УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИЕ НАУКИ

2018, Т. 160, кн. 2 ISSN 2541-7746 (Print)
С. 309–316 ISSN 2500-2198 (Online)

UDK 519.226.3

LOWER BOUNDS FOR THE EXPECTED SAMPLE SIZE IN THE CLASSICAL AND *d*-POSTERIOR STATISTICAL PROBLEMS

I.A. Kareev, I.N. Volodin

Kazan Federal University, Kazan, 420008 Russia

Abstract

In this report, the problem of construction of lower boundaries for the expected sample size of statistical inference procedures has been considered. The general methodology for construction of the lower bounds and the review of the main results for the classical statistical problems have been presented along with the analysis of the new and earlier results on adoption of the technique to the d-posterior approach. Namely, the hypothesis testing problem has been considered.

Keywords: expected sample size, lower bounds, efficiency, d-posterior approach, Bayesian paradigm, hypothesis testing

Introduction

In mathematical statistics, we have some inequalities determining lower boundaries for various components of the risk functions of estimating and hypotheses distinguishing procedures. Rao–Cramér's inequality is the most famous one. It determines a lower boundary for the estimation variance constricted by a sample with the fixed size when the distribution of an observed random variable satisfies certain regularity conditions. Various generalizations and modifications of this inequality were developed by A. Bhattachryya, L. Bolshev, E. Barankin, J. Chipman, H. Robbins, et al. J. Wolfowitz generalized Rao–Cramer's inequality for sequential sampling.

Another well-known inequality was introduced by A. Wald. He determined a lower boundary for the expected sample size in any sequential procedure regarding the problem of distinguishing between two simple hypotheses with the given limits on the probabilities of type-I and type-II errors. W. Hoeffding and G. Simons generalized this inequality for the case of distinguishing between more than two hypotheses (see [1]). Later, in the 1960s-1880s, I. Volodin [2–12], as well as some other authors, have established several analogous inequalities for the expected total sample size in the problems of hypothesis testing, classification, selection, etc. The essential similarity of all these inequalities is that they are only simple implications of a single important property of the Kullback-Leibler divergence: data contained in the statistic set do not exceed those contained in the sample.

Several uses can be distinguished for such lower boundaries:

1) they can be used as a robust criterion of sample size insufficiency — if the expected sample size is less than the lower boundary, then there is no appropriate procedure for solving the statistical problem with the given limits on the risks;

2) they can be used to measure the efficiency of existing procedures by comparing their needed sample size to some theoretical optimal one;

3) they can be used as some another measure of difficulty of a problem.

This paper provides an overview of the obtained lower boundaries for the average sample size with regard to many classical problems of mathematical statistics (section 1) and presents the new and earlier results on adaptation of the lower bounds construction methods for Bayesian problems, namely on hypothesis testing in the d-posterior approach (in section 2).

1. Volodin's lower bounds in the general form and their applications for the classical statistical problems

In his earliest investigations, I.N. Volodin introduced a general method for the construction of lower boundaries for the expected sample size of statistical inference procedures [5]. The method allows to obtain the closed form of lower boundaries for a wide range of statistical problems. Here, we provide Malyutov's modification [10] of that method, which gives more precise lower bounds in problems when several independent populations are involved.

1.1. The lower bound in the general form (see [5, 10]). Let us denote the Kullback–Leibler divergence by

$$\mathrm{KL}(F_1, F_2) = \int \ln \frac{\mathrm{d}F_1}{\mathrm{d}F_2}\, \mathrm{d}F_1$$

for some distributions F_1 and F_2. When $F_1 = F(\theta)$, $F_2 = F(\vartheta)$ (i.e., they coincide up to the value of parameter θ) we denote $\mathrm{KL}(\theta, \vartheta; F) = \mathrm{KL}(F(\theta), F(\vartheta))$.

