COLLOQUIUM MATHEMATICUM

Online First version

CALCULATION OF THE DEFICIENCY OF SOME STATISTICAL ESTIMATORS CONSTRUCTED FROM SAMPLES WITH RANDOM SIZES

BY

V. E. BENING (Moscow), V. YU. KOROLEV (Moscow and Hangzhou) and A. I. ZEIFMAN (Vologda and Moscow)

Abstract. The purpose of this paper is to use deficiency to compare estimators constructed from samples with random size with those constructed from samples with non-random size. The deficiency can be a characteristic of a possible loss of accuracy of statistical inference if a random-size sample is erroneously regarded as a sample with non-random size. It is heuristically shown that if the asymptotic distribution of the sample size normalized by its expectation is not degenerate, then the deficiency of a statistic constructed from a sample with random size of expectation n with respect to the same statistic constructed as if the sample size were non-random and equal to n, grows almost linearly as n grows. A non-trivial behavior of the deficiency is only possible if the random sample size is asymptotically degenerate. This is the case considered in this paper, where we study the deficiencies of statistics constructed from samples whose sizes have the Poisson, binomial and special three-point distributions. We also give some basic properties of estimators based on samples with random sizes.

1. Introduction

1.1. Statistics from samples with random sizes motivation. In classical problems of mathematical statistics, the size of the sample, i.e., the number of available observations, is assumed to be deterministic. In asymptotic settings it plays the role of an infinitely increasing *known* parameter. At the same time, in practice very often the data to be analyzed are collected or registered during a certain period of time and the flow of events bringing consecutive observations is a random point process. Therefore, the number of observations is unknown till the end of the process of their registration and must also be treated as a (random) observation. For example, this is so in insurance statistics where during different accounting periods different numbers of insurance events (insurance claims and/or insurance

2010 *Mathematics Subject Classification*: Primary 62F12; Secondary 62F05.
Key words and phrases: estimator, risk function, deficiency, asymptotic deficiency, sample with random size, asymptotic expansion, Poisson distribution, binomial distribution.
Received 16 June 2017; revised 30 June 2018.
Published online 8 April 2019.

DOI: 10.4064/cm7327-7-2018

contracts) occur, and in high performance information systems where due to the stochastic character of the intensities of information flows, the size of data available for statistical analysis can often be regarded as random. Say, the statistical algorithms applied in high-frequency financial applications must take into account that the number of events in a limit order book during a time unit essentially depends on the intensity of order flows. Contemporary statistical procedures of insurance and financial mathematics do take this circumstance into account as one of the possible ways of dealing with heavy tails. However, in other fields such as medical statistics or quality control this approach has not become conventional yet although the number of patients with a certain disease varies from month to month due to seasonal factors or from year to year due to some epidemic reasons, and the number of failed items varies from lot to lot. In these cases the number of observations as well as the observations themselves are unknown beforehand and should be treated as random to avoid underestimation of risks or error probabilities.

In asymptotic settings, statistics constructed from samples with random size are special cases of random sequences with random indices. The randomness of indices usually leads to the limit distributions for the corresponding random sequences being heavy-tailed even in the situations where the distributions of non-randomly indexed random sequences are asymptotically normal (see, e.g., [2], [3], [4], [7] and [10]). For example, if a statistic which is asymptotically normal in the traditional sense is based on a sample with random size having a negative binomial distribution, then instead of the expected normal law, the Student distribution with power-type decreasing heavy tails appears as the asymptotic law for this statistic [2], [7].

At the same time, according to the conventional logic of statistical analysis, the distributions of statistics (estimators, tests, etc.) to be used for statistical inference should be known before the actual sample is observed in order to calculate critical values or thresholds. As a rule, asymptotic approximations by limit distributions of statistics are used instead of exact distributions because the former are much easier to compute. Therefore, in limit theorems of probability theory and mathematical statistics, centering and normalization of random variables are used to obtain non-trivial asymptotic distributions. It should be stressed that to obtain a reasonable approximation to the distribution of basic random variables, both centering and normalizing values should be non-random. Otherwise the approximate distribution becomes random itself and, say, the problem of evaluation of quantiles required for the calculation of critical values or confidence intervals becomes senseless.

Throughout the paper we use conventional notation: $\mathbb{N} = \{1, 2, \ldots\}$, $h(n) \sim f(n)$ $(n \to \infty)$ means $\lim_{n\to\infty} h(n)/f(n) = 1$. The symbols $\stackrel{d}{=}$ and

$\Rightarrow$ denote the coincidence of distributions and convergence in distribution, respectively.

Consider a family $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ of probability measures, each defined on a measurable space $(\Omega, \mathfrak{F})$. Consider a sequence of independent random variables (r.v.'s) $X_1, X_2, \dots$ defined on $(\Omega, \mathfrak{F})$. Let $N_1, N_2, \dots$ be a sequence of non-negative integer random variables defined on the same measurable space so that for each $n \geq 1$ the random variable N_n is independent of the sequence $X_1, X_2, \dots$ with respect to any measure P_θ from $\mathcal{P}$. A random sequence $N_1, N_2, \dots$ is said to be *infinitely increasing* ($N_n \to \infty$) *in probability* P_θ if $\mathsf{P}_\theta(N_n \leq M) \to 0$ as $n \to \infty$ for any $M \in (0, \infty)$.

