On the asymptotic behavior of the deficiency of some statistical estimators based on samples with random sizes.

Bening VE*

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia

Abstract

Due to the stochastic character of the intensities of information flows in high performance information systems, the size of data available for the statistical analysis can be often regarded as random. The purpose of this paper is to present some means for the comparison of the quality of estimators constructed from samples with random sizes with that of estimators constructed from samples with non-random sizes. As this means it is proposed to use the deficiency. It can be an illustrative characteristic of a possible loss of the 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 whose expectation equals *n* with respect to the same statistic constructed as if the sample size was 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 considered in the paper where the deficiencies of statistics constructed from samples whose sizes have the Poisson, binomial and special three-point distributions, respectively, are considered. Some basic results dealing with some properties of estimators based on the samples with random sizes are also presented.

Keywords: Estimator, Risk function, Deficiency, Asymptotic deficiency, Sample with random size, Asymptotic expansions, Poisson distribution, Binomial distribution, Three-point distribution.

Accepted on October 26, 2018, 2018

Introduction

Motivation for the consideration of statistics constructed from samples with random sizes

In most cases related to the analysis of experimental data, the number of random factors which influence observed objects is random and changes from one observation to anorher. Due to the stochastic character of the intensities of information flows in high performance information systems, the size of data available for the statistical analysis can be often regarded as random. In classical problems of mathematical statistics, the size of the available sample, i. e., the number of available observations, is traditionally assumed to be deterministic. In the asymptotic settings it plays the role of infinitely increasing known parameter. At the same time, in practice very often the data to be analyzed is collected or registered during a certain period of time and the flow of informative events each of which brings a next observation forms a random point process. Therefore, the number of available observations is unknown till the end of the process of their registration and also must 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 or insurance 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 the statistical analysis can be often regarded as random. Say, the statistical algorithms applied in high-frequency financial applications must take into consideration that the number of events in a limit order book during a time unit essentially depends on the intensity of order flows. Moreover, contemporary statistical procedures of insurance and financial mathematics do take this circumstance into consideration as one of 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 available 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 sizes are special cases of random sequences with random indices. The randomness of indices usually leads to that the limit distributions for the corresponding random sequences are heavy-tailed even in the situations where the distributions of non-randomly indexed random sequences are asymptotically normal [1-3]. For example, if a statistic which is asymptotically normal in the traditional sense, is constructed on the basis of a sample with random size having negative binomial distribution, then instead of the expected normal law, the Student distribution with power-type decreasing heavy tails appears as an asymptotic law for this statistic [1,4].

At the same time, according to the conventional logics of the statistical analysis, the distributions of the statistics (estimators, tests, etc.) to be used for the 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 the exact distributions because the former are considerably easier computable than the latter. As this is so, in limit theorems of probability theory and mathematical statistics the centering and normalization of random variables are used to obtain non-trivial asymptotic distributions. It should be especially noted that to obtain reasonable approximation to the distribution of the 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: $\mathbb{R}$ is the set of real numbers, $\mathbb{N}$ is the set of natural numbers, *h(n) ~ f(n)*, $n \to \infty$ if and only if $\lim_{n\to\infty} h(n)/f(n) = 1$. The symbols $\overset{d}{=}$, $\Rightarrow$ and $\square$ denote the coincidence of distributions, convergence in distribution and the end of the proof, respectively.

Consider a family of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ each of which is defined on a measurable space $(\Omega, \mathfrak{F})$. Consider a sequence of random variables (r.v.'s) $X_1, X_2, \ldots$ defined on a measurable space $(\Omega, \mathfrak{F})$. Everywhere in what follows consider the random variables $X_1, X_2, \ldots$ to be independent and identically distributed (i.i.d) with common distribution P_θ. Let $N_1, N_{2,\ldots}$ be a sequence of nonnegative integer random variables with common distribution *P* defined on the same measurable space so that for each $n \geq 1$ the random variable N_n is independent of the sequence $X_1, X_2, \ldots$ with respect to any measure P_θ from $\mathcal{P}$. A random sequence $N_1, N_{2,\ldots}$ (N_i with distribution *P*, *i = 1, 2,...*) is said to be infinitely increasing ($N_n \to \infty$) in probability *P*, if $P(N_n \leq M) \to 0$ as $n \to \infty$ for any M ϵ $(0, \infty)$. For $n \geq 1$, let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic, that is, a measurable function of the r.v.'s $X_1, \ldots, X_n$. For each $n \geq 1$ define the r.v. T_{Nn} by letting:

