relation

$$\frac{P\{a<\theta<b\}}{P\{c<\theta<d\}} = \frac{0}{0},$$

that is, represents indeterminacy, and thus we cannot make a statement about the chances of θ being in some finite pair of finite intervals.

The second rule of Jeffrey's touches upon the parameters whose nature lets us make an assumption on their having a value lying in the interval $[0,\infty)$ for example, a standard deviation. He proposes for such a parameter its logarithm having a uniform distribution, that is, if one puts $\vartheta = \log\theta$, then the prior d.d. for ϑ will be chosen in the form

$$h(\vartheta)\,d\vartheta \backsim d\vartheta, \quad -\infty < \vartheta < \infty. \tag{2.26}$$

Since $d\vartheta = d\theta/\theta$ (2.26) yields

$$h(\theta) \backsim \theta^{-1}, \quad 0 \leqslant \theta < \infty. \tag{2.27}$$

corresponding to the absence of information about the parameter θ.

An important property of (2.27) is its invariance with respect to the transformations $K = \theta^n$. Actually,

$$dK = n\,\theta^{n-1}d\theta \implies \frac{dK}{K} \backsim \frac{d\theta}{\theta}.$$

This property is very important because some research parameterizes the models in terms of the standard deviation σ, others in terms of a variance σ^2 or in terms of the parameter of exactness $\tau = \sigma^2$. It is easy to show that, if we choose the quantity $d\sigma/\sigma$ as the prior d.d. for σ, the relation will be

$$\frac{d\sigma}{\sigma} \backsim \frac{d\sigma^2}{\sigma^2} \backsim \frac{d\tau}{\tau},$$

a logical implication. This prior distribution is also improper, whence we may conclude that the relation $P\{0<\theta<a\}/P\{a<\theta<\infty\}$ is indeterminacy, that is, we can say nothing about the chances of the parameter being in the intervals $(0,a)$ and (a,∞). Indeterminacy similar to this one is considered again a formal representation of ignorance.

A question of representation of an improper prior p.d.f. is considered by Akaike [2–4]. He proposes the following interpretation: an improper prior p.d.f. can be represented in the form of a limit of a proper prior p.d.f. in such a way that the corresponding posterior p.d.f. converges point wisely to the posterior p.d.f., responding to an improper prior distribution. A mutual entropy is considered their nearness measure. It is shown also that the most appropriate choice is a choice of approximating eigenvalues, depending on a sample.

In spite of the fact that prior p.d.f.s, in accordance with the assumption of Jeffrey's are improper, corresponding to them the posterior distributions are proper and allow us to obtain the desired estimates. Let $(x_1, x_2, \ldots, x_n)$ be a sample from $N(\mu, \sigma)$ where μ and σ are unknown. If we have no prior information about μ and σ we may apply the principles of Jeffrey's and use as a prior d.d. a function of the form

$$h(\mu,\sigma)\,d\mu\,d\sigma \backsim \frac{1}{\sigma}\,d\mu\,d\sigma, \quad -\infty < \mu < \infty,\ 0 < \sigma < \infty.$$

Using the Bayes theorem, we can easily obtain the posterior d.d. That is,

$$\begin{aligned} \hbar(\mu,\sigma \mid x) &\backsim h(\mu,\sigma)\,\ell(\mu,\sigma \mid x) \\ &\backsim \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[vs^2 + n(\mu-\hat{\mu})^2\right]\right\}, \end{aligned} \tag{2.28}$$

where $\ell(\mu,\sigma \mid x) = \sigma^{-n}\exp\left[-\sum(x_i-\mu)^2/(2\sigma^2)\right]$ is a likelihood function, $v = n-1$, $\hat{\mu} = \frac{1}{n}\sum x_i$, $vs^2 = \sum(x_i-\hat{\mu})^2$. The posterior density we have written is proper. In the same manner, a marginal posterior d.d. of the parameter μ

$$\begin{aligned} \hbar(\mu \mid x) &= \int_0^\infty \hbar(\mu,\sigma \mid x)\,d\sigma \\ &\backsim \int_0^\infty \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[vs^2 + n(\mu-\hat{\mu})^2\right]\right\} d\sigma \\ &\backsim \left[vs^2 + n(\mu-\hat{\mu})^2\right]^{-\frac{v+1}{2}} \end{aligned} \tag{2.29}$$

is proper. As seen from (2.29), $\hbar(\mu \mid x)$ has a form of Student d.d. with a mean value $\hat{\mu}$. Analogously to $\hbar(\sigma \mid x)$,

$$\hbar(\sigma \mid x) \backsim \frac{1}{\sigma^{v+1}} \exp\left(-\frac{vs^2}{2\sigma^2}\right).$$

This posterior d.d. for σ has a form of the inverse gamma-distribution.