For $n \geq 1$ let $T_n = T_n(X_1, \dots, X_n)$ be a *statistic*, that is, a measurable function of the r.v.'s $X_1, \dots, X_n$. For each $n \geq 1$ define the r.v. T_{N_n} by letting

$$T_{N_n}(\omega) = T_{N_n(\omega)}\big(X_1(\omega), \dots, X_{N_n(\omega)}(\omega)\big)$$

for every $\omega \in \Omega$. Assume that for each $\theta \in \Theta$ the expectation $\mathsf{E}_\theta T_n \equiv g(\theta)$ exists. We will say that the statistic T_n is *asymptotically normal* if

$$\mathsf{P}_\theta\big(\sqrt{n}(T_n - g(\theta)) < x\big) \Rightarrow \Phi(x) \quad (n \to \infty) \tag{1}$$

for each $\theta \in \Theta$.

The following statement describes the change of the limit law of an asymptotically normal statistic when the sample size is replaced by a r.v.

LEMMA 1. *Assume that $N_n \to \infty$ in probability P_θ as $n \to \infty$ for each $\theta \in \Theta$. Suppose the statistic T_n is asymptotically normal in the sense of* (1). *Then a distribution function $F(x)$ such that*

$$\mathsf{P}_\theta\big(\sqrt{n}(T_{N_n} - g(\theta)) < x\big) \Rightarrow F(x) \quad (n \to \infty)$$

exists if and only if there exists a distribution function $Q(x)$ satisfying the conditions $Q(0) = 0$ and

$$F(x) = \int_0^\infty \Phi(x\sqrt{y})\, dQ(y), \quad x \in \mathbb{R}, \qquad \mathsf{P}_\theta(N_n < nx) \Rightarrow Q(x) \quad (n \to \infty).$$

This lemma is a particular case of Theorem 3 in [10], the proof of which is in turn based on general theorems on convergence of superpositions of independent random sequences [11]. See also [8, Theorem 3.3.2].

This result was extended to more general normal mixtures, in particular to normal variance-mean mixtures in [12].

1.2. The concept of deficiency. Before turning to the general case of statistics constructed from samples with random size, which is the main aim of the present paper, let us recall the notion of deficiency of a statistical estimator for the traditional case where the sample size is non-random.

Suppose that $T_n^*(X_1,\dots,X_n)$ and $T_n(X_1,\dots,X_n)$ are two competing estimators of $g(\theta)$, $\theta\in\Theta$, based on n observations $X_1,\dots,X_n$ and let their expected squared errors (risk functions) be denoted by $R_n^*(\theta)$ and $R_n(\theta)$, respectively. An interesting quantitative comparison can be obtained by taking a viewpoint similar to that of the asymptotic relative efficiency (ARE) of estimators, and asking for the number $m(n)$ of observations needed by the estimator $T_{m(n)}(X_1,\dots,X_{m(n)})$ to match the performance of $T_n^*(X_1,\dots,X_n)$ (based on n observations). The asymptotic (as $n\to\infty$) comparison of the two estimators involves the comparison of $m(n)$ with n, and this can be carried out in various ways. Although the difference $m(n)-n$ seems to be a natural quantity to examine, historically the ratio $n/m(n)$ was preferred by almost all authors in view of its simpler behavior. The first general investigation of $m(n)-n$ was carried out by Hodges and Lehmann [9]. They name $m(n)-n$ the *deficiency* of T_n with respect to T_n^* and denote it as

$$d_n = m(n)-n. \tag{2}$$

Suppose that as $n\to\infty$, the ratio $n/m(n)$ tends to a limit b, the *asymptotic relative efficiency* of $T_n(X_1,\dots,X_n)$ with respect to $T_n^*(X_1,\dots,X_n)$. If $0<b<1$, we have $d_n\sim(b^{-1}-1)n$, and further asymptotic information about d_n is not particularly revealing. On the other hand, if $b=1$, the asymptotic behavior of d_n, which may now vary from $o(1)$ to $o(n)$, does provide important additional information.

If $\lim_{n\to\infty} d_n$ exists, it is called the *asymptotic deficiency* of T_n with respect to T_n^* and denoted d. At points where no confusion is likely, we shall simply call d the *deficiency* of T_n with respect to T_n^*.

The deficiency of T_n relative to T_n^* indicates how many observations one loses by insisting on T_n, and thereby provides a basis for deciding whether or not the price is too high. If the risk functions of these two estimators are

$$R_n(\theta)=\mathsf{E}_\theta(T_n-g(\theta))^2,\qquad R_n^*(\theta)=\mathsf{E}_\theta(T_n^*-g(\theta))^2,$$

then, by definition, $d_n(\theta)\equiv d_n=m(n)-n$, for each n, may be found from

$$R_n^*(\theta)=R_{m(n)}(\theta). \tag{3}$$

To solve (3), one has to treat $m(n)$ as a continuous variable. This can be done in a satisfactory manner by defining $R_{m(n)}(\theta)$ for non-integer $m(n)$ as

$$R_{m(n)}(\theta)=\big(1-m(n)+[m(n)]\big)R_{[m(n)]}(\theta)+\big(m(n)-[m(n)]\big)R_{[m(n)]+1}(\theta)$$

(see [9]).