$$T_{N_n}(\omega) = T_{N_n(\omega)}\left(X_1(\omega), \ldots, X_{N_n(\omega)}(\omega)\right)$$

for every elementary outcome $\omega \,\epsilon\, \Omega$. Assume that for each $\theta \,\epsilon\, \Theta$ there exists:

$$E_\theta \, T_n \equiv g(\theta),$$

where $E_\theta \equiv E_{\theta,n}$ is the expectation w.r.t. distribution $P_\theta \equiv P_{\theta,n}$ of T_n. We will say that the statistic T_n is asymptotically normal,

$$T_n \sim N(g(\theta), \sigma^2(\theta)), \sigma^2(\theta) > 0, \ n \to \infty,$$

if

$$P_\theta\left(\sqrt{n}\,\sigma(\theta)(T_n - g(\theta)) < x\right) \Rightarrow \Phi(x), \quad n \to \infty (1.1)$$

for each $\theta \,\epsilon\, \Theta$.

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. (Theorem 3.3.2) [5].

Lemma 1.1. *Assume that $N_n \to \infty$ in probability P as $n \to \infty$. Let the statistic T_n be asymptotically normal in the sense of (1.1). Then a distribution function $F(x)$ such that*

$$P_\theta\left(\sqrt{n}\,\sigma(\theta)(T_{N_n} - g(\theta)) < x\right) \Rightarrow F(x), \quad n \to \infty,$$

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

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

The concept of deficiency

Before turning to the general case of statistics constructed from samples with random size, that is the main aim of the present paper, let us recall the notion of a deficiency of a statistical estimator for the traditional case where the sample size is non-random [6].

Suppose that $T_n(X_1, \ldots, X_n)$ and $T_n(X_1, \ldots, X_n)$ are two competing estimators of $g(\theta)$, $\theta \,\epsilon\, \Theta$ based on *n* observations $X_1, \ldots, X_n$ 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 estimator $T_{m(n)}(X_1, \ldots, X_{m(n)})$ to match the performance of $T_n^*(X_1, \ldots, X_n)$ (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 very 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 [5]. They name *m(n) - n* the deficiency of T_n with respect to T_n^* and denote it as:

$$d_n = m(n) - n.(1.2)$$

Suppose that for $n \to \infty$, the ratio *n / m(n)* tends to a limit *b*, the asymptotic relative efficiency of T_n $(X_1, \ldots, X_n)$ with respect to $T_n^*(X_1, \ldots, X_n)$. 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 be varying 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^* will then indicate 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:

$$R_n(\theta) = E_\theta(T_n - g(\theta))^2, \quad R_n^*(\theta) = E_\theta(T_n^* - g(\theta))^2,$$

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

$$R_n^*(\theta) = R_{m(n)}(\theta).(1.3)$$

In order to solve (1.3), *m*(*n*) has to be treated as a continuous variable. This can be done in a satisfactory manner by defining $R_{m(n)}(\theta)$ for non-integer *m*(*n*) as:

$$R_{m(n)}(\theta) = (1 - m(n) + [m(n)])R_{[m(n)]}(\theta) + (m(n) - [m(n)])R_{[m(n)]+1}(\theta) \ [6].$$

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:

$$R_n^* = \frac{a(\theta)}{n^r} + \frac{b(\theta)}{n^{r+s}} + o(n^{-(r+s)}),(1.4)$$

$$R_n = \frac{a(\theta)}{n^r} + \frac{c(\theta)}{n^{r+s}} + o(n^{-(r+s)}),(1.5)$$

for certain $a(\theta)$, $b(\theta)$ and $c(\theta)$ not depending on n and certain constants $r > 0$, $s > 0$. The leading term in both expansions is the same in view of the fact that ARE is equal to one. From (1.2) - (1.5) is now easily follows that [6]

$$d_n(\theta) = \frac{c(\theta) - b(\theta)}{r\, a(\theta)} n^{(1-s)} + o(n^{(1-s)}).(1.6)$$