Some authors decide not to use improper d.d., preferring instead to introduce "locally-uniform" and "sloping" distribution densities. Box and Tico [28] propose the functions which are "sufficiently-sloping" in that domain where a likelihood function takes greater values. Outside of this domain, the form of the curve of a prior p.d.f. doesn't matter,

since, if one finds a kernel of the posterior p.d.f., he multiplies it by the small values of a likelihood function.

The most interesting and important peculiarity of the proposed improper prior p.d.f. is the property of invariance. Jeffrey's gives it an interesting interpretation. He proves the following statement. Suppose that a prior p.d.f. for the vector θ is chosen in the form

$$h(\theta) \backsim \left|\mathrm{Inf}_\theta\right|^{1/2}. \tag{2.30}$$

Here Inf_θ is a Fisher information matrix for the vector of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$, that is,

$$\mathrm{Inf}_\theta = -E_X\left[\frac{\partial^2 \log f(X \mid \theta)}{\partial\theta_i\,\partial\theta_j}\right],$$

where the mean value is taken over the random variable X. Then a prior p.d.f. of the form (2.30) will be invariant in the following sense. If a researcher parameterizes his model with the help of the component of the vector η, where $\eta = F(\theta)$, and F is single-valued differentiable transformation of the components of the vector θ, and chooses a prior p.d.f. for θ so that

$$h(\eta) \backsim |\mathrm{Inf}_\eta|^{1/2}.$$

then the posterior probability statements, obtained in this way, don't contradict the posterior statements obtained with the help of parameterization of the components of the vector θ and a prior p.d.f. of the form (2.30). The proof of this statement can be found in the book by Zelner [275]. Hartigan [101] develops the idea of representing a prior p.d.f. in the form (2.30) and formulates six properties of invariance. The property (2.30) is a particular case among them. Hardigan's interpretation of invariance is more common and includes the invariance relative to transformations of a sample space, repeated performance of samples, and contraction of a space for the parameter θ.

2.4.3 *Choice of a prior distribution with the help of information criteria*

This is the subject of investigation in many works devoted to the Bayes estimation. In this connection, [46, 60, 65, 108, 275] should be distinguished.

The approach proposed by Zelner [275] may be interpreted as follows. As information measure, contained in the p.d.f. of the observation $f(x \mid \theta)$ for a given θ, is used in the integral

$$I_x(\theta) = \int_\Omega f(x \mid \theta) \ln f(x \mid \theta)\, dx. \tag{2.31}$$

A priori mean information contents is defined as

$$\bar{I}_x = \int_\Theta I_x(\theta) h(\theta)\, d\theta.$$

If now, from a prior information contents $\bar{I}_x$ associated with the observation x, we subtract information contents of the prior information, then it is possible to represent a measure of information gained by

$$G = \bar{I}_x - \int_\Theta h(\theta) \ln h(\theta)\, d\theta.$$

Then it is assumed (in the situation of there being no exact information about a prior p.d.f.) to choose $h(\theta)$ from the maximization condition for G. Zelner calls such a function a prior p.d.f. with "minimal information". Now consider the following example. Suppose

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\theta)^2}{2}\right], \quad x \in (-\infty, \infty).$$

Then it is easy to obtain

$$I_x(\theta) = -\int_{-\infty}^{\infty} f(x \mid \theta) \ln f(x \mid \theta)\, d\theta = -\frac{1}{2}(\ln 2\pi + 1),$$

that is, $I_x(\theta)$ is independent of θ, hence for the proper $h(\theta)$

$$I_x = -\frac{1}{2}(\ln 2\pi + 1)$$

and

$$G = -\frac{1}{2}(\ln 2\pi + 1) - \int_\Theta h(\theta) \ln h(\theta)\, d\theta.$$

The value of G will be maximal if one minimizes a portion of the information contained in a prior distribution

$$I_\theta = \int_\Theta h(\theta) \ln h(\theta)\, d\theta$$

The solution of this problem is a uniform p.d.f. on Θ, that is, $h(\theta) \backsim \text{const}$. It should be noted that this result is in full accord with the form of a prior p.d.f. obtained using the rule of Jeffrey's [114].

