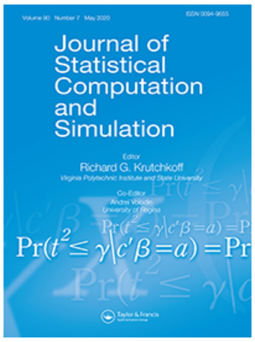


Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Estimation of mean value of a normal distribution with constraints on the relative error and *d*-risk

Rustem Salimov, Su-Fen Yang, Andrei Volodin & Igor Volodin

To cite this article: Rustem Salimov, Su-Fen Yang, Andrei Volodin & Igor Volodin (2020) Estimation of mean value of a normal distribution with constraints on the relative error and *d*-risk, Journal of Statistical Computation and Simulation, 90:7, 1286-1300, DOI: 10.1080/00949655.2020.1724292

To link to this article: https://doi.org/10.1080/00949655.2020.1724292

Published online: 07 Feb 2020.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2020, VOL. 90, NO. 7, 1286–1300
https://doi.org/10.1080/00949655.2020.1724292

Estimation of mean value of a normal distribution with constraints on the relative error and *d*-risk

Rustem Salimov[a], Su-Fen Yang[b], Andrei Volodin[c] and Igor Volodin[a]

[a]Department of Mathematical Statistics, Kazan Federal University, Kazan, Russian Federation; [b]Department of Statistics, National Chengchi University, Taipei City, Taiwan, Republic of China; [c]Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada

ABSTRACT

We consider a problem of mean value estimation for a normal distribution with prior knowledge of its randomness and extreme smallness. We continue our investigation presented in article [Volodin et al. Estimation of the mean value for the normal distribution with constrains on d-risk. Lobachevskii J Math. 2018;39:377–387], where the constraints have been made on the absolute estimation error. It is more appropriate to control its relative estimation error, not the absolute. In this article, we consider not only estimates based on a fixed number of observations, but also the sequential procedure of estimation. Both estimators guarantee the given constraints on their *d*-risks (the so-called *d*-guarantee procedure). We present a simple method for calculating the minimal sample size that guarantees the given constraints on the *d*-risk of the relative error when the estimators with uniformly minimal *d*-risk and Bayesian are applied. A sequential guarantee estimation procedure is also proposed, and the distribution of the corresponding stopping time is illustrated by the results of statistical simulations. As a practical application of the proposed statistical procedures, the problem of estimating the concentration of arsenic in drinking water is considered.

ARTICLE HISTORY
Received 26 September 2019
Accepted 28 January 2020

KEYWORDS
Parameter estimation; normal mean; *d*-posterior approach; exponential prior distribution; sequential estimation; estimate with the minimal *d*-risk; minimal sample size; empirical Bayesian approach

1. Introduction

We consider a typical statistical Quality Control problem of percentage estimation for some substance (contaminant or pollutant) in a product. The amount of this contaminant should be within acceptable limits. If the production process is stable (under control), then, from the Mathematical Statistics' point of view, we are dealing with a sequence of statistical experiments. In each of the experiments, the value of the corresponding parameter θ is estimated. This parameter can be considered as an index for a family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ of possible distributions of the random sample X on a measurable sample space $(\mathcal{X}, \mathcal{A})$. It is assumed that the corresponding sequence $\theta_1, \theta_2, \ldots$ of unknown (estimated) values of the parameter θ in each statistical experiment can be interpreted as a value of the sequence $\vartheta_1, \vartheta_2, \ldots$ of independent identically distributed random variables with distribution G_λ

CONTACT Su-Fen Yang ✉ yang@mail2.nccu.tw, health887459@gmail.com 📇 Department of Statistics, National Chengchi University, No.64, Sec. 2, Zhinan Rd., Wenshan Dist., Taipei City, Taiwan, Republic of China

that belongs to a family $\mathcal{G} = \{G_\lambda, \lambda \in \Lambda\}$ of so-called prior distributions of the random parameter ϑ.

It is natural to solve this statistical problem of θ estimation in the framework of the Bayesian paradigm. Let $L(\theta, d)$, $(\theta, d) \in \Theta^2$ be a loss function. It indicates some costs for a statistician in the case where a decision d is accepted when the true value of θ is different from d. The task of setting the risk function (the mean value of losses) is crucial for a solution of the statistical inference guarantee problem. Our new approach for an estimation of the mean value of normal distribution is based on the Bayesian paradigm that was initially presented in article [1]. This is the so-called d-posterior approach to the problem of statistical inference guarantee. Contrary to the classical approach, here the value of mean losses under wrong decisions is calculated from experiments that have ended in the acceptance of the same decision d, instead of from statistical experiments with the same value of the parameter of interest θ. Therefore, the solution of the guarantee problem is based on the calculation of the d-risk function: the conditional mathematical expectation of the loss function with respect to the σ-algebra generated by the decision function $\delta(X)$. Note that in our case $\delta(X)$ is the estimate of the parameter itself.

If we assume that the prior distribution G is known, or we have its estimate from the results of a real sequence of statistical experiments, then the prior risk of the statistical procedure $\mathbf{R}_G = \mathbf{E}L(\vartheta, \delta(X))$ (the mean value is calculated by the joint distribution of X and ϑ) is too rough as a characteristic of the quality for the estimate $\delta(X)$. The natural analogue of the risk function $R(\theta) = \mathbf{E}_\theta L(\theta, \delta(X))$ for a solution of the guarantee problem of statistical inference is the d-risk function

$$\mathfrak{R}(d) = \mathbf{E}\{L(\vartheta, d) \mid \delta(X) = d\}, \quad d \in \Theta.$$

The function $\mathfrak{R}(d)$ can be interpreted as the mean value of losses from the statistical experiment that resulted in the acceptance of the same decision $d \in \mathcal{D}$.

Therefore the guarantee estimation procedure is based not on the probability of deviation of the estimate $\delta(X)$ from the true (unknown) value of the parameter θ, but on the probability that for the completed observations, the value of the random parameter ϑ is different from the decision made $d = \delta(x)$ (estimate of θ by the result x of random sample X).