Hence,

$$d(\theta) = d = \begin{cases} \pm\infty, \quad 0 < s < 1, 1mm \dfrac{c(\theta) - b(\theta)}{ra(\theta)}, \quad s = 1, 1mm 0, \quad s > 1.(1.7) \end{cases}$$

A useful property of deficiencies is the following (transitivity): if a third estimator $\overline{T}_n$ is given, for which the risk $\overline{R}_n(\theta)$ also has an expansion of the form (1.5), the deficiency d of $\overline{T}_n$ with respect to T_n^* satisfies the relation $d = d_1 + d_2$, where d_1 is the deficiency of $\overline{T}_n$ 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 one. Hodges and Lehmann [6] demonstrate the use of deficiency in a number of simple examples for which this is the case (for testing problems see also [7-10]).

The purpose and structure of the paper

The purpose of this paper is to present some means for the comparison of the quality of estimators constructed from samples with random sizes with that of estimators constructed from samples with non-random sizes. As this means we propose to use the deficiency. It can be an illustrative characteristic of a possible loss of the 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 [3] and presents a number of applications of the deficiency concept in problems of point estimation in the case when the number of observations is random.

Section 2 contains main results. First, in Section 2.1 we heuristically show that if the d.f. $Q(x)$ in Lemma 1.1 is not degenerate, then the deficiency of a statistic constructed from a sample with random size whose expectation equals n with respect to the same statistic constructed as if the sample size was 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 considered in Sections 2.3, 2.4 and 2.5 where the deficiencies of statistics constructed from samples whose sizes have the Poisson, binomial and special three-point distributions, respectively, are considered. Section 2.2 contains some preliminary basic results dealing with some properties of estimators based on the samples with random sizes. Sections 3 - 5 contain results concerning deficiencies of asymptotic quantiles.

In this paper we focus on the case where the sample size is independent of the r.v.'s forming the sample. This assumption, first, is made for the sake of simplicity of the methods used to obtain the 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 is 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,... is not crucial since basic Lemma 1.1 can be proved without this assumption [5]. Third, most papers considering non-independent sample sizes deal with the case of asymptotically degenerate indexes. 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, the results of this paper can be extended for the non-independent case.

Deficiencies of Some Estimators Based on the Samples with Random Size

The asymptotic behavior of the deficiency of a statistic constructed from a sample with random size

The interpretation of the deficiency as the number of additional observations required to attain the same quality here needs to be refined since this number becomes random in random-size-samples problems. In order to circumvent this difficulty assume that the r.v.'s N_1, N_2,... are parameterized by their expectations:

$$E\,N_n = n, \quad n \in \mathbb{N}.$$

This assumption will enable us, instead of comparing random variables, to compare their easily tractable parameters.

Before we construct the exact formulas for the deficiencies so tractable, we have to make some important heuristic comments concerning the boundedness of the deficiency as a function of the parameter n. By X without any indexes we will denote a r.v. with the standard normal distribution $N(0, 1)$. Let T_n be an asymptotically normal (1.1) (with $\sigma(\theta) = 1$) statistic constructed from the sample $X_1,\ldots,X_n$, T_{N_n} be (the same) statistic constructed from the random-size-sample $X_1,\ldots,X_{N_n}$. Assume that $E_\theta\, T_n = g(\theta)$, $n \in \mathbb{N}$, implying $n \in \mathbb{N}$, $n \in \mathbb{N}$ (Theorem 2.1). Denote,

$$R_n^*(\theta) = E_\theta\,(T_n - g(\theta))^2, \quad R_n(\theta) = E_\theta\,(T_{N_n} - g(\theta))^2.$$

From Lemma 1.1, for n large enough we have the approximate relations:

$$T_n = g(\theta) + \frac{X}{\sqrt{n}} + o(n^{-1/2}),\ T_{N_n} = g(\theta) + \frac{X}{\sqrt{Un}} + o(n^{-1/2}),$$

Where,

$$P(U < x) = Q(x), \quad x \in \mathbb{R},$$

and the r.v.'s X and U are independent. Therefore,

$$R_n^*(\theta) = E_\theta\,(\frac{X}{\sqrt{n}} + o(n^{-1/2}))^2 = E\,(\frac{X}{\sqrt{n}})^2 + o(n^{-1}) = \frac{1}{n} + o(n^{-1}),$$

$$R_n(\theta) = E_\theta\,(\frac{X}{\sqrt{Un}} + o(n^{-1/2}))^2 = E\,(\frac{X}{\sqrt{Un}})^2 + o(n^{-1}) = \frac{E\,U^{-1}}{n} + o(n^{-1}).$$