Generally $R_n^*(\theta)$ and $R_n(\theta)$ are not known exactly and we have to use approximations. Here these are obtained by observing that $R_n^*(\theta)$ and $R_n(\theta)$

will typically satisfy asymptotic expansions (a.e.) of the form

$$R_n^* = \frac{a(\theta)}{n^r} + \frac{b(\theta)}{n^{r+s}} + o(n^{-(r+s)}), \tag{4}$$

$$R_n = \frac{a(\theta)}{n^r} + \frac{c(\theta)}{n^{r+s}} + o(n^{-(r+s)}), \tag{5}$$

for certain $a(\theta)$, $b(\theta)$ and $c(\theta)$ not depending on n and certain constants $r, s > 0$. The leading term in both expansions is the same since ARE is 1. From (2)–(5) it now easily follows that (see [9])

$$d_n(\theta) \equiv \frac{c(\theta) - b(\theta)}{ra(\theta)} n^{1-s} + o(n^{1-s}). \tag{6}$$

Hence

$$d(\theta) \equiv d = \begin{cases} \pm\infty, & 0 < s < 1, \\ \dfrac{c(\theta) - b(\theta)}{ra(\theta)}, & s = 1, \\ 0, & s > 1. \end{cases} \tag{7}$$

A useful property of deficiency is transitivity: if a third estimator $\bar{T}_n$ is given, for which the risk $\bar{R}_n(\theta)$ also has an expansion of the form (5), the deficiency d of $\bar{T}_n$ with respect to T_n^* satisfies the relation $d = d_1 + d_2$, where d_1 is the deficiency of $\bar{T}_n$ with respect to T_n and d_2 is the deficiency of T_n with respect to T_n^*.

The situation where $s = 1$ seems to be the most interesting. Hodges and Lehmann [9] demonstrate the use of deficiency in a number of simple examples with $s = 1$ (for testing problems see also [1]).

1.3. The purpose and structure of the paper. The purpose of this paper is to use deficiency to compare estimators constructed from samples with random size with those constructed from samples with non-random size. Deficiency can indicate a possible loss of accuracy of statistical inference if a random-size sample is erroneously regarded as a sample with non-random size. The present paper develops the research started in [5] and presents a number of applications of deficiency in problems of point estimation when the number of observations is random.

Section 2 contains our main results. First, in Section 2.1 we heuristically show that if the d.f. $Q(x)$ in Lemma 1 is not degenerate, then the deficiency of a statistic constructed from a sample with random size of expectation n with respect to the same statistic constructed as if the sample size were non-random and equal to n, grows almost linearly as n grows. A non-trivial behavior of deficiency is only possible if the random sample size is asymptotically degenerate. This is the case considered in Sections 2.3–2.5 where we find the deficiencies of statistics constructed from samples whose sizes

have the Poisson, binomial and special three-point distributions. Section 2.2 contains some preliminary basic results on properties of estimators based on samples with random size.

In this paper we focus on the case where the sample size is independent of the r.v.'s forming the sample. This assumption is made, first, in order to be able to use simple methods to obtain qualitative results. Second, in many applied problems this assumption does not contradict the essence of the problem. For example, this is so when the data are accumulated within a prescribed time interval (a month, a year, etc.), but the informative events form a stochastic flow. This situation is typical for financial and insurance practice or any other field of activities with accounting periods. Moreover, the independence of $X_1, X_2, \dots$ is not crucial since the basic Lemma 1 can be proved without this assumption [10]. Third, most papers considering non-independent sample sizes deal with the case of asymptotically degenerate indices. This is just the case yielding non-trivial results in the present paper. It seems that using martingale techniques or imposing some concrete conditions on the character of dependence between the sample elements and the sample size, one could extend the results of this paper to the non-independent case.

2. Deficiencies of estimators based on samples with random size

2.1. Asymptotic behavior of deficiency for samples with random size. The interpretation of deficiency as the number of additional observations required to attain the same quality needs to be refined here since this number becomes random in random-size-sample problems. In order to circumvent this difficulty assume that the r.v.'s $N_1, N_2, \dots$ are parameterized by their expectations: $\mathsf{E}_\theta N_n = n$, $n \in \mathbb{N}$. This assumption will enable us, instead of comparing random variables, to compare their easily tractable parameters.

Before giving exact formulas for deficiencies, we have to make some important heuristic comments concerning the boundedness of deficiency as a function of n. By X without any indices we will denote a r.v. with the standard normal distribution. Let T_n be an asymptotically normal (see (1)) statistic constructed from the sample $X_1, \dots, X_n$, and let T_{N_n} be (the same) statistic constructed from the random-size sample $X_1, \dots, X_{N_n}$. Assume that $\mathsf{E}T_n = g(\theta)$, $n \in \mathbb{N}$, implying $\mathsf{E}T_{N_n} = g(\theta)$, $n \in \mathbb{N}$ (see Theorem 1 below). Denote $R_n^*(\theta) = \mathsf{E}(T_n - g(\theta))^2$ and $R_n(\theta) = \mathsf{E}(T_{N_n} - g(\theta))^2$.

From Lemma 1, for n large enough we have the approximate relations

$$T_n = g(\theta) + \frac{X}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \qquad T_{N_n} = g(\theta) + \frac{X}{\sqrt{Un}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where $\mathsf{P}(U < x) = Q(x)$, $x \in \mathbb{R}$, and the r.v.'s X and U are independent.