Let Δ be a constant that we use in order to control the relative error of our estimate. In this article, the estimation problems of the mean value θ of a normal (θ, σ^2) distribution is solved with the prior exponential distribution

$$G(\theta) = 1 - \exp\{-\lambda(\theta - \theta_0)\}, \quad \theta > \theta_0, \ \lambda > 0$$

of the random parameter ϑ and loss function $L(\theta, d) = 0$, if

$$\frac{1}{1+\Delta} < \frac{\vartheta}{d} < 1 + \Delta,$$

and $L(\theta, d) = 1$ otherwise. Such a loss function shows how many times the decision made is different from the true value of the parameter; that is, the loss function corresponds to the relative error for θ estimation. A similar estimate has been considered in [2] with a somewhat different loss function. The minimax estimation for θ has been constructed, and a formula for the minimal sample size is derived under the guarantee constraint on the classical risk of the estimate.

Therefore, contrary to similar investigations presented in article [3], in the current investigation we assume that the true value of θ cannot be zero (the estimated contaminant is always present in the product). If we assume that $\theta_0 = 0$, then the d-risk of the Bayesian estimate becomes unbounded, which is typical for the problem of the guarantee estimation with constraints on the relative error (in connection with this see articles [2,4]).

The article is organized in the following way. In the second section, a formula for the posterior reliability of the estimate is derived. This is a more convenient approach than the posterior risk for constructing an estimation procedure. The Bayesian estimate of θ is also presented, and the algorithm for the estimate with a uniformly minimal d-risk is established (see articles [5,6] for methods of constructing such estimates). It is essential in our case that such an estimate possesses the d-minimax property. Consequently, it gives the decision rule corresponding to the minimal sample size for which there is an estimation procedure of θ with given constraints on the relative error and d-risk (for the proof of such statements for different risk definitions of a statistical inference procedure we refer to article [4]). The third section is devoted to the derivation of formulae for d-risks of the Bayesian estimate, and the estimate with uniformly minimal d-risk. In the fourth section, we suggest methods of calculating the minimal sample size required to obtain an estimate with guarantee constraints on the absolute error and d-risk. In Section 5, a sequential procedure for the estimation of θ value is constructed and its properties are investigated by the method of statistical simulation. Section 6 is devoted to the empirical estimate of prior distribution parameters. In the same section, as an example of the θ estimation sequential procedure, we consider the problem of estimating the concentration of arsenic in drinking water. Parameters of the model are chosen in accordance with the State regulations for laboratory testing of this ecologically important characteristic of drinking water. All formulae obtained for characteristics of the estimation procedures are illustrated numerically and graphically on exactly these parameters of the probability model.

2. Bayesian and uniformly minimizing d-risk estimates of parameter θ

The probability model we are considering consists of a normal (θ, σ^2) distribution and an exponential prior distribution with the rate parameter λ and shift parameter θ_0. For this model, the distribution family of the random sample $X^{(n)} = (X_1, \dots, X_n)$ possesses the sufficient statistic

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

Therefore, the statistical model is reduced to the normal $(\theta, \sigma^2/n)$ distribution with the density function

$$p(x \mid \theta) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n}{2\sigma^2}(x-\theta)^2\right\}, \quad \theta \in \mathrm{R}, \ \ x \in \mathrm{R}.$$

Prior exponential distribution is defined by the density function

$$g(\theta) = \lambda \exp\{-\lambda(\theta - \theta_0)\}, \quad \theta \geq \theta_0.$$

The problem we are interested in solving consists of constructing a procedure $(\nu, \widehat{\theta_\nu})$ for parameter θ estimation with the stopping time ν and decision function (estimate) $\widehat{\theta_\nu}$. The procedure should guarantee the given constraint on the relative error and d-risk of the estimate:

$$\sup_{d \geq \theta_0} P\left((1+\Delta)^{-1} \leq \frac{\vartheta}{d} < 1 + \Delta \mid \widehat{\theta_\nu} = d\right) \geq 1 - \beta,$$

where $\beta \in (0,1)$ is a given constant. It is assumed that the variance σ^2 and parameter λ are known, or are estimated by the data archive in the framework of the empirical Bayesian approach.

For the derivation of the posterior distribution, we note that the joint distribution of $\overline{X}$ and ϑ has the density function

$$u(x,\theta) = p(x \mid \theta) g(\theta) = \frac{\lambda\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{n}{2\sigma^2}(x-\theta)^2 - \lambda(\theta - \theta_0)\right\},$$

where $x \in \mathrm{R}$, $\theta \geq \theta_0$.

The posterior distribution is defined as the conditional distribution of the random parameter θ with respect to the sufficient statistic $\overline{X}$; that is, it has the density function

$$h(\theta \mid \overline{X}) = \frac{p(\overline{X} \mid \theta) g(\theta)}{\int_0^\infty p(\overline{X} \mid u) g(u)\, \mathrm{d}u}.$$

In the framework of the declared probability model, the posterior density

$$h(\theta \mid \overline{X}) = \frac{\exp\{-n(\overline{X}-\theta)^2/(2\sigma^2) - \lambda(\theta-\theta_0)\}}{\int_0^\infty \exp\{-n(\overline{X}-u)^2/(2\sigma^2) - \lambda(u-\theta_0)\}\, \mathrm{d}u}.$$

The marginal (unconditional) density function of the sample mean

$$\begin{aligned} p_G(x) &= \int_{\theta_0}^\infty p(x \mid \theta) g(\theta) d\theta \\ &= \lambda \exp\left\{-\frac{n}{2\sigma^2}(x^2 - t^2) + \lambda\theta_0\right\} \cdot \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int_{\theta_0}^\infty \exp\left\{-\frac{n}{2\sigma^2}(\theta - t)^2\right\} \mathrm{d}\theta \\ &= \lambda \exp\left\{-\frac{n}{2\sigma^2}(x^2 - t^2) + \lambda\theta_0\right\} \left[1 - \Phi\left((\theta_0 - t)\frac{\sqrt{n}}{\sigma}\right)\right], \quad x \in \mathbf{R}, \end{aligned}$$

where $t = x + \lambda\sigma^2/n$.