Let us consider the general problem of statistical inference where we observe m populations $X_1, \ldots, X_m$ independently. The lower boundary for the expected total sample size $\nu = \nu_1 + \cdots + \nu_m$ is given by the inequality:

$$\mathbb{E}(\nu \,|\, \theta) \geq \inf_{\varphi} \sup_{\vartheta \in \Theta} \mathrm{KL}(\theta,\, \vartheta;\, \delta_\varphi) \Big/ \sum_{i=1}^{m} \pi_i^{\varphi} \mathrm{KL}(\theta,\, \vartheta;\, X_i) \quad \forall\, \theta \in \Theta, \tag{1}$$

where δ_φ is a random variable denoting the decision made by the procedure φ after the experiment is over; $\pi_i(\theta) = \mathbb{E}\big(\nu_i^{\varphi} \Big/ \sum_{j=1}^{m} \nu_j^{\varphi}\big)$ is the expected ratio of observations, which φ takes from the i-th population.

When $\operatorname{dom} \delta \in \{d_0, d_1\}$ (a bivalued random variable), then

$$\mathrm{KL}(\theta,\, \vartheta\,;\, \delta_\varphi) = w\big(\, \psi(d_0 \,|\, \theta)\,,\, 1 - \psi(d_0 \,|\, \vartheta)\,\big),$$

where

$$w(x, y) = x \ln \frac{x}{1-y} + (1-x) \ln \frac{1-x}{y}, \qquad \psi(d \,|\, \theta) = \mathbb{P}(\delta_\varphi = d \,|\, \theta).$$

1.2. Multiple simple hypotheses testing (see [2]). Consider the problem of distinguishing between $m \geq 3$ simple hypotheses

$$H_i : \theta = \theta_i, \quad i = 1, \ldots, m,$$

about the distribution of a population $X \sim F(\theta)$, $\theta \in \Theta \subset \mathbb{R}$.

For this problem, the inequality (1) gives us

$$\mathbb{E}_{\theta_i} \nu \geq \max_{j \neq i} \sum_{k=1}^{r} \alpha_{ik} \ln(\alpha_{ik}/\alpha_{jk}) \Big/ \mathrm{KL}(\theta_i,\, \theta_j;\, X), \quad i = 1, \ldots, m,$$

where $\|\alpha_{ij}\| = \|\, \psi(d_j \,|\, \theta_i)\, \|$ is a matrix consisting of values of the operating characteristic (the *strength* of the procedure φ).

1.3. Goodness-of-fit test (see [6]). Let $\mathcal{F}$ be a family of mutually absolutely continuous distributions F in the same measurable space. We consider the problem of testing the null hypothesis (for some $\Delta > 0$)

$$H_0\colon F = F_0 \quad \text{against} \quad H_1\colon \sup_{A\in\mathcal{A}} |F(A) - F_0(A)| \geq \Delta$$

about the distribution F of a population X with the given limitation α_0, α_1 on probabilities of type-I and type-II errors. Here and elsewhere, the $\mathcal{A}$ is the algebra of the problem's probabilistic space $(\mathcal{X}, \mathcal{A})$.

For this problem, we obtained the lower boundaries on the expected sample size when H_1 is true:

$$\mathbb{E}(\nu \mid F \in H_1) \geq \frac{w(\alpha_1, \alpha_0)}{\mathrm{KL}(F, F_0)}.$$

When H_0 is true, the lower bound is:

$$\mathbb{E}(\nu \mid F \in H_0) \geq \frac{w(\alpha_0, \alpha_1)}{\inf\limits_{F\in H_1} \mathrm{KL}(F_0, F)} = \frac{w(\alpha_0, \alpha_1)}{-h(1/2 - 2\Delta/3) - C\cdot\Delta^8},$$

where

$$h(p) = p \ln \frac{p+\Delta}{p} + (1-p)\ln\frac{1-p-\Delta}{1-p}$$

and

$$0 \leq C \leq \frac{1024}{3645}\left(1 + \frac{8\Delta^6}{91}\right).$$

1.4. Homogeneity test (see [6]). Let $\mathcal{F}$ be a family of mutually absolutely continuous distributions F on the same measurable space. Let $X_1 \sim F_1$ and $X_2 \sim F_2$ be the population which can be observed in an arbitrary way, so $\nu = \nu_1 + \nu_2$. We consider the problem of testing the null hypothesis (for some $\Delta > 0$)

$$H_0\colon F_1 = F_2 \in \mathcal{F}_0 \quad \text{against} \quad H_1\colon \sup_{A\in\mathcal{A}} |F_1(A) - F_2(A)| \geq \Delta$$

subject to the limitations α_0, α_1 on probabilities of type-I and type-II errors.