Therefore,

$$R_n^*(\theta) = \mathsf{E}_\theta\left[\frac{X}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)\right]^2 = \mathsf{E}\left(\frac{X}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right) = \frac{1}{n} + o\left(\frac{1}{n}\right),$$

$$R_n(\theta) = \mathsf{E}_\theta\left[\frac{X}{\sqrt{Un}} + o\left(\frac{1}{\sqrt{n}}\right)\right]^2 = \mathsf{E}\left(\frac{X}{\sqrt{Un}}\right)^2 + o\left(\frac{1}{n}\right) = \frac{1}{n}\mathsf{E}\frac{1}{U} + o\left(\frac{1}{n}\right).$$

Equating $R_n^*(\theta)$ and $R_{m(n)}(\theta)$ we obtain

$$\frac{1}{n} + o\left(\frac{1}{n}\right) = \frac{1}{n+d_n}\mathsf{E}\frac{1}{U},$$

or

$$d_n/n = D + o(1) \qquad (n \to \infty),$$

where $D = \mathsf{E}U^{-1} - 1$. So, in general, if $\mathsf{E}U^{-1} \geq 1$, then $d_n = O(n)$. And the only possibility for d_n to be $o(n)$, and in particular to remain bounded, is $\mathsf{E}U^{-1} = 1$. In general, if in addition to the assumptions of Lemma 1, the family $\{N_n/n\}_{n\geq 1}$ is uniformly integrable, then the assumptions of Lemma 1 and $\mathsf{E}_\theta N_n = n$ imply that $\mathsf{E}U = 1$, so that by the Jensen inequality we have $\mathsf{E}U^{-1} \geq 1$ with equality attainable if and only if $\mathsf{P}(U = 1) = 1$.

In other words, for the deficiency d_n to be bounded in n, it is necessary that the sample size N_n be asymptotically degenerate in the sense that

$$N_n/n \to 1 \quad \text{in probability as } n \to \infty.$$

This property holds for sample sizes with the Poisson, binomial and special three-point distributions considered in the present paper.

It is worth noting that an example of geometrically distributed N_n for which the limit r.v. U has the exponential distribution vividly illustrates that the deficiency may be unbounded since in this case the Fréchet distribution of the r.v. U^{-1} has infinite first moment.

Summarizing, if the d.f. $Q(x)$ in Lemma 1 is not degenerate, then the deficiency of a statistic constructed from a sample with random size of expectation n with respect to the same statistic constructed as if the sample size were non-random and equal to n, grows almost linearly as n grows. A non-trivial behavior of the deficiency is possible only if the random sample size is asymptotically degenerate. This is the case to be considered in the present paper.

2.2. Estimators based on samples with random size. Assume that for each $n \geq 1$ the r.v. N_n takes only natural values (i.e., $N_n \in \mathbb{N}$) and is independent of the sequence $X_1, X_2, \ldots$. Everywhere in what follows the r.v.'s $X_1, X_2, \ldots$ are assumed to be independent and identically distributed with distribution depending on $\theta \in \Theta \subset \mathbb{R}$.

Recall that we assume that $\mathsf{E}_\theta N_n = n$, that is, the expected sample size equals the sample size for the case where it is non-random, that is, the r.v. N_n is parameterized by its expectation n.

THEOREM 1.

(1) *If* $\mathsf{E}_\theta T_n = g(\theta)$, $\theta \in \Theta$, *then*

$$\mathsf{E}_\theta T_{N_n} = g(\theta), \quad \theta \in \Theta.$$

(2) *Let* $R_n^*(\theta) = \mathsf{E}_\theta\big(T_n - g(\theta)\big)^2$ *and* $R_n(\theta) = \mathsf{E}_\theta(T_{N_n} - g(\theta))^2$. *Assume that there exist numbers* $a(\theta)$, $b(\theta)$*and* $C(\theta), \alpha, r, s > 0$ *such that*

$$\left| R_n^*(\theta) - \frac{a(\theta)}{n^r} - \frac{b(\theta)}{n^{r+s}} \right| \le \frac{C(\theta)}{n^{r+s+\alpha}}.$$

Then

$$|R_n(\theta) - a(\theta)\mathsf{E}_\theta N_n^{-r} - b(\theta)\mathsf{E}_\theta N_n^{-r-s}| \le C(\theta)\mathsf{E}_\theta N_n^{-r-s-\alpha}.$$

Proof. We use the total probability formula:

$$\mathsf{E}_\theta T_{N_n} = \sum_{k=1}^{\infty} \mathsf{E}_\theta T_k \mathsf{P}_\theta(N_n = k) = \sum_{k=1}^{\infty} g(\theta)\mathsf{P}_\theta(N_n = k)$$

$$= g(\theta) \sum_{k=1}^{\infty} \mathsf{P}_\theta(N_n = k) = g(\theta), \quad \theta \in \Theta,$$

and

$$|R_n(\theta) - a(\theta)\mathsf{E}_\theta N_n^{-r} - b(\theta)\mathsf{E}_\theta N_n^{-r-s}|$$

$$= \left| \sum_{k=1}^{\infty} \mathsf{E}_\theta (T_k - g(\theta))^2 \mathsf{P}_\theta(N_n = k) - a(\theta) \sum_{k=1}^{\infty} \frac{\mathsf{P}_\theta(N_n = k)}{k^r} - b(\theta) \sum_{k=1}^{\infty} \frac{\mathsf{P}_\theta(N_n = k)}{k^{r+s}} \right|$$

$$= \left| \sum_{k=1}^{\infty} \left[\mathsf{E}_\theta (T_k - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}} \right] \mathsf{P}_\theta(N_n = k) \right|$$

$$\le \sum_{k=1}^{\infty} \left| \mathsf{E}_\theta (T_k - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}} \right| \mathsf{P}_\theta(N_n = k)$$

$$\le \sum_{k=1}^{\infty} \frac{C(\theta)}{k^{r+s+\alpha}} \mathsf{P}_\theta(N_n = k) = C(\theta)\mathsf{E}_\theta N_n^{-r-s-\alpha}. \blacksquare$$