The density function of the posterior distribution ϑ is

$$h(\theta \mid \overline{X}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma\Phi\left((T-\theta_0)\sqrt{n}/\sigma\right)} \exp\left\{-\frac{n}{2\sigma^2}(\theta_0 - T)^2\right\}, \quad \theta > \theta_0,$$

where $T = \overline{X} + \lambda\sigma^2/n$ and $\Phi(\cdot)$ is the standard normal distribution function. Hence, the posterior distribution of the random parameter ϑ is the normal $(T, \sigma^2/n)$ distribution truncated at the point θ_0.

Proposition 2.1: *The Bayesian estimate of θ takes the form*

$$\hat{\theta}_n = \hat{\theta}_n(T) = \max\{\theta_0(1+\Delta), \hat{a}_n\},$$

where

$$\hat{a}_n = \hat{a}_n(T) = \frac{T}{(1+\Delta)+(1+\Delta)^{-1}} + \sqrt{\frac{T^2}{[(1+\Delta)+(1+\Delta)^{-1}]^2} + \frac{4\sigma^2 \ln(1+\Delta)}{n[(1+\Delta)^2-(1+\Delta)^{-2}]}}. \tag{1}$$

Proof: Construction of the Bayesian estimate is based on the maximization of the posterior reliability function

$$H(a) = P\left((1+\Delta)^{-1} \le \frac{\vartheta}{d} < 1+\Delta \mid \overline{X}\right)$$

as a function of some decision $d = a$. Let

$$u_1(a) = \max\{\theta_0, a/(1+\Delta)\}, \quad u_2(a) = \max\{\theta_0, a(1+\Delta)\}, \tag{2}$$

and write its integral representation

$$\begin{aligned} H(a) &= \int_{u_1(a)}^{u_2(a)} h(\theta \mid \overline{X})\, \mathrm{d}\theta \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma \Phi\left((\theta_0 - T)\sqrt{n}/\sigma\right)} \int_{u_1(a)}^{u_2(a)} \exp\left\{-(\theta - T)^2\right\} \mathrm{d}\theta \\ &= \frac{\Phi\left[(u_2(a) - T)\sqrt{n}/\sigma\right] - \Phi\left[(u_1(a) - T)\sqrt{n}/\sigma\right]}{\Phi\left((T-\theta_0)\sqrt{n}/\sigma\right)}. \end{aligned} \tag{3}$$

For the Bayesian estimate, it is necessary to find the point where the maximum of the posterior reliability function is achieved by a, or, which is the same, for the function

$$D(a) = \int_{u_1(a)}^{u_2(a)} \exp\left\{-(\theta - T)^2\right\} \mathrm{d}\theta, \quad a > \theta_0.$$

Finding the point of the maximum of the posterior reliability is fulfilled by its maximization in the following three regions of a-values.

If $a/(1+\Delta) < a(1+\Delta) < \theta_0$, then $D(a)$ is obviously zero.

If $a/(1+\Delta) < \theta_0 < a(1+\Delta)$, then

$$D(a) = \int_{\theta_0}^{a(1+\Delta)} \exp\left\{-\frac{n}{2\sigma^2}(\theta - T)^2\right\} \mathrm{d}\theta,$$

and, in this case, the maximum of $D(a)$ is achieved at the point

$$\hat{a} = \max\{a : a/(1+\Delta) < \theta_0 < a(1+\Delta)\};$$

that is, $a = \theta_0(1+\Delta)$. Note that $\hat{a}$ can be the Bayesian estimate if T is sufficiently small (as will be shown in what follows, this happens for $T < \theta_0(1+\Delta)$).

If, however $\theta_0 < a(1+\Delta)^{-1} < a(1+\Delta)$, then the point of the maximum of the reliability function is found by the method of differential calculus. Differentiating the function $D(a)$ by a and equating the expression obtained to zero:

$$\frac{\mathrm{d}D(a)}{\mathrm{d}a} = (1+\Delta)\exp\left\{-\frac{n}{2\sigma^2}(a(1+\Delta)-T)^2\right\} - \frac{1}{1+\Delta}\exp\left\{-\frac{n}{2\sigma^2}\left(\frac{a}{1+\Delta}-T\right)^2\right\} = 0.$$

Elementary calculations reduce this equation to the form

$$a^2\left[(1+\Delta)^2 - \frac{1}{(1+\Delta)^2}\right] - 2aT\left[(1+\Delta) - \frac{1}{1+\Delta}\right] - \frac{4\sigma^2\ln(1+\Delta)}{n} = 0.$$

Its solution $\hat{a}_n = \hat{a}_n(T)$ has the form (1). We choose the root of the quadratic equation with 'plus' because $a > \theta_0(1+\Delta) > 0$); and therefore, it gives the desired Bayesian estimate of θ. The proposition is proved. ■

Concerning the estimate with the uniformly minimal d-risk (the methods of constructing such estimates are discussed in article [5]), it is impossible to write it in the closed form, and we present only the algorithm of its numerical realization.

For this, consider the posterior reliability as a function of the value t of the statistic $T = T(\overline{X})$ and some decision a:

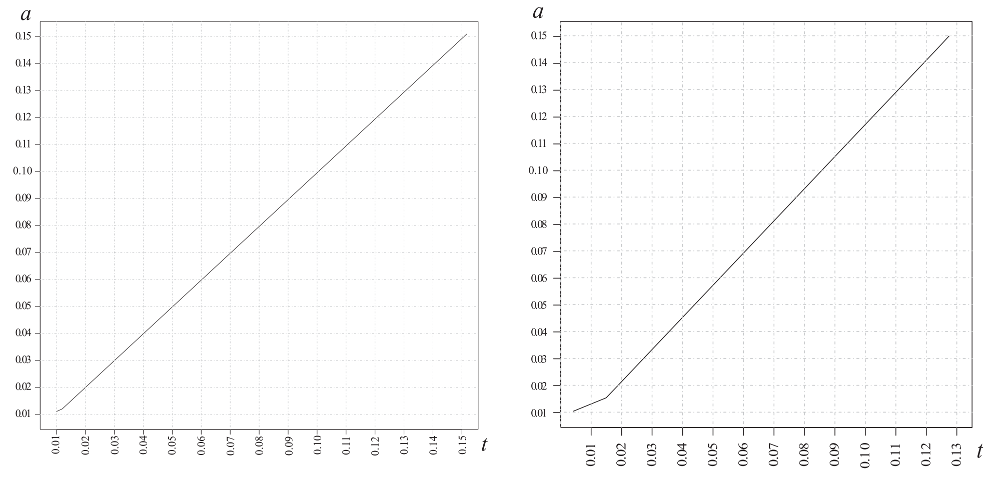
$$H(a,t) = \frac{\Phi\left[(u_2(a)-t)\sqrt{n}/\sigma\right] - \Phi\left[u_1(a)-t)\sqrt{n}/\sigma\right]}{\Phi\left((t-\theta_0)\sqrt{n}/\sigma\right)}.$$