If one considers (see Lindley [143]) a functional G in the asymptotical form

$$G_A = \int_\Theta h(\theta) \ln \sqrt{n|\mathrm{Inf}_\theta|}\, d\theta - \int_\Theta h(\theta) \ln h(\theta)\, d\theta,$$

where n is the number of independent samples from a general population, distributed by the probability law $f(x \mid \theta)$, and finding a prior p.d.f., $h(\theta)$ maximizing G_A under the condition $\int \theta\, h(\theta) d\theta = 1$, then we can obtain

$$h(\theta) \backsim |\mathrm{Inf}_\theta|^{1/2},$$

that is, a prior p.d.f. corresponding to the generalized rule of Jeffrey's, considered above, giving the invariant p.d.f. At the same time, as was shown by Zelner [275], if G is represented in the nonasymptotical form, then an invariant prior p.d.f. of Jeffrey's does not always appear to be a p.d.f. with "minimal information". In the case when a prior p.d.f. of Jeffrey's doesn't maximize G, its use makes us bring additional information into the analysis in contrast to the case when it uses prior information to maximize G. As can be seen from the above conclusions, the desire of Jeffrey's to ensure the property of invariance of the statistical deductions with respect to the parameter transformation deviates from the principle of "scantiness of knowledge". Convert [46], Deely, Tierney and Zimmer [60], Jaynes [113] investigate the question about the choice of a prior distribution with the minimization of the direct portion of information I_θ (Shennon), contained in a prior p.d.f. A rule of choice of $h(\theta)$ from the condition $I_\theta \longrightarrow \min$ is called an *entropy maximum principle* because the entropy $S_\theta = -I_\theta$ is used instead of I_θ. They introduce the term "*a least favorable distribution*". If $\mathscr{H}$ is a family of prior distributions, then $H \in \mathscr{H}$ is the least favorable distribution under the condition, that is, its corresponding minimum of the expected losses is greater than that one for the other elements of a family.

It should be noted that the estimate obtained in such a way coincides with a minimax one [257].

Deely, Tierney and Zimmer consider the use of a maximum entropy principle for the choice of a prior distribution in the binomial and exponential models. They show, in particular, that a least favorable distribution may be the best one in accordance with a maximum entropy principle.

Jaynes [113] modified this principle to a more general form. He introduces the measure

$$S_K = -\int_\Theta h(\theta) \ln \left[\frac{h(\theta)}{K(\theta)}\right] d\theta,$$

where K, as noted by El-Sayyad [73], is a suitable monotonic function. El-Sayyad proposes the use of a group theory approach for a choice of $K(\theta)$. Such an approach results in a change of the parameters of $h(\theta)$, which is very essential and doesn't change the entropy measure. For example, if θ_1 is a position parameter, θ_2 is a scale parameter ($\theta_2 > 0$), then the prior density is found so that $h(\theta_1, \theta_2) = ag(\theta_1 + b, a\theta_2)$, and the solution takes the form $h(\theta_1, \theta_2) \backsim \theta_2^{-1}$. Thus, we again have obtained the invariant prior p.d.f. of Jeffrey's.

For the binomial model, as shown by Jaynes [113], the use of the generalized maximum entropy principle gives the equation $\theta(1-\theta)h'(\theta) = (2\theta - 1)h(\theta)$, whence $h(\theta) \backsim \left[\theta(1-\theta)\right]^{-1}$.

Devjatirenkov [65] uses the information criterion

$$I(\hat{\theta}) = \int_{\Theta}\int_{\Omega} \hbar(\theta \mid x)p(x)\ln q(\theta \mid \hat{\theta})dxd\theta,$$

where $q(\theta \mid \hat{\theta})$ is the p.d.f. of the parameter with the given estimate, for the determination of the best (in the sense of minimum of I estimate $\hat{\theta}$. The estimate obtained in such a way appears to be less sensitive (in comparison with the usual one) with respect to the deviation of a prior distribution. It is interesting that a variance of the estimate for the Gaussian distribution attains the lower limit of the Cramer-Rao inequality.