COROLLARY 1. *Let* $R_n^*(\theta)$ *and* $R_n(\theta)$ *be as in Theorem* 1. *Assume that there exist numbers* $a(\theta)$, $b(\theta)$ *and* $r, s > 0$ *such that*

$$R_n^*(\theta) = \frac{a(\theta)}{n^r} + \frac{b(\theta)}{n^{r+s}}.$$

Then

$$R_n(\theta) = a(\theta)\mathsf{E}_\theta N_n^{-r} + b(\theta)\mathsf{E}_\theta N_n^{-r-s}.$$

Consider some examples.

1. Let observations $X_1, \dots, X_n$ have expectation $\mathsf{E}_\theta X_1 = g(\theta)$ and variance $\mathsf{D}_\theta X_1 = \sigma^2(\theta)$. The customary estimator for $g(\theta)$ based on n observations

is

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{8}$$

This estimator is unbiased and consistent, and its variance is

$$R_n^*(\theta) = \mathsf{D}_\theta T_n = \sigma^2(\theta)/n. \tag{9}$$

If this estimator is based on a sample with random size, then (see Corollary 1)

$$R_n(\theta) = \mathsf{D}_\theta T_{N_n} = \sigma^2(\theta)\mathsf{E}_\theta N_n^{-1}. \tag{10}$$

2. Now, if $g(\theta)$ is given, for $\sigma^2(\theta)$ we consider the estimator

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - g(\theta))^2. \tag{11}$$

This estimator is unbiased and consistent, and its variance is

$$\overline{R}_n^*(\theta) = \mathsf{D}_\theta \overline{T}_n = \frac{\mu_4(\theta) - \sigma^4(\theta)}{n}, \tag{12}$$

where $\mu_4(\theta) = \mathsf{E}_\theta(X_1 - g(\theta))^4$. For this estimator based on a sample with random size we have

$$\overline{R}_n(\theta) = \mathsf{D}_\theta \overline{T}_{N_n} = (\mu_4(\theta) - \sigma^4(\theta))\mathsf{E}_\theta N_n^{-1}. \tag{13}$$

3. In the preceding example suppose that $g(\theta)$ is unknown and instead of (11) we consider any estimator of the form

$$\widetilde{T}_n \equiv \widetilde{T}_n(\gamma) = \frac{1}{n+\gamma}\sum_{i=1}^{n}(X_i - T_n)^2, \qquad \gamma \in \mathbb{R}, \tag{14}$$

with T_n defined in (8). If $\gamma \neq -1$, this estimator is not unbiased but may have smaller expected squared error than the unbiased estimator with $\gamma = -1$. One easily obtains (see [9, (3.6)])

$$\begin{aligned}\widetilde{R}_n^*(\theta) &= \mathsf{E}_\theta(\widetilde{T}_n - \sigma^2(\theta))^2 \\ &= \frac{\sigma^4(\theta)}{n(n+\gamma)^2}\left[(n-1)\left(\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)(n-1) + 2\right) + n(\gamma+1)^2\right]\end{aligned}$$

and hence

$$\begin{aligned}\widetilde{R}_n^*(\theta) = \sigma^4(\theta)&\left[\frac{1}{n}\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right) + \frac{\gamma+1}{n^2}\left((\gamma+1) + 2 - 2\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\right)\right] \\ &+ O(n^{-3}).\end{aligned} \tag{15}$$

Using Theorem 1 we have

$$\widetilde{R}_n(\theta) = \mathsf{E}_\theta(\widetilde{T}_{N_n} - \sigma^2(\theta))^2 = \sigma^4(\theta)\Bigg[\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\mathsf{E}_\theta N_n^{-1} + (\gamma+1)\left((\gamma+1) + 2 - 2\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\right)\mathsf{E}_\theta N_n^{-2}\Bigg] + O(\mathsf{E}_\theta N_n^{-3}). \tag{16}$$

2.3. Deficiencies of estimators for Poisson-distributed sample size. When the deficiencies of statistical estimators constructed from samples of random size $N_{m(n)}$ and the corresponding estimators constructed from samples of non-random size n (under the condition $\mathsf{E}_\theta N_n = n$) are evaluated, we actually compare the expected size $m(n)$ of a random sample with n by means of the quantity $d_n = m(n) - n$ and its limit value.

We will now apply the results of Section 2.2 to the three examples. We begin with the case of Poisson-distributed sample size. Let M_n be the Poisson r.v. with parameter $n-1$, $n \geq 2$, i.e.

$$\mathsf{P}_\theta(M_n = k) = e^{1-n}\frac{(n-1)^k}{k!}, \qquad k = 0, 1, \ldots,$$

Define the random sample size as $N_n = M_n + 1$. Then obviously, $\mathsf{E}_\theta N_n = n$ and

$$\mathsf{E}_\theta N_n^{-1} = e^{1-n}\sum_{k=0}^{\infty}\frac{(n-1)^k}{(k+1)!} = \frac{1-e^{1-n}}{n-1}.$$

Expanding the exponent in a Taylor series, we easily obtain

$$\mathsf{E}_\theta N_n^{-1} = \frac{1}{n} + \frac{1}{n^2} + o(n^{-2}). \tag{17}$$

The deficiency d of T_{N_n} relative to T_n (see (8)) is given by (9), (10), (17) and (7) with $r = s = 1$, $a(\theta) = \sigma^2(\theta)$, $b(\theta) = 0$, $c(\theta) = \sigma^4(\theta)$, and hence

$$d = 1.$$

Similarly, the deficiency $\overline{d}$ of $\overline{T}_{N_n}$ relative to $\overline{T}_n$ (see (11)) is given by (12), (13), (17) and (7) with $r = s = 1$, $a(\theta) = c(\theta) = \mu_4(\theta) - \sigma^4(\theta)$, $b(\theta) = 0$, and hence

$$\overline{d} = 1.$$

Now consider the third example (see (14)). We have

$$\mathsf{E}_\theta N_n^{-2} = e^{1-n}\sum_{k=0}^{\infty}\frac{(n-1)^k}{(k+1)^2 k!} = \frac{e^{1-n}}{n-1}\sum_{k=1}^{\infty}\frac{(n-1)^k}{kk!} = \frac{e^{1-n}}{n-1}\int_0^{n-1}\frac{e^x-1}{x}\,dx.$$

Using the Bernoulli–L'Hôpital principle we obtain

$$\int_0^{n-1}\frac{e^x-1}{x}\,dx \sim \frac{e^{n-1}}{n-1}, \qquad n \to \infty,$$

and

$$\mathsf{E}_\theta N_n^{-2} \sim \frac{1}{n^2}, \qquad n \to \infty. \tag{18}$$

Now the deficiency $\widetilde{d}$ of $\widetilde{T}_{N_n}$ with respect to $\widetilde{T}_n$ (see (14)) is given by (15), (16), (18) and (7) with $r = s = 1$, and hence

$$\widetilde{d} = 1,$$

whereas the deficiency $\widetilde{d}_{\gamma_1,\gamma_2}$ of $\widetilde{T}_{N_n}(\gamma_1)$ with respect to $\widetilde{T}_{N_n}(\gamma_2)$ (see (14)) is given by (17), (18) and (7) with $r = s = 1$, and hence

$$\widetilde{d}_{\gamma_1,\gamma_2} = (\gamma_1 - \gamma_2)\left(\frac{\gamma_1 + \gamma_2 + 2}{\mu_4(\theta)/\sigma^4(\theta) - 1} - 2\right).$$

Thus, the classical $\widetilde{T}_{N_n}(0)$ is better than $\widetilde{T}_{N_n}(-1)$ if

$$\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1 > \frac{1}{2},$$

with the situation reversed if

$$\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1 < \frac{1}{2}.$$

In particular, if X_1 is normal, then

$$\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1 = 2 \quad \text{and} \quad \widetilde{d}_{\gamma_1,\gamma_2} = \frac{1}{2}(\gamma_1 - \gamma_2)(\gamma_1 + \gamma_2 - 2).$$

One can therefore save an expected 3/2 observations by using the biased estimator $\widetilde{T}_{N_n}(0)$. The best value of γ in the normal case is $\gamma = 1$, for which $\widetilde{d}_{0,1} = 2$ and which therefore provides an additional saving of 1/2 observations.

These examples illustrate the following statement.

THEOREM 2. *Assume that there exist numbers $a(\theta)$, $b(\theta)$ and k_1, k_2 such that*

$$R_n^*(\theta) = \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + o(n^{-2})$$

and

$$\mathsf{E}_\theta N_n^{-1} = \frac{1}{n} + \frac{k_1}{n^2} + o(n^{-2}), \qquad \mathsf{E}_\theta N_n^{-2} = \frac{k_2}{n^2} + o(n^{-2}), \qquad \mathsf{E}_\theta N_n^{-3} = o(n^{-2}).$$

Then the asymptotic deficiency of T_{N_n} with respect to T_n is equal to

$$d(\theta) = \frac{k_1 a(\theta) + b(\theta)(k_2 - 1)}{a(\theta)}.$$

This follows from Theorem 1, (6) and (7).

2.4. Deficiencies of estimators for binomially distributed sample size. In this section the results obtained above will be applied to the calculation of the deficiencies of the estimators T_n, $\overline{T}_n$, $\widetilde{T}_n$ (see (8), (11) and (14)) constructed from samples whose size is binomially distributed.

Using the definition of the binomial distribution we directly obtain the following statement.

LEMMA 2. *Let the r.v. B_n have the binomial distribution with parameters $m(n-1)$, $n \geq 2$ and $p = 1/m$, where $m \geq 2$ is a fixed natural number. Define the r.v. N_n as*

$$N_n = B_n + 1.$$

Then, as $n \to \infty$,

$$\mathsf{E}_\theta N_n = n, \qquad \mathsf{E}_\theta N_n^{-1} = \frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3}),$$
$$\mathsf{E}_\theta N_n^{-3/2} = \frac{1}{n^{3/2}} + O(n^{-5/2}), \qquad \mathsf{E}_\theta N_n^{-2} = \frac{1}{n^2} + O(n^{-3}),$$
$$\mathsf{E}_\theta N_n^{-5/2} = O(n^{-3}), \qquad \mathsf{E}_\theta N_n^{-3} = O\biggl(\frac{(1-1/m)^n}{n+1}\biggr).$$

Lemma 2 and relations (10), (13) and (16) yield the following result.

THEOREM 3. *Let the r.v. B_n have the binomial distribution with parameters $m(n-1)$, $n \geq 2$ and $p = 1/m$, where $m \geq 2$ is a fixed natural number. Put $N_n = B_n + 1$. Then*

$$R_n(\theta) = \mathsf{E}_\theta(T_{N_n} - g(\theta))^2 = \sigma^2(\theta)\biggl(\frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3})\biggr),$$
$$\overline{R}_n(\theta) = \mathsf{E}_\theta(\overline{T}_{N_n} - \sigma^2(\theta))^2 = (\mu_4(\theta) - \sigma^4(\theta))\biggl(\frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3})\biggr),$$
$$\begin{aligned}\widetilde{R}_n(\theta) &= \mathsf{E}_\theta(\widetilde{T}_{N_n} - \sigma^2(\theta))^2 \\ &= \sigma^4(\theta)\biggl\{\frac{1}{n}\biggl(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\biggr) \\ &\quad + \frac{1}{n^2}\biggl[(\gamma+1)^2 + 2\biggl(\frac{m-1}{m} - 2\gamma - 1\biggr)\biggl(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\biggr)\biggr]\biggr\} + O(n^{-3}).\end{aligned}$$

COROLLARY 2. *Under the assumptions of Theorem 3 the asymptotic deficiency of the estimators T_{N_n}, $\overline{T}_{N_n}$ and $\widetilde{T}_{N_n}$ with respect to the corresponding estimators T_n, $\overline{T}_n$ and $\widetilde{T}_n$ is*

$$d = \frac{m-1}{m}.$$

2.5. Deficiencies of estimators for sample size having a three-point symmetric distribution. In this section we will consider the case

where the random sample size N_n has the symmetric distribution of the form

$$\mathsf{P}_\theta(N_n = n - h_n) = \mathsf{P}_\theta(N_n = n) = \mathsf{P}_\theta(N_n = n + h_n) = 1/3, \tag{19}$$

where the sequence of natural numbers $h_n < n$ satisfies the condition

$$\lim_{n\to\infty} \frac{h_n}{n} = 0, \tag{20}$$

that is, $h_n = o(n)$ as $n \to \infty$. It is easy to see that (19) and (20) imply that $N_n/n \to 1$ in probability as $n \to \infty$.

LEMMA 3. *Let the r.v. N_n have distribution* (19) *under condition* (20). *Then $\mathsf{E}_\theta N_n = n$ and, as $n \to \infty$,*

$$\mathsf{E}_\theta \frac{1}{N_n} = \frac{1}{n} + \frac{2}{3n}\left(\frac{h_n}{n}\right)^2 + O\left(\frac{1}{n}\left(\frac{h_n}{n}\right)^4\right),$$

$$\mathsf{E}_\theta \frac{1}{N_n^{3/2}} = \frac{1}{n^{3/2}} + O\left(\frac{1}{n^{3/2}}\left(\frac{h_n}{n}\right)^2\right),$$

$$\mathsf{E}_\theta \frac{1}{N_n^{2}} = \frac{1}{n^{2}} + O\left(\frac{1}{n^{2}}\left(\frac{h_n}{n}\right)^2\right),$$

$$\mathsf{E}_\theta \frac{1}{N_n^{5/2}} = \frac{1}{n^{5/2}} + O\left(\frac{1}{n^{5/2}}\left(\frac{h_n}{n}\right)^2\right),$$

$$\mathsf{E}_\theta \frac{1}{N_n^{3}} = \frac{1}{n^{3}} + O\left(\frac{1}{n^{3}}\left(\frac{h_n}{n}\right)^2\right).$$

This follows from the easily verified equalities

$$\begin{aligned}
\mathsf{E}_\theta \frac{1}{N_n} &= \frac{3n^2 - h_n^2}{3n(n^2 - h_n^2)} = \frac{1}{n}\left(1 - \frac{h_n^2}{3n}\right)\left(1 + \frac{h_n^2}{n^2} + O\left(\frac{h_n^4}{n^4}\right)\right) \\
&= \frac{1}{n} + \frac{2}{3n}\left(\frac{h_n}{n}\right)^2 + O\left(\frac{1}{n}\left(\frac{h_n}{n}\right)^4\right), \\
\mathsf{E}_\theta \frac{1}{N_n^{3/2}} &= \frac{1}{3n^{3/2}}\left(\frac{1}{(1 - h_n/n)^{3/2}} + 1 + \frac{1}{(1 + h_n/n)^{3/2}}\right) \\
&= \frac{1}{n^{3/2}} + O\left(\frac{1}{n^{3/2}}\left(\frac{h_n}{n}\right)^2\right), \\
\mathsf{E}_\theta \frac{1}{N_n^{2}} &= \frac{1}{3n^{2}}\left(\frac{1}{(1 - h_n/n)^{2}} + 1 + \frac{1}{(1 + h_n/n)^{2}}\right) = \frac{1}{n^2} + O\left(\frac{1}{n^2}\left(\frac{h_n}{n}\right)^2\right).
\end{aligned}$$

The asymptotic formulas for $\mathsf{E}_\theta N_n^{-5/2}$ and $\mathsf{E}_\theta N_n^{-3}$ are established in a similar way. ∎

This lemma and formulas (10), (13) and (16) directly imply

THEOREM 4. *Let the r.v. N_n have distribution* (19) *under condition* (20). *Then*

$$R_n(\theta) = \mathsf{E}_\theta(T_{N_n} - g(\theta))^2 = \sigma^2(\theta)\left(\frac{1}{n} + \frac{2h_n^2}{3n^3}\right) + o(n^{-2}),$$

$$\overline{R}_n(\theta) = \mathsf{E}_\theta(\overline{T}_{N_n} - \sigma^2(\theta))^2 = (\mu_4(\theta) - \sigma^4(\theta))\left(\frac{1}{n} + \frac{2h_n^2}{3n^3}\right) + o(n^{-2}),$$

$$\begin{aligned}\widetilde{R}_n(\theta) &= \mathsf{E}_\theta(\widetilde{T}_{N_n} - \sigma^2(\theta))^2 \\ &= \sigma^4(\theta)\left\{\frac{1}{n}\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\right. \\ &\quad + \frac{1}{n^2}\left[2 + (\gamma+1)\left(\gamma + 1 - 2\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\right)\right] \\ &\quad \left. + \frac{2h_n^2}{3n^3}\left(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1\right)\right\} + o(n^{-2}).\end{aligned}$$

COROLLARY 3. *Suppose the assumptions of Theorem* 4 *hold and*

$$h_n^2/n \to h > 0, \qquad n \to \infty.$$

Then the asymptotic deficiency d of the estimators T_{N_n}, $\overline{T}_{N_n}$ and $\widetilde{T}_{N_n}$ with respect to the corresponding estimators T_n, $\overline{T}_n$ and $\widetilde{T}_n$ is

$$d = 2h/3.$$

It is worth noting that in Corollary 3, h can be arbitrarily large. Therefore the *finite* asymptotic deficiency d considered in Corollary 3 can be arbitrarily large. This is in full correspondence with the conclusion of Section 2.1.

Acknowledgements. Sections 1.1, 1.3 and 2.1 were written by V. Korolev and A. Zeifman who were financially supported by the Russian Science Foundation, project 18-11-00155.

REFERENCES

[1] V. E. Bening, *Asymptotic Theory of Testing Statistical Hypotheses: Efficient Statistics*, Optimality, Power Loss, and Deficiency, De Gruyter, Berlin, 2011.

[2] V. E. Bening and V. Yu. Korolev, *On an application of the Student distribution in the theory of probability and mathematical statistics*, Theory Probab. Appl. 49 (2005), 377–391.

[3] V. E. Bening and V. Yu. Korolev, *Some statistical problems related to the Laplace distribution*, Informatics Appl. 2 (2008), 19–34.

[4] V. E. Bening and V. Yu. Korolev, *Generalized Poisson Models and their Applications in Insurance and Finance*, De Gruyter, Berlin, 2012.

[5] V. E. Bening, V. Yu. Korolev, V. A. Savushkin and A. I. Zeifman, *On the deficiency of some estimators constructed from samples with random sizes*, AIP Conf. Proc. 1648 (2015), 250009.
[6] H. Cramér, *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton, 1946.
[7] B. V. Gnedenko, *On estimation of unknown parameters from a random number of independent observations*, Trans. Razmadze Tbilisi Math. Inst. 92 (1989), 146–150.
[8] B. V. Gnedenko and V. Yu. Korolev, *Random Summation. Limit Theorems and Applications*, CRC Press, Boca Raton, FL, 1996.
[9] J. L. Hodges and E. L. Lehmann, *Deficiency*, Ann. Math. Statist. 41 (1970), 783–801.
[10] V. Yu. Korolev, *Convergence of random sequences with independent random indices. II*, Theory Probab. Appl. 40 (1995), 907–910.
[11] V. Yu. Korolev, *A general theorem on the limit behavior of superpositions of independent random processes with applications to Cox processes*, J. Math. Sci. 81 (1996), 2951–2956.
[12] V. Yu. Korolev and A. I. Zeifman, *On convergence of the distributions of random sequences with independent random indexes to variance-mean mixtures*, Stochastic Models 32 (2016), 414–432.

V. E. Bening
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
and
Institute of Informatics Problems
Federal Research Center «Informatics and Control»
Russian Academy of Sciences
E-mail: bening@yandex.ru

V. Yu. Korolev
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
and
Institute of Informatics Problems
Federal Research Center «Informatics and Control»
Russian Academy of Sciences
and
Hangzhou Dianzi University
E-mail: victoryukorolev@yandex.ru

A. I. Zeifman (corresponding author)
Vologda State University
Institute of Informatics Problems
Federal Research Center «Informatics and Control»
Russian Academy of Sciences
and
Vologda Research Center
Russian Academy of Sciences
E-mail: a_zeifman@mail.ru