For each fixed value of a find the point $t(a)$ of achieving the maximum by the variable t of function $H(a,t)$ and present the graph of the inverse function $a = a(t),\ t \in \mathrm{R}$.

The statistic $\theta_n^*(\overline{X}) = a(T(\overline{X}))$ serves as the estimate of θ with the uniformly minimal d-risk. Therefore, if $t = t(\overline{x})$ is the observed value of the statistic $T(\overline{X})$, then the observed value of the estimate $\theta_n^*(\overline{X})$ equals $a(t(\overline{x}))$. It is easy to see that for each fixed value of a, the function $H(a,t)$ has unique maximin and the function $a = a(t),\ t \in \mathrm{R}$, monotonically increases by t. Figure 1 provides graphs of the function $a(\cdot)$ for the values of the parameters from the probability model for the example in Section 6. These graphical illustrations and some additional numerical examinations of the function $a(\cdot)$ show that for sufficiently small values of Δ, the function is well approximated by the linear relation $a(t) = t$.

3. The *d*-risk function for the Bayesian and uniformity minimal *d*-risk estimates of parameter *θ*

The d-risk functions of the estimates can be obtained relatively simply by the substitution of these estimates into the posterior reliability $H(a)$.



(a) (b)

Figure 1. Graphs for the construction of the uniformly minimal *d*-risk estimates. (a) $\Delta = 0.1$, $n = 2$. (b) $\Delta = 0.2$, $n = 2$.

Proposition 3.1: *The d-risk function of the Bayesian estimate is defined for all* $d \geq \theta_0 (1 + \Delta)$ *and can be written as*

$$\mathfrak{R}_B(d) = 1 - \left[\Phi\left(\left(\frac{d^2 - U^2/n}{2dV} - \theta_0\right)\frac{\sqrt{n}}{\sigma}\right)\right]^{-1} \times \left[\Phi\left(\left((1+\Delta)d - \frac{d^2 - U^2/n}{2dV}\right)\frac{\sqrt{n}}{\sigma}\right) - \Phi\left(\left(\frac{d}{1+\Delta} - \frac{d^2 - U^2/n}{2dV}\right)\frac{\sqrt{n}}{\sigma}\right)\right],$$

where

$$V = [(1+\Delta) + (1+\Delta)^{-1}]^{-1}, \quad U^2 = \frac{4\sigma^2 \ln(1+\Delta)}{(1+\Delta)^2 - (1+\Delta)^{-2}}.$$

The d-risk function of the estimate with the uniformly minimal d-risk is defined for all values $d \in \mathrm{R}$ *and can be written as*

$$\mathfrak{R}_M(d) = 1 - \frac{\Phi\left[(u_2(d) - t(d))\sqrt{n}/\sigma\right] - \Phi\left[(u_1(d) - t(d))\sqrt{n}/\sigma\right]}{\Phi\left((t(d) - \theta_0)\sqrt{n}/\sigma\right)}.$$

Proof: The d-risk function for the Bayesian estimate $\hat{\theta}_n(T) = \max\{\theta_0(1+\Delta), \hat{a}_n\}$ is obtained by its substitution into the expression of the posterior reliability

$$\begin{aligned}\mathrm{H}_n(T) &= \max_{a \in \mathrm{R}} H(a)\\ &= \left[\Phi\left(\frac{A - T}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{B - T}{\sigma}\sqrt{n}\right)\right] \times \left[\Phi\left(\frac{(T - \theta_0)}{\sigma}\sqrt{n}\right)\right]^{-1},\end{aligned}$$

where

$$\begin{aligned}A &= \max\{\theta_0, (1+\Delta)\max\{\theta_0(1+\Delta), \hat{a}_n\}\},\\ B &= \max\{\theta_0, (1+\Delta)^{-1}\max\{\theta_0(1+\Delta), \hat{a}_n\}\}.\end{aligned}$$

Note that for $\hat{a}_n \le \theta_0(1+\Delta)$ the expression in the first square brackets has the form

$$\Phi\left(\frac{\theta_0(1+\Delta)^2 - T}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{\theta_0 - T}{\sigma}\sqrt{n}\right),$$

and for $\hat{a}_n \ge \theta_0(1+\Delta)$

$$\Phi\left(\frac{(1+\Delta)\hat{a}_n - T}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{\hat{a}_n/)1+\Delta) - T}{\sigma}\sqrt{n}\right).$$

Hence, the posterior reliability of the Bayesian estimate is defined only for the values of T that satisfy the inequality $\hat{a}_n \ge \theta_0(1+\Delta)$ and are equal to

$$\mathrm{H}_n(T) = \left[\Phi\left(\frac{(1+\Delta)\hat{a}_n - T}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{\hat{a}_n/(1+\Delta) - T}{\sigma}\sqrt{n}\right)\right] \cdot \left[\Phi\left(\frac{(T-\theta_0)}{\sigma}\sqrt{n}\right)\right]^{-1}.$$

From the obtained formula of the posterior reliability $\mathrm{H}_n(T)$ of the Bayesian estimate $\hat{\theta}_n(T) = \max\{\theta_0(1+\Delta), \hat{a}_n\}$, we derive the desired formula for the d-risk

$$\mathfrak{R}_B(d) = E\left\{1 - \mathrm{H}_n(T \mid \hat{\theta}_n(T) = d\right\}$$

of this estimate, which is true for all $d \ge \theta_0(1+\Delta)$.