For this problem, there is the following adaptation of the general lower boundary (1). When H_0 is true:

$$\mathbb{E}(\nu \mid H_0) \geq \frac{2w(\alpha_0, \alpha_1)}{-\ln(1-\Delta)}. \tag{2}$$

When H_1 is true:

$$\mathbb{E}(\nu \mid H_1) \geq \frac{2w(\alpha_1, \alpha_0)}{\ln(1-\Delta^2) + \Delta\ln\dfrac{1+\Delta}{1-\Delta}}. \tag{3}$$

1.5. Test for invariance to a group of transformations (see [7]). Let G be a group of transformations and consider two sets of distributions:

$$\mathcal{F}_0 = \{F\colon F(A) = F(gA) \quad \forall A \in \mathcal{A}, \quad \forall g \in G\},$$

$$\mathcal{F}_1 = \left\{ F\colon \exists g \in G, \quad \exists \mathcal{A}_0 \subseteq \mathcal{A} \quad \sup_{A\in\mathcal{A}_0} |F(A) - F(gA)| \geq \Delta \right\}, \quad \Delta > 0.$$

The problem of invariance to a group of transformation consists in testing the null hypothesis

$$H_0\colon F \in \mathcal{F}_0 \quad \text{against} \quad H_1\colon F \in \mathcal{F}_1$$

about the distribution F of the population subject to the limits α_0, α_1 on the risk.

Surprisingly, the adaptation of the general lower boundary (1) yields the same form as of the lower boundaries for the homogeneity testing problem (2), (3).

1.6. Selection problem (see [13–15]). Let us have $m \geq 3$ populations $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$, $i = 1, \ldots, m$ with the same known σ^2. The problem is to select the population with the highest value of θ, i.e., to select one of the hypotheses

$$H_i\colon \theta_i \geq \theta_j + \Delta \quad \forall\, j \neq i, \quad \Delta > 0,$$

subject to the limit α on the wrong decision probability. The populations might be observed in an arbitrary way, so the total sample size $\nu = \sum_{i=1}^{m} \nu_i$.

The lower boundary for the least favorable case:

$$\sup_{\theta_1,\ldots,\theta_m} \mathbb{E}(\nu \mid \theta_1, \ldots, \theta_m) \geq \frac{(\sqrt{m-1}+1)^2}{2\Delta^2}\sigma^2 w(\alpha, \alpha).$$

1.7. Ranking problem (see [14, 16]). Let us have $m \geq 3$ populations $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$, $i = 1, \ldots, m$ with the same known σ^2. It is known that $|\theta_i - \theta_j| \geq \Delta$, $\Delta > 0$. The problem is to place the population in an ascending order of values θ_i subject to the limit α on the probability of wrong ordering. The populations can be observed in an arbitrary way, so the total sample size $\nu = \sum_{i=1}^{m} \nu_i$.

The lower bound for the least favorable case:

$$\sup_{\theta_1,\ldots,\theta_m} \mathbb{E}(\nu \mid \theta_1, \ldots, \theta_m) \geq \frac{m-1}{\Delta^2}\sigma^2 w(\alpha, \alpha).$$

2. The lower bounds for hypotheses testing in the *d*-posterior approach

We consider the following Bayesing problem. Let $X \sim F(\vartheta)$, where $\vartheta \in \Theta \subseteq \mathbb{R}$ is the unknown random parameter of interest; let $\vartheta \sim G$. The problem is to distinguish between the null hypotheses

$$H_0\colon \vartheta \in \Theta_0, \quad H_1\colon \vartheta \in \Theta_1$$

based on the observations from X, where $\Theta_0 + \Theta_1 = \Theta$. Let us put $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$.