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

$$\frac{1}{n} + o(n^{-1}) = \frac{E\,U^{-1}}{(n + d_n)} + o((n + d_n)^{-1})$$

or

$$\frac{d_n}{n} = D + o(1), \quad n \to \infty,$$

Where,

$$D = E\,U^{-1} - 1.$$

So, in general, if $EU^{-1} \geq 1$, then $d_n = O(n)$. And the only possibility for d_n to be $o(n)$ and, in particular, to remain bounded, is the case:

$$E\,U^{-1} = 1.$$

In general, if in addition to the conditions of Lemma 1.1, the family $\{N_n / n\}_{n \geq 1}$ is uniformly integrable, then the conditions of Lemma 1.1 and $E\,N_n = n$ imply that $EU = 1$, so that by the Jensen inequality we have $EU^{-1} \geq 1$ with the equality attainable if and only if

$$P\,(U = 1) = 1.$$

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

$$\frac{N_n}{n} \to 1$$

in probability as $n \to \infty$. This property is inherent in 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 as the exponential distribution vividly illustrates the possibility of the deficiency to be unbounded since in this case the Fréchet distribution of the r.v. U^{-1} has the infinite first moment.

Summarizing the abovesaid we conclude that if the d.f. $Q(x)$ in Lemma 1.1 is not degenerate, then the deficiency of a statistic constructed from a sample with random size whose expectation equals n with respect to the same statistic constructed as if the sample size was 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.

Some properties of estimators based on the samples with random sizes

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 independent and identically distributed with distribution depending on $\theta \in \Theta \in \mathbb{R}$.

Recall that we assume that,

$$E\,N_n = n,$$

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 2.1.

1. *If*

$$E_\theta\,T_n = g(\theta), \quad \theta \in \Theta,$$

Then,

$$E_\theta\,T_{N_n} = g(\theta), \;\; \theta \in \Theta.$$

2. Let

$$R_n^*(\theta) = E_\theta\,(T_n - g(\theta))^2, \;\; R_n(\theta) = E_\theta\,(T_{N_n} - g(\theta))^2.$$

Assume that there exist numbers $a(\theta)$, $b(\theta)$, $C(\theta) > 0$, $\alpha > 0$, $r > 0$ and $s > 0$ such that

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

Then,

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

Proof: The desired relations can be easily obtained by the formula of total probability formula. Namely, we obviously have

$$E_\theta\,T_{N_n} = \sum_{k=1}^{\infty} E_\theta\,T_k\,P(N_n = k) = \sum_{k=1}^{\infty} g(\theta)P(N_n = k) =$$

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

and

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

$$= |\sum_{k=1}^{\infty} E_\theta\,(T_k - g(\theta))^2 P(N_n = k) - a(\theta)\sum_{k=1}^{\infty}\frac{P(N_n = k)}{k^r} - b(\theta)\sum_{k=1}^{\infty}\frac{P(N_n = k)}{k^{r+s}}\,| =$$

$$= |\sum_{k=1}^{\infty}[E_\theta\,(T_k - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}}]\,P(N_n = k)\,| \leqslant$$

$$\leqslant \sum_{k=1}^{\infty}|\,E_\theta\,(T_k - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}}\,|\,P(N_n = k) \leqslant$$

$$\leqslant \sum_{k=1}^{\infty}\frac{C(\theta)}{k^{r+s+\alpha}}P(N_n = k) = C(\theta)E\,N_n^{-r-s-\alpha}. \quad \square$$

Corollary 2.1. *Let* $R_n^*(\theta) = E_\theta\,(T_n - g(\theta))^2$, $R_n(\theta) = E_\theta\,(T_{N_n} - g(\theta))^2$ *. Assume that there exist numbers* $a(\theta)$, $b(\theta)$, $r > 0$ *and* $s > 0$ *such that*

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

Then,

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

Consider some examples.

1. Let observations $X_1, \ldots, X_n$