Really, since $\hat{a}_n(T)$ is a monotonically increasing function of statistic T, then the d-risk is defined by the substitution of $d = \hat{a}_n(T)$ and the change of variable T in $\bar{H}_n(T)$ by the root $T(d)$ of the equation

$$\hat{a}_n(T) = VT + \sqrt{V^2T^2 + U^2/n} = d.$$

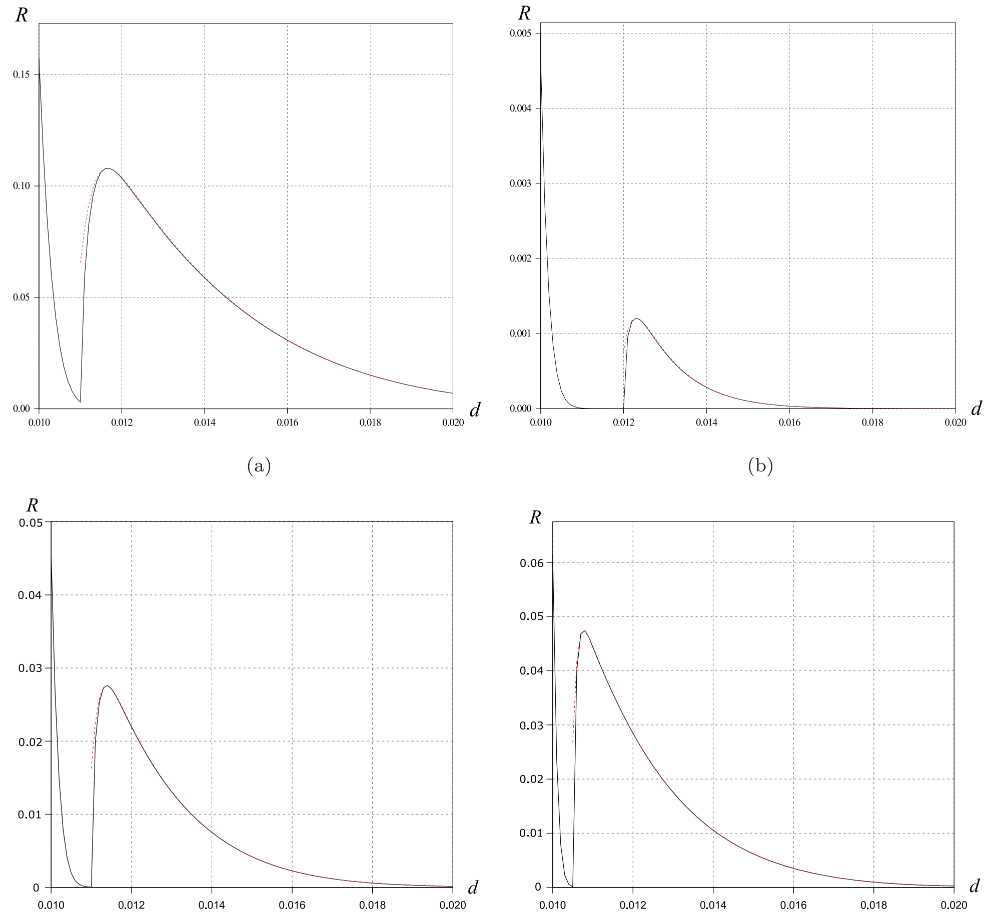
This equation is equivalent to $V^2T^2 - 2dVT + d^2 = V^2T^2 + U^2/n$, hence

$$T = T(d) = \frac{d^2 - U^2/n}{2dV}.$$

To sum up, the substitutions of $\hat{a}_n = d,\ T = T(d)$ into $1 - \bar{H}_n(T)$ provide the desired expression for the d-risk of the Bayesian estimate, which is true for all $d \ge \theta_0(1+\Delta)$.

The formula for the function $\mathfrak{R}_M(d),\ d \in \mathrm{R}$ that gives the d-risk of the uniformly minimal d-risk estimate follows directly from the algorithm of this estimate construction and monotonicity of the function $t(a),\ a \in \mathrm{R}$. The proposition is proved. ■

Figure 2 provides graphs for the d-risks of the Bayesian (solid line) and uniformly minimal d-risk (dotted line) estimates with the parameter values of the probability model from the example given in Section 6. These functions are practically identical on the joint part ($d \ge \theta_0(1+\Delta)$) of their domains. This is a corollary of the well known statement on the asymptotic equivalence of the maximum likelihood, Bayesian and uniformly minimal d-risk estimates (see [7–9]). Obviously, an application of the results from these articles



(a)

(b)

(c)

(d)

Figure 2. Graphs of d-risk functions. (a) $\Delta = 0.1,\ n = 2$. (b) $\Delta = 0.2,\ n = 2$. (c) $\Delta = 0.1,\ n = 4$. (d) $\Delta = 0.05,\ n = 14$.

to our probability model requires a check of corresponding regularity conditions; however, in this case, everything is quite trivial because all three estimates are equivalent to $\widehat{\theta}_n = T(\overline{X}) = \overline{X} + \lambda\sigma^2/n$ as $n \to \infty$.

In connection with these graphical illustrations, it should also be noted that the support of the Bayesian estimate is $[\theta_0(1+\Delta), \infty)$, while the support of the minimal risk estimate is somehow wider: $[\theta_0, \infty)$. From the point of view of the theory of estimates with uniformly minimal risk [5,6], d-risks of these estimates can be compared only on the intersection of the supports of these distributions. The fact that the estimate with the uniformly minimal risk has a bigger maximum value does not mean that it is worse than Bayesian. The maximum of the estimate with the uniformly minimal risk is achieved outside of the region of the d-risk values for the Bayesian estimate; it is the best from the d-risk point of view among all estimates with the distribution that have support equals to the interval $[\theta_0, \infty)$. This remark plays an important role in the comparison of estimates by the minimal sample size that guarantee the given constraint on the d-risk.