Let $\operatorname{dom} \delta_\varphi \in \{d_0, d_1\}$, where d_0 denotes the selection of H_0 by a procedure φ after an experiment, and d_1 is the selection of H_1. In the d-posterior approach, type-I (on the left) and type-II (on the right) d-risks are considered:

$$\mathbf{P}(\vartheta \leq 0 |\ \delta = d_1), \quad \mathbf{P}(\vartheta > 0 |\ \delta = d_0).$$

Type-I d-risk is a probability of that H_0 is correct among all experiments for which the procedure φ selected H_1 by the results of the experiment. On the contrary, type-II d-risk is a probability of that H_1 is correct among all experiments for which H_0 was selected.

For the considered hypotheses, the testing problem we bring in the constraints – the type-I and type-II d-risk must be less than the prescribed limits β_0 and β_1:

$$\mathbf{P}(\vartheta \leq 0 |\ \delta = d_1) \leq \beta_0, \quad \mathbf{P}(\vartheta > 0 |\ \delta = d_0) \leq \beta_1. \tag{4}$$

We suggest the following lower boundary for the expected sample size when H_0 is true:

$$\frac{1}{G_0} \int\limits_{\theta \in \Theta_0} \mathbf{E}_\theta \nu \, \mathrm{d}G(\theta) \geq L,$$

where

$$L = \inf_{\varphi} \frac{1}{G_0 G_1} \int\limits_{\theta\in\Theta_0} \mathrm{d}G(\theta) \int\limits_{\vartheta\in\Theta_1} \mathrm{d}G(\vartheta) \frac{\mathrm{KL}(\theta,\vartheta;\,\delta)}{\mathrm{KL}(\theta,\vartheta;\,X)}, \tag{5}$$

$G_k = G(\Theta_k)$, and $\inf\limits_{\varphi}$ is taken subject to restrictions (4) on the d-risks for φ. Note that

$$\mathrm{KL}(\theta,\vartheta;\,\delta) = w(\psi(\theta), 1-\psi(\vartheta)) = \psi(\theta)\ln\frac{\psi(\theta)}{\psi(\vartheta)} + (1-\psi(\theta))\ln\frac{1-\psi(\theta)}{1-\psi(\vartheta)},$$

where $\psi(\theta) = \mathbb{P}(\delta = d_0 |\, \theta)$. Namely, the lower bound depends on the procedure φ only through its operative characteristic. Thus, we can consider inf by ψ rather than inf by φ.

Lemma 1. *The constrains on d-risks*

$$\mathbf{P}(\vartheta \le 0|\ \delta = d_1) = \beta_0, \quad \mathbf{P}(\vartheta > 0|\ \delta = d_0) = \beta_1.$$

are equivalent to the following constraints on the "Bayesian" risks:

$$\int\limits_{\Theta_0} \psi(\theta)\,\mathrm{d}G(\theta) = a_0, \quad \int\limits_{\Theta_1} \psi(\theta)\,\mathrm{d}G(\theta) = a_1, \tag{6}$$

where

$$a_0 = \frac{(1-\beta_1)(G_0-\beta_0)}{1-\beta_0-\beta_1}, \quad a_1 = \frac{\beta_1(G_0-\beta_0)}{1-\beta_0-\beta_1} \tag{7}$$

Proof. We put

$$\psi_k = \int\limits_{\Theta_k} \psi(\theta)\,\mathrm{d}G(\theta), \quad k = 0,1.$$

The equations on d-risks can be rewritten in terms of ψ_k as

$$\frac{G_0-\psi_0}{1-\psi_0-\psi_1} = \beta_0, \quad \frac{\psi_1}{\psi_0+\psi_1} = \beta_1,$$

from which we easily obtain the statement of the lemma. □

By swapping the order of inf and the integration, we obtain the basic estimate for the lower boundary.