In the work [46] information quality was used not as a criterion of a choice of a prior distribution, but as a method of indeterminacy elimination for the determination of the parameters of a prior distribution density. Arguments given in [46] are of an intuitive nature, but seem to be reasonable and may be used in practice. We give the main results of [46]. Suppose $f(x \mid \theta)$ is p.d.f. of observation x and $h(\theta;\gamma)$ is a prior p.d.f. with the parameter γ.

Analogous to a Fisher information $I_x(\theta)$ one introduces the so called *Bayes information*, contained in $h(\theta;\gamma)$:

$$BI_{\gamma}(\theta) = E_{\theta}\left[\left|\frac{\partial \ln h(\theta;\gamma)}{\partial\theta}\right|^2\right]. \tag{2.32}$$

Next one determines a weight significance w of a prior information with respect to an empirical one: $w = BI_{\gamma}(\theta)/I_x(\theta)$. Suppose now that θ_p is a prior estimate of the parameter θ. If the vector γ consists of two parameters γ_1 and γ_2, then we need to solve a system of two equations:

$$E_{\gamma}[\theta] = \theta_p,$$
$$BI_{\gamma}(\theta) = wI_x(\theta).$$

In the case when the number of parameters γ exceeds two, it is necessary to use additional arguments (see [46]). For example, for the binomial model

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and a prior p.d.f. beta $h(\theta) \backsim \theta^{a-1}(1-\theta)^{b-1}$, we have

$$I_x(\theta) = \frac{n}{\theta(1-\theta)}.$$

It also follows from (2.32),

$$BI_{\gamma}(\theta) = \frac{(a+b-4)(a+b-2)(a+b-1)}{(a-2)(b-2)}.$$

The system of equations for the determination of a and b has the form

$$\frac{a}{a+b} = \theta_p,$$
$$BI_\gamma(\theta) = wI_x(\hat{\theta}),$$

where θ_p and $\hat{\theta}$ is prior and empirical estimates of the parameter θ.

The methods for a choice of a prior distribution which are based on an information criterion fall outside the limits of the traditional Bayes approach and are drawn either to minimax or to empirical Bayes methods. In the works of some authors there are many efforts to construct fundamental theories directed to justification of a choice of prior distributions. We single out the works by Japanese statistician Akaike [2–4] who proposes the methods of effective use of Bayes models. The goal of constructions proposed by him is a change of the role of a prior distribution in the Bayes models. Akaike [2] proposes to use prior distributions adaptive to empirical data. They are called *modificators*.

In the problem of prediction of a density for the distribution $g(z)$ of future observations based on a sample of obtained data, which is being solved with the help of Bayes theory, a prior distribution (modificator) is chosen from the correspondence between the $g(z)$ and estimate $\hat{g}(z)$, expressed with the help of *Kulbak information measure*

$$B(g,\hat{g}) = -\int \ln[g(z)/\hat{g}(z)]g(z)\,dz.$$

In this capacity a mean value of entropy is used, that is $E_x[B(g(z), g(z \mid x))]$, where,

$$g(z \mid x) = \int_\Theta f(z \mid \theta)\hbar(\theta \mid x)d\theta \quad \text{and} \quad \hbar(\theta \mid x) \backsim f(x \mid \theta)h(\theta).$$

Thus, a prior p.d.f. is chosen by minimization of the mean value of entropy. It is interesting to note that Akaike's method gives the improper prior p.d.f. of Jeffrey's in some cases (in particular, in the problem of prediction of a distribution of the Gauss random vector).

Another interesting attempt to exclude the arbitrariness in the choice of a prior p.d.f. is proposed by Bernardo [20]. He recommends choosing standard prior and posterior distributions that describe a situation of "insignificant prior information", and deficient information is found from empirical data. His criterion of a choice of a prior p.d.f., $h(\theta)$ is constructed with the help of expected information about θ proposed by Lindley:

$$I^\theta\{\in, h(\theta)\} = \int_\Omega f(x)dx \int_\Theta \hbar(\theta \mid x) \ln \frac{\hbar(\theta \mid x)}{h(\theta)}\,d\theta,$$

where $\in$ denotes an experiment during which a random variable X with p.d.f., $f(x \mid \theta)$, $\theta \in \Theta$, is observed. It is also assumed that $h(\theta)$ belongs to the class of admissible prior distributions $\mathscr{H}$. The main idea is interpreted in the following way. Consider a random