4. The sample size that guarantees given constraints on the relative accuracy and the *d*-risk of the estimate

Sample size planning is usually associated with the risk of a certain estimate, considering it as a function of the sample size n. The minimal sample size that guarantees the given accuracy and reliability of the estimate is defined as the smallest n for which the risk of the employed estimate does not exceed the given constraint. There is much literature, including monographs, devoted to this subject. We refer to one of the recent monographs [10], which considers myriad procedures of statistical inference. It also suggests methods for sample size determination with the help of normal approximations for the estimate distribution or test statistics without discussing mathematical aspects of the validity of using these approximations. The literature cited in the monograph is not exhaustive, but is extensive; it contains nearly a thousand references.

However, there is a different aspect to the problem of sample size planning. It is connected to a definition of the *necessary sample size* as the smallest n for which exists the guarantee procedure of statistical inference. As shown in article [4], the decision rules that correspond to such critically minimal sample size should be minimax for the loss function normed by the risk constraint. In our case of the mean value estimation with given constraints on the relative error and d-risk, such a rule does exist; it is the estimate with the uniformly minimal d-risk.

Therefore, *for the estimates with the distribution support that is the interval* $[\theta_0, \infty)$, the necessary sample size for the estimate of θ with the given constraint Δ on the relative error of the estimate and the constraint β on d-risk is defined as the smallest integer $n^* = n^*(\Delta, \beta)$, which satisfies the inequality

$$\max_{d \geq \theta_0} \frac{\Phi\left[(u_2(d) - t(d))\sqrt{n}/\sigma\right] - \Phi\left[(u_1(d) - t(d))\sqrt{n}/\sigma\right]}{\Phi\left((t(d) - \theta_0)\sqrt{n}/\sigma\right)} \geq 1 - \beta,$$

where u_1 and u_2 have the form (2).

The following lax and purely heuristic reasonings led to a surprisingly accurate closed form formula for n^*.

Assign the value 1 to the normal distribution function in the denominator of the d-risk. By this we only increase the d-risk. Exchange $(1 + \Delta)^{-1}$ by the smaller $1 - \Delta$, increasing d-risk again. Make use of the fact that our numerical illustrations for the function $t(d)$ indicate the possibility of its exchange by d. Let d_0 be the point where the maximum of the d-risk function is achieved. After these simplifications and substitutions, the inequality for n^* determination takes the form

$$2\Phi\left(d_0 \frac{\Delta(1 + \Delta)\sqrt{n}}{\sigma}\right) - 1 \geq 1 - \beta.$$

Finally, assuming that $d_0 = \theta_0$, we obtain the desired approximation for the necessary sample size as the smallest integer value of

$$\widetilde{n} \geq \left(\frac{\Phi^{-1}(1 - \beta/2)\sigma}{\theta_0 \Delta(1 + \Delta)}\right)^2.$$

This formula has the same form as its counterpart in article [2], where the constraints α were imposed on the classical risk function for parameter θ and the classical minimax

Table 1. The minimal sample size needed for the guarantee of estimation.

Δ	0,05	0,06	0,07	0,08	0,09	0,10	0,11	0,12	0,13	0,16	0,17
n^*	16	11	8	7	5	4	4	3	3	2	2
n_B	14	10	7	6	4	4	3	3	2	2	1
$\widetilde{n}$	14	10	7	6	4	4	3	3	2	2	1

estimate of θ was established. As in that article, we pay attention to the rapid increase of the necessary sample size when θ_0 approaches the value of zero.

We note once more that n^* is the necessary sample size only for the d-guarantee procedures θ with the distribution support that equals the interval $[\theta_0, \infty)$. According to the theory of estimates with the minimal d-risk [5,6] any other estimate θ will have a bigger d only on the intersection of its distribution support with the interval $[\theta_0, \infty]$. The minimal sample size that provides its d-guarantee property of any estimate with a different support of its distribution, can differ from n^* in any direction. The Bayesian estimate takes values only in the interval $[\theta_0(1+\Delta), \infty)$, has the smallest d-risk in this interval (see Figure 2), but has a different minimal sample size n_B than n^*. The value n_B is defined as the smallest integer n that satisfies the inequality (see Proposition 3.1)

$$\min_{d \geq \theta_0(1+\Delta)} \mathfrak{R}_B(d) \leq \beta.$$

The closeness of the d-risk functions of this interval assures the same approximation $\tilde{n}$ for n_B when the values of Δ are small. To confirm this, it is enough to repeat the arguments similar to the derivation of the approximation $\tilde{n}$ for n^*, first making sure that with the accuracy up to $O(\Delta^2)$, the functions $V \sim 1/2$, $U^2 \sim \sigma^2$ and $(d^2 - U^2/n)/(2dV) \sim d$ for $n \to \infty$ (that is, if we neglect the value U^2/n).

How accurately $\widetilde{n}$ serves as an approximation for n^* and n_B, can be seen by the values presented in Table 1.

5. Sequential procedure for θ estimation

Let β be a fixed constant. A sequential procedure of the 'First Crossing' has been defined in article [1] in the framework of the general problem of the statistical inference guarantee. For our goal of θ estimation, this procedure is defined by the stopping time

$$\nu = \min\{n : H(\hat{a}_n(T)) \geq 1 - \beta\}$$

with the subsequent decision for the parameter value with the help of the Bayesian estimate (see formula (3)): $\hat{\theta}_\nu = \hat{\theta}_\nu(T) = \max\{\theta_0(1+\Delta), \hat{a}_\nu\}$. Therefore, the observations are stopped when the random process

$$\left\{T_n = \overline{X} - \lambda\sigma^2/n,\ n \geq 1\right\}$$

reaches the boundary of the region $\{W(n),\ n \geq 1\}$, which is defined as the solution of the equation $\mathrm{H}_n(T) = 1 - \beta$ by T. After stopping the experiment at the step $\nu = n$, the Bayesian estimate $\hat{\theta}_n = \max\{\theta_0(1+\Delta), \hat{a}_n\}$ is calculated.