Theorem 1. *Let us suppose that the* inf *in (5) is reached on the procedure with the non-increasing operating characteristic $\psi(\theta)$. Then,*

$$L \ge \frac{1}{G_0 G_1} \int\limits_{\theta\in\Theta_0} \mathrm{d}G(\theta) \int\limits_{\vartheta\in\Theta_1} \mathrm{d}G(\vartheta)\,\mathbb{I}\big(h_0(\theta) \ge h_1(\vartheta)\big) \frac{w\big(h_0(\theta), 1-h_1(\vartheta)\big)}{\mathrm{KL}(\theta,\vartheta;\,X)},$$

where

$$h_0(\theta) = \frac{a_0 - G(\theta)}{G_0 - G(\theta)}, \quad h_1(\vartheta) = \frac{a_1}{G(\vartheta) - G_0},$$

and a_0, a_1 are as in (7):

$$a_0 = \frac{(1-\beta_1)(G_0-\beta_0)}{1-\beta_0-\beta_1}, \quad a_1 = \frac{\beta_1(G_0-\beta_0)}{1-\beta_0-\beta_1}.$$

Proof. Let us consider a set H of all non-increasing functions $h(\theta)$, such that $h(-\infty)=1,\ h(\infty)=0$. Then,

$$L \geq \frac{1}{G_0G_1}\int\limits_{\theta\in\Theta_0} \mathrm{d}G(\theta)\int\limits_{\vartheta\in\Theta_1} \mathrm{d}G(\vartheta)\frac{\inf_{h\in H} w\big(h(\theta),1-h(\vartheta)\big)}{\mathrm{KL}(\theta,\vartheta;\,X)}$$

For fixed $\theta\in\Theta_0, \vartheta\in\Theta_1$, the minimum by $h(\theta)$ of $w\big(h(\theta),1-h(\vartheta)\big)$ subject to the constrains (6) is reached on a step-function of form

$$h(x)=\begin{cases}1, & x\leq\theta;\\ y_0, & \theta<x\leq 0;\\ y_1, & 0<x\leq\vartheta;\\ 0, & \vartheta<x.\end{cases}$$

Now, minimizing the expression by y_0, y_1 gives the statement of the theorem. □

Conclusions

In this paper, the results of the studies on the problems of construction of lower bounds for the expected sample size of statistical inference procedures are presented. As the review shows, the problem of construction of lower bounds for the classical statistical problems is sufficiently developed.

On the other hand, little has been done for the case of the Bayesian paradigm. In this paper, we present some new basic results on adaptation of the technique of lower bounds construction to the d-posterior approach. As the study shows, the construction of lower boundaries in this setting can involve solving integral minimization. Apparently, one of the best approaches is to apply calculus of the variations methods. Another one is to provide some additional assumptions and simplifications.

Acknowledgements. This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities (project no. 1.7629.2017/8.9).

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Simons G. Lower bounds for average sample number of sequential myltihypothesis tests. *Ann. Math. Stat.*, 1967, vol. 38, no. 5, pp. 1343–1364.
2. Volodin I.N. Estimates of the necessary sample size in problems of statistical classifications. II. *Theory Probab. Its Appl.*, 1977, vol. 22, no. 4, pp. 730–745. doi: 10.1137/1122086.
3. Volodin I.N. Optimum sample size in statistical inference procedures. *Izv. Vyssh. Uchebn. Zaved., Mat.*, 1978, vol. 21, no. 12, pp. 33–45. (In Russian)
4. Volodin I.N. Bounds for the necessary sample size in statistical classification problems. I. *Theory Probab. Its Appl.*, 1977, vol. 22, no. 2, pp. 339–348. doi: 10.1137/1122037.
5. Volodin I.N. Lower bounds for average sample size and efficiency of procedures of statistical inference procedures. *Theory Probab. Its Appl.*, 1979, vol. 24, no. 1, pp. 120–129. doi: 10.1137/1124009.
6. Volodin I.N. Lower bounds for the mean sample size in goodness-of-fit and homogeneity tests. *Theory Probab. Its Appl.*, 1980, vol. 24, no. 3, pp. 640–649. doi: 10.1137/1124079.

7. Volodin I.N. Lower bounds for the mean sample size in invariance tests. *Theory Probab. Its Appl.*, 1980, vol. 25, no. 2, pp. 3569–360. doi: 10.1137/1125043.