variable $I^{\theta}\{\in (k), h(\theta)\}$ determining a portion of information about θ, expected in k recurrences of an experiment. We may achieve, by infinite recurrence, the exact value of θ. Thus, $I^{\theta}\{\in (\infty), h(\theta)\}$ measures a portion of deficient information about θ with a prior p.d.f., $h(\theta)$. The standard prior distribution $\pi(\theta)$, which corresponds to "indefinite prior knowledge", is defined as minimizing deficient information in the class $\mathscr{H}$. The standard posterior distribution after the observation x is defined with the help of Bayes theorem:

$$\pi(\theta \mid x) \backsim \pi(\theta) f(x \mid \theta).$$

Since the exact knowledge of a real number requires the knowledge of an infinite quantity of information, in the continuous case we obtain $I^{\theta}\{\in (\infty), h(\theta)\} = \infty$, for all $h(\theta) \in \mathscr{H}$. Standard posterior distributions for this case are defined with the help of a limit passing:

$$\pi(\theta \mid x) = \lim_{k \to \infty} \pi_k(\theta \mid x);$$

moreover, $\pi_k(\theta \mid x) \backsim \pi_k(\theta) f(x \mid \theta)$, and $\pi_k(\theta) = \arg\max I^{\theta}\{\in (k), h(\theta)\}$.
For the case of binomial trials a standard prior p.d.f., $\pi(\theta) \backsim \theta^{-1}(1-\theta)^{-1}$.
That is, we have the same result as the one in the work by Jaynes [113] obtained with the help of maximum of entropy.
The attractive feature of Bernardo theory is that this theory is free of some difficulties which are peculiar to a standard Bayes approach, in particular, the use of standard prior distributions doesn't give any marginal paradoxes (see the works by Stone and Dawid [243], and by Dawid, Stone and Zidek [56]) peculiar to non-informative prior distributions.
An interesting approach for a choice of prior distributions, based on geometrical probabilities, is proposed by Fellenberg and Pilz [79]. They consider a problem of the choice of a prior distribution in the estimation of a mean value of the time-to-failure for the exponential distribution with the cumulative distribution function $F(t;\lambda) = 1 - \exp(-\lambda t)$. For prior information a segment of uncertainty $[\lambda_1, \lambda_2]$ is used for the parameter A corresponding to the space of c.d.f. $\mathscr{B} = \{F(t;\lambda_1) \leqslant F(t;\lambda) \leqslant F(t;\lambda_2)\}$, consisting of the unknown cumulative distribution function $F(t;\lambda)$. It is assumed that a probability of λ being in $[\lambda_1, \lambda_2]$ is equal to unity. A prior p.d.f., $h(\lambda)$, is determined from the condition that a probability of a parameter λ getting into the interval $[\lambda_1, x]$ equals the probability of $F(t;\lambda)$ getting into the space $\mathscr{B}_x = \{F(t;\lambda_1) \leqslant F(t;\lambda) \leqslant F(t;\lambda_2)\}$. The equality of probabilities

$$P\{\lambda_1 \leqslant \lambda \leqslant x\} = P\{F(t;\lambda) \in \mathscr{B}_x\}$$

is valid because of monotonicity of $F(t;\lambda)$ on the parameter λ. The probability $P\{F(t;\lambda) \in \mathscr{B}_x\}$ is determined from geometrical reasoning (having used a principle of equal chances)

as a ration of the area contained between $F(t;\lambda_1)$ and $F(t;x)$. The resulting expression for the prior density of the parameter A has the form

$$h(\lambda) = \frac{\lambda_1\lambda_2}{\lambda_2 - \lambda_1} \cdot \frac{1}{\lambda_2} \quad \lambda_1 \leqslant \lambda \leqslant \lambda_2.$$

2.5 The general procedure of reliability estimation and the varieties of relevant problems

This section is connected conceptually with the preceding four paragraphs. However, there are no references for the ideas discussed in them. The reader who doesn't want to learn the Bayes approach from the formal mathematical positions may skip them and start with the consideration of the general problems of the Bayes theory of estimation discussed below.

2.5.1 *Reliability estimation*

The setting of a problem of reliability estimation consists of the following four elements:

a) *distribution of probabilities of the basic random variable* characterizing the reliability of a technical device or system (for example, a cumulative distribution function of time-to-failure $F(t;\theta)$, where θ is a vector of parameters;
b) *a prior probability distribution*, represented, for example, in the form of p.d.f., $h(\theta)$, of the vector of parameters θ, characterizing the uncertainty of the given prior information I_a about reliability;
c) *loss function*, $L(\hat{R},R)$ characterizing the losses involved when one replaces the reliability R by its estimate $\hat{R}$;
d) *testing plan*, II, prescribing the method of obtaining experimental data I_e.

The problem may be interpreted as follows: we need to find the estimates of reliability R by using a priori information I_a and experimental data I_e. We consider below two forms of representation of the estimate for the reliability

a) the set of the point estimate R and standard deviation (s.d.) which is a characteristic of exactness of the estimate $\hat{R}$;
b) confidence interval $\left[\underline{R},\overline{R}\right]_\gamma$ with a given confidence probability .

The principal scheme for obtaining Bayes estimates is represented in Fig. 2.1. It consists of three steps.

Step 1 *Composition of likelihood function* $\ell(\theta, I_e)$. To do this we use some statistical model describing the distribution of the basic random variable $F(t;\theta)$ and experimental data I_e obtained after a realization of a testing plan II

Step 2 *Construction of a posterior distribution* $h(\theta \mid I_a, I_e)$. Here we use the Bayes formula

$$h(\theta \mid I_a, I_e) = \frac{h(\theta \mid I_a)\ell(\theta \mid I_e)}{\int_\Theta h(\theta \mid I_a)\ell(\theta \mid I_e)d\theta}, \tag{2.33}$$

where Θ is a range of the parameter θ.

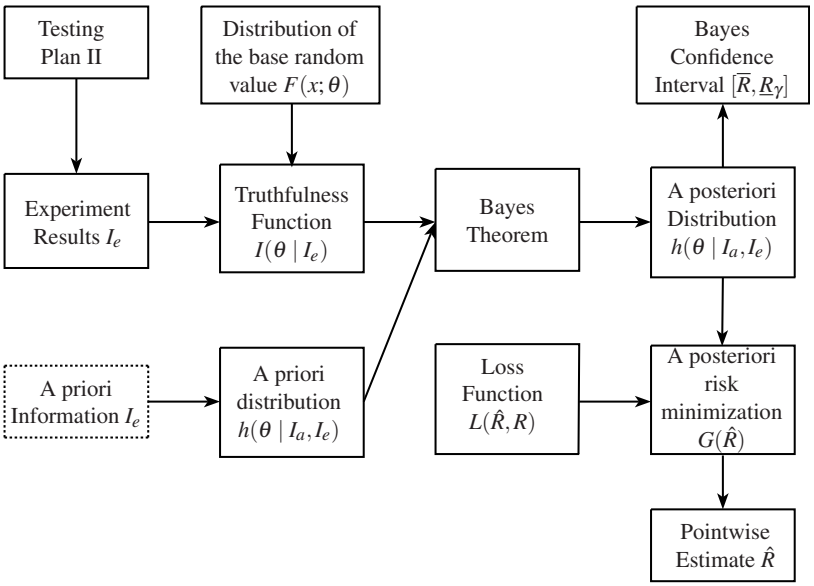


Fig. 2.1 The general scheme of obtaining the Bayes estimates

Step 3 *Obtaining Bayes estimates.* The Bayes confidence interval is defined by the condition $P\{\underline{R} \leqslant R \leqslant \overline{R}\} = \gamma$, or

$$\int_{\underline{R} \leqslant R(\theta) \leqslant \overline{R}} \hbar(\theta \mid I_a, I_e)d\theta = \gamma. \tag{2.34}$$

To obtain the Bayes point estimate $\hat{R}^*$, we should write the function of the posterior risk

$$G(\hat{R}) = \int_\Theta (\hat{R}, R(\theta))\hbar(\theta \mid I_a, I_e)d\theta \tag{2.35}$$

and choose among all estimates $\hat{R}$ such that one minimizes the function (2.35), that is,

$$\hat{R}^* = \arg \min_{\hat{R}\in[0,1]} G(\hat{R}). \tag{2.36}$$

If a squared-error loss function is chosen of the form $L(\hat{R},R) = (R-\hat{R})^2$, then the Bayes estimate is determined in the form of the posterior mean value

$$\hat{R}^* = \int_\Theta R(\theta)\hbar(\theta \mid I_a, I_e)d\theta. \tag{2.37}$$

The error of estimation of the value of $\hat{R}^*$ is assessed by the posterior s.d., $\sigma_{\hat{R}^*}$, which satisfies the relation

$$\sigma^2_{\hat{R}^*} = \int_\Theta R^2(\theta)\hbar(\theta \mid I_a, I_e)d\theta - \hat{R}^{*2}. \tag{2.38}$$

Example 2.1. The device time-to-failure is subjected to the exponential distribution with the probability density

$$f(t;\lambda) = \lambda\, \mathrm{e}^{-\lambda t}, \quad t \geqslant 0, \;\; \lambda \geqslant 0. \tag{2.39}$$

As a prior distribution, a gamma-distribution with the likelihood kernel [202] has been chosen. The distribution density of the parameter λ has the form

$$h(\lambda \mid I_a) = \frac{\rho^\delta \lambda^{\delta-1}\mathrm{e}^{-\rho\lambda}}{\Gamma(\delta)}, \quad \lambda \geqslant 0, \;\; \delta \geqslant 0, \;\; \rho \geqslant 0, \tag{2.40}$$

and the parameters δ and ρ are assumed to be known. A mean square-error loss function, $L(\hat{R},R) = (R-\hat{R})^2$. As a testing plan $[n,U,T]$ (see [91]) has been chosen. We shall discuss the solution of the problem in detail including finding the estimates of the failure rate λ. The solution will be given in the form of the three steps:

1) Composing a likelihood function we bear in mind the fact that as a result of testing by the plan $[n,U,T]$ we have observed a censored sample.

   Suppose, for definiteness, that d of the tests have ended by failures at the instants $t_1^*, t_2^*, \ldots, t_d^*$ and that $n-d$ of those remaining were interrupted before failure after T units of time from the beginning. The likelihood function describing this situation is chosen (see [91]) in the form

$$\ell(\lambda \mid I_e) = \prod_{i=1}^{d} f(t_i^*;\lambda) \prod_{j=1}^{n-d} \int_T^\infty f(x;\lambda)\, dx. \tag{2.41}$$

   Substitution of p.d.f. (2.39) into (2.41) and simplifying, we have

$$\ell(\lambda \mid I_e) = \prod_{i=1}^{d} \lambda\, \mathrm{e}^{-\lambda t_i^*} \left(\mathrm{e}^{-\lambda T}\right)^{n-d} = \lambda^d \mathrm{e}^{\lambda K}, \tag{2.42}$$

   where $K = t_1^* + t_2^* + \cdots + t_d^* + (n-d)T$. Thus, the sufficient Bayes testing statistics appears to be as a pair of numbers (d,K).

2) Substituting the likelihood function (2.42) and a prior p.d.f. (2.40) in the Bayes formula (2.33), we have the posterior distribution density

$$\hbar(\lambda \mid I_a, I_e) = \frac{(\rho+K)^{\delta+d}}{\Gamma(\delta+d)} \lambda^{d+\delta+1} \mathrm{e}^{-(\rho+K)}. \tag{2.43}$$

The obtained expression is a gamma-function. Consequently, the chosen prior distribution is conjugated.

3) Determine the point estimate $\hat{\lambda}^*$. For the squared-error loss function, the minimum of the posterior risk $G(\hat{\lambda})$ is attained at the point of the posterior mean value of the estimated parameter, that is,

$$\hat{\lambda}_* = \int_0^\infty \lambda \hbar(\lambda \mid I_a, I_e) d\lambda = \frac{d+\delta}{\rho+K}.$$

The posterior s.d., which may be interpreted as exactness characteristic of the estimate λ^*, is defined by the variance of the gamma-distributed random variable, that is,

$$\sigma_{\hat{\lambda}^*} = \frac{\sqrt{d+\delta}}{\rho+K}.$$

An interval estimate of the failure-rate is often used as the upper λ confidence limit. Define the Bayes analog of this estimate $\bar{\lambda}_\gamma$ and putting $\underline{\lambda}_\gamma = 0$. By definition, $P\{\lambda \leqslant \bar{\lambda}_\gamma\}$ and since the posterior distribution is a gamma-distribution, the transformation $z = 2\lambda(\rho+K)$ gives us the chi-square distribution with $2(\delta+d)$ degrees of freedom, we have

$$P\{2\lambda(\rho+K) \leqslant X_{\gamma;2(\delta+d)}\} = \gamma, \tag{2.44}$$

where $X_{\gamma;2(\delta+d)}$ is a quantile of the chi-square distribution of the probability . From this relation (2.44) we finally obtain

$$\bar{\lambda}_\gamma = \frac{X_{\gamma;2(\delta+d)}}{2(\rho+K)}.$$

2.5.2 *Varieties of problems of Bayes estimation*

When one solves practical problems, it is very difficult to establish all the above-mentioned elements for Bayesian reliability analysis and modeling. Therefore, we need to improve the classification of all possible varieties. From the methodological point-of-view, the most essential elements can be classified as follows:

1) the forming of a reliability model;
2) completeness of information on a prior distribution (or completeness of a prior uncertainty);

3) completeness of information on the main random variable.

We consider, in brief, each of these classifications. From the point-of-view of the first characteristic, we will distinguish two types of reliability models, giving the base for a corresponding approach for reliability determination. The first one (which is used more often) we will name "formal" and write it in the form

$$\xi > t_{\text{req}} \tag{2.45}$$

where ξ is a random time to-failure, t_{req} is the time which is necessary, in accordance with technological assignment of the device or system functioning. The model (2.45) doesn't touch upon real technical peculiarities of the device or system and enables us to analyze different (by nature) devices in a standard way, based only on the observable properties of the device (results of trials).

In addition to this model, we will use a model which is based on the mathematical description of real processes of the device collapse. This model, represented in the form

$$Z_j(t) = \phi(X(t)) > 0, \quad t \geqslant t_{\text{req}}, \quad j = 1, 2, \ldots, m, \tag{2.46}$$

will be named "functional". $Z_j(t)$ denotes a random process of the initial device variables (loading factors, physically-mechanical characteristics, geometrical parameters, etc.); the function $\phi(\cdot)$ is called the *survival function*.

The set of conditions (2.46) symbolizes the time of the device being operable. There are many works devoted to the investigation of reliability of technical devices in the framework of this modeling which don't use the Bayes approach [26, 32, 89, 135, 173, and 260].

The questions connected with using the formal models of the device reliability will be discussed in the following five chapters, from the third to the seventh one, inclusive. Chapters 8–10 are devoted to the methods in reliability analysis based on the functional models.

From the point-of-view of a prior uncertainty (the second classification property) we will discuss the following cases:

(A1) The case of a complete prior determinacy, when the prior distribution density is determined uniquely;

(A2) The case of an incomplete priori determinacy, when a prior density is not given, and is determined only by finite number of restrictions, imposed on some functions defined on a prior distribution (for example, only a mean prior value of the parameter θ may be given);

(A3) The case of a complete a priori indeterminacy, when only the finite number of estimates of the parameter θ are known.

The first case is the most frequently encountered and will be discussed in Chapters 3, 4, 5, 8, 9. The case (A2) is the least studied in Bayes literature.

In Chapter 6 we give the description of the general formal representation of a partial a priori information and solve many concrete problems. The case (A3) is known as empirical Bayes estimation and is discussed in Chapter 7.

With respect to the third property, we will discuss the following two cases:

(C1) Parametric, when a parametric family is given for the cumulative distribution function of the main random variable, that is, $F(t) = F(t;\theta)$, where $\theta \in \Theta$;

(C2) Nonparametric, when this cumulative distribution function is determined on some nonparametric class (for example, a class of all continuous cumulative distribution functions).

Furthermore, we will use, as a rule, parametric estimates, since they are simpler and broadly applied. Nonparametric Bayes estimates will be studied in Chapter 4 and partially in Chapter 7. In Chapter 5 we will use the so called quasiparametric estimates. They use different methods for an approximate solution of a problem of estimation which is set as a parametric one, but are solved by means of nonparametric methods.



http://www.springer.com/978-94-91216-13-8

Bayesian Theory and Methods with Applications

Savchuk, V.; Tsokos, C.P.

2011, XIV, 318 p., Hardcover
ISBN: 978-94-91216-13-8
A product of Atlantis Press