The graphical representation of the First Crossing procedure is given in Figure 3. This figure illustrates the remarkable property of the suggested sequential procedure for θ

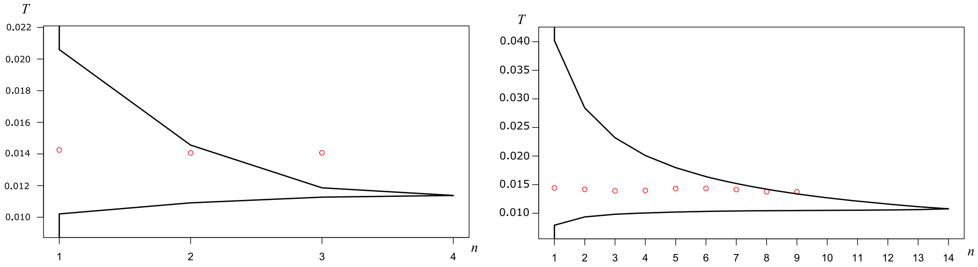


Figure 3. The experiment continuation area with $\Delta = 0.1$ and $\Delta = 0.05$. The solid line is the stopping boundary $W(n)$. The points are the result of process path modeling T_n. (a) $\Delta = 0.1$, $\beta = 0.05$. (b) $\Delta = 0.05$, $\beta = 0.05$.

estimation: as does its counterpart in article [3], it has the stopping time $\nu_B \leq n_B$.We discovered a similar phenomenon for the case of parameter estimation of different probability models; in particular for the success probability estimation of Bernoulli trials with the quadratic loss function and prior Beta distribution. With this comes the assumption that the First Crossing sequential procedure always has the stopping time ν for which $P(\nu \leq n_B) = 1$; that is, from the point of view of estimation sequential procedures, the procedure with a fixed number of observations is the worst one because it requires the largest sample size. At this moment, we do not have a rigorous mathematical proof of this statement.

An interesting fact for practical application of the First Crossing procedures can be noted from the shape of the region of continuation of observations for this procedure: if the observed value of the statistic T appears to be too small or too large, then only one observation for the guarantee estimate θ is required.

For the probability model parameters that correspond to the example considered in the next section, the properties of the First Crossing procedure have been investigated by the method of statistical simulations. We ran 10^5 iterations of the First Crossing procedure, and for each iteration, the stopping time ν was recorded and the value of Bayesian estimate calculated. The results of simulations are presented in Figures 3 and 4 as tables of the relative frequencies of stopping the experiment at the corresponding steps.

The effectiveness of an sequential procedure is usually measured by the ratio of the necessary sample size to the mean value of the stopping time of the procedure: $\mathcal{E} = n^*/E\nu$. For our sequential procedure the mean value of the stopping time depends on β and Δ and approximately two times less than the necessary sample size. For the frequency data of stopping times for the stopping times of our procedure presented in Figures 3 and 4, the effectiveness values are $\mathcal{E}_{0.05} = 2.2$ and $\mathcal{E}_{0.1} = 2.1$, respectively.

6. The empirical counterpart of the Bayesian estimate

Practical application of the suggested estimation procedures is possible only if the true values of the parameters σ^2, λ and θ_0 are known. However, laboratories systematically estimating θ for each incoming product for an analysis; usually have an archive of observations that consists of a sufficiently large number N of samples $X_{i1}, \ldots, X_{in}$, $i = 1, \ldots, N$. With such data, it is possible to estimate sufficiently accurate the variance σ^2 and θ_0 of the

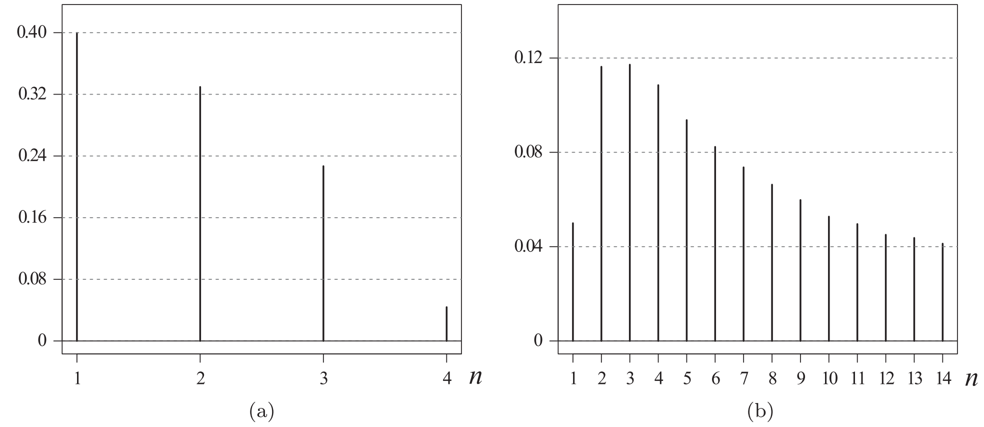


Figure 4. Statistical experiment stopping frequencies for steps $n = 1, 2, \ldots$. (a) $\Delta = 0.1$, $\beta = 0.05$. (b) $\Delta = 0.05$, $\beta = 0.05$.

random variable of interest. As for the parameter λ of the prior distribution, its value is estimated from the archive data by the standard methods of mathematical statistics, namely the method of maximum likelihood or the method of moments.

To use these methods, one must find the marginal distribution of the sufficient statistic $\overline{X}$ knowing its conditional normal $(\vartheta, \sigma^2/n)$ distribution and the prior two-parameter exponential distribution of ϑ. The density function of $\overline{X}$ is located in the denominator of the posterior density of ϑ and is equal to

$$p_G(x) = \lambda \exp\left\{-\frac{n}{2\sigma^2}(x^2 - t^2) + \lambda\theta_0\right\} \cdot \left[1 - \Phi\left((\theta_0 - t)\frac{\sqrt{n}}{\sigma}\right)\right], \quad x \in \mathbf{R}.$$

Obviously, finding the maximum point by λ of the likelihood function is computationally cumbersome because of its complex relationship to this parameter. Nevertheless, the value of λ can be simply estimated by the method of moments because

$$\mathbf{E}\overline{X} = \lambda \int_{\theta_0}^{\infty} \theta \exp\{-\lambda(\theta - \theta_0)\}\, \mathrm{d}\theta = \frac{1}{\lambda} + \theta_0.$$

Hence the estimates of θ_0 and λ by the sample means $\overline{X}_1, \ldots, \overline{X}_N$ reduced by the data archive, have the form

$$\widehat{\theta}_{0,N} = \min_{1 \le i \le N} \overline{X}_i, \quad \widehat{\lambda}_N = \left[\frac{1}{N}\sum_{i=1}^{N} \overline{X}_i\right]^{-1} + \widehat{\theta}_{0,N}.$$

However, it is possible to estimate the parameters θ_0 and λ without using a data archive, which will be shown by the following example.