8. Volodin I.N. Lower bounds for sample size sufficient for procedures of guaranteed equivariant estimation. *Sov. Math.*, 1982, vol. 26, no. 3, pp. 15–20.

9. Khamdeev I.I. Lower bounds for sufficient sample size in procedures for guaranteed equivariant estimation (the case of general transformation group). *Sov. Math.*, 1983, vol. 27, no. 11, pp. 101–104.

10. Malutov M.B. Lower bounds for mean duration of a sequentially programmable experiment. *Sov. Math.*, 1983, vol. 27, no. 11, pp. 21–47.

11. Volodin I.N. Guaranteed statistical inference procedures (determination of the optimal sample size). *J. Sov. Math.*, 1989, vol. 44, no. 5, pp. 568–600.

12. Galtchouk L.I., Maljutov M.B. One bound for the mean duration of sequential testing homogeneity. In: Kitsos C.P., Muller W.G. (Eds.) *MODA 4 – Advances in Model-Oriented Data Analysis. Contributions to Statistics*. Heidelberg, Physica. 1995, pp. 49–56. doi: 10.1007/978-3-662-12516-8_5.

13. Kareev I.A. Lower bounds for average sample size and efficiency of sequential selection procedures. *Theory Probab. Its Appl.*, 2013, vol. 57, no. 2, pp. 227–242. doi: 10.1137/S0040585X97985935.

14. Kareev I. Lower bounds for the expected sample size of sequential procedures for selecting and ranking of binomial and Poisson populations. *Lobachevskii J. Math.*, 2016, vol. 37, no. 4, pp. 455–465. doi: 10.1134/S1995080216040119.

15. Kareev I. Lower bounds for expected sample size of sequential procedures for the multinomial selection problems. *Commun. Stat.: Theory Methods*, 2017, vol. 46, no. 19, pp. 1–8. doi: 10.1080/03610926.2016.1222429.

16. Kareev I.A. Lower bound for the average sample size and the efficiency of ranking sequential procedures. *Theory Probab. Its Appl.*, 2014, vol. 58, no. 3, pp. 503–509. doi: 10.1137/S0040585X97986709.

Received
December 14, 2017

Kareev Iskander Amirovich, Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematical Statistics

Kazan Federal University
ul. Kremlevskaya, 18, Kazan, 420008 Russia
E-mail: *kareevia@gmail.com*

Volodin Igor Nikolaevich, Doctor of Physical and Mathematical Sciences, Professor of the Department of Mathematical Statistics

Kazan Federal University
ul. Kremlevskaya, 18, Kazan, 420008 Russia

УДК 519.226.3

Нижние границы для среднего объёма выборки
для классических и d-апостериорных задач

И.А. Кареев, И.Н. Володин

Казанский (Приволжский) федеральный университет, г. Казань, 420008, Россия

Аннотация

В работе рассмотрена проблема построения нижних границ для среднего объёма наблюдений для процедур статистического вывода. Приведены общая методология построения нижних границ и обзор основных результатов, полученных для классических статистических задач. Представлены новые результаты по адаптации этой методологии к задачам, сформулированным согласно d-апостериорному подходу. В частности, рассмотрена задача построения проверки сложной гипотезы в d-апостериорной формулировке.

Ключевые слова: средний объём выборки, нижние границы, эффективность, d-апостериорный подход, Байесовская парадигма, проверка гипотез

Поступила в редакцию
14.12.17

Кареев Искандер Амирович, кандидат физико-математических наук, доцент кафедры математической статистики

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *kareevia@gmail.com*

Володин Игорь Николаевич, доктор физико-математических наук, профессор кафедры математической статистики

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия]

⟨ ***For citation*:** Kareev I.A., Volodin I.N. Lower bounds for the expected sample size in the classical and d-posterior statistical problems. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, 2018, vol. 160, no. 2, pp. 309–316. ⟩

⟨ ***Для цитирования*:** *Kareev I.A., Volodin I.N.* Lower bounds for the expected sample size in the classical and d-posterior statistical problems // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2018. – Т. 160, кн. 2. – С. 309–316. ⟩