7. An application of the *d*-guarantee procedure to real life data

Consider the described procedure applied to the problem of estimating the arsenic concentration in drinking water.

Regulatory documents, used in everyday quality testing of drinking wanter, state standards, ecological norms and regulations determine the mandatory requirements for water quality (see Russian Federation State Standards $GOSTZISO5725-6$ and $SanPIN2.1.4.1074-0.1$). Our consultations with fellow workers at the Department of Applied Ecology at the Kazan Federal University and their recommendations for the usage of the real data obtained in the laboratory at a particular water intake from Volga river, allowed us to set the following values for the parameters of probability model.

Extensive archival data with everyday measurements of arsenic content show that with large reliability (not less than 0.99), the value of arsenic content (mg/litre) lies within 0.01–0.05 with a clear tendency of the distribution skewness towards small values. The value 0.05 is the maximum allowable concentration (MAC) of the arsenic, the value $\theta_0 = 0.01$ has been defined as the lower 99% tolerant limit, and besides this value practically coincides with the smallest value for all sample data of the archive. Standard deviation $\sigma = 0.001$ has been defined by the Russian Federation State Standards $GOSTZISO5725-6$ (Section 6): 20% of the smallest content 0.01 equals 2σ.

Assuming that the intake quality level of the drinking water is not less than $Q_{in} = 0.99$, we can derive the parlayer λ value of the prior distribution from the equation $\exp\{-\lambda(0.05-0.01)\} = 0.99$. In our case $\lambda \approx 100$. Accuracy properties illustration for the proposed estimation procedures and sample sizes were held for the reliability value $1-\beta = 0.95$ and two values $\Delta = 0.1$ and 0.2.

Remark: It should be noted that for this particular problem, the constraint on relative accuracy should be used in the following 'nonsymmetric shape'

$$\frac{1}{1+\Delta_1} < \frac{\vartheta}{d} < 1+\Delta_2$$

with $\Delta_1 \ll \Delta_2$ because the decision based on a small (acceptable) amount of arsenic in the case when the water actually contains far more arsenic, causes more severe consequences than the decision based on the amount that exceeds the true value of θ. All constructions presented in our article can be easily extended to the case of such a loss function.

8. Concluding remarks

In this article, we obtained a solution to the statistical problem of a normal mean value estimation with the prior knowledge of a positive small value of the estimated parameter. The sample size is determined by the given constraints of the relative error and d-risk of the estimate. The defining moment in the solution of this well studied problem is in the essentially new approach to the definition of notion reliability and designing an estimate, which is different from the classical and Bayesian estimates. We do not require a small probability α for deviations (Δ) of an estimate (random variable or statistic) from the fixed (given) value θ, but we ensure the guarantee probability $1-\beta$ of the random parameter ϑ falling into the given neighbourhood of the estimate value obtained by the statistical experiment. This is the so-called d-posterior approach to the guarantee problem of statistical inference, which was developed by statisticians at the Kazan University in the late 1970s.

In the article we provide the method and approximate formula for the minimal sample size determination for which there exists a procedure that guarantees the required relative accuracy and 'd-reliability' $1 - \beta$ for the θ value decision. This method provides the value of so-called 'necessary sample size' only for the estimates with the support that coincides with the support $[\theta_0, \infty)$ of the prior distribution. The estimation procedure that corresponds to this minimal sample size is obtained. A sequential d-guarantee procedure for θ estimation is also suggested. The numerical illustrations show that the sample size (stopping time) for the sequential procedure is always smaller than the minimal sample size for the Bayesian guarantee estimation by the sample of a fixed size. The distribution of the stopping time of the sequential procedure (the random sample size) is illustrated by using data from the practical problem of estimating the arsenic content in drinking water. The results of statistical simulation show that, in some cases, the gain in sample size may be quite significant.

The results obtained in the article provide new directions for designing quality control procedures and certifying manufactured products.

Acknowledgments

The authors are grateful to the referee for carefully reading the manuscript and for offering substantial suggestions that enabled them to improve their paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Volodin IN. Guarantied procedures of statistical inference (sample size determination). J Soviet Math. 1989;44:568–600.

[2] Volodin IN. On an estimation of the mean value of the normal distribution with guarantee constraints on the relative error. (Russian) Publisher Metallurgiya. M.: Zavodskaya Laboratoriya. 1978;44:69–72.

[3] Volodin IN, Salimov RF, Turilova EA, et al. Estimation of the mean value for the normal distribution with constrains on d-risk. Lobachevskii J Math. 2018;39:377–387.

[4] Volodin IN. Optimum sample size in statistical inference procedures. Soviet Math (Iz VUZ). 1980;22:68–78.

[5] Simushkin SV, Volodin IN. Statistical inference with minimal d-risk. Berlin–New York: Springer; 1983. p. 629–636. (Lecture notes in mathematics; 1021).

[6] Simushkin SV, Volodin IN. Statistical inference with minimal d-risk. J Soviet Math. 1988;42:1464–1472.

[7] Ibragimov IA, Has'minskiĭ RZ. The limiting behavior of Bayesian estimates and estimates of maximal plausibility. Soviet Math Dokl. 1971;12:831–834.

[8] Volodin IN, Novikov AA. Statistical estimates with asymptotically minimal d-risk. Theory Probab Appl. 1993;38:118–128.

[9] Zaikin AA. Estimates with the asymptotically minimal d-risk (Russian). Theory Probab Appl. 2018;63:609–618.

[10] Ryan TP. Sample size determination and power. Hobeken, NJ; John Wiley and Sons; 2013. (Wiley series in probability and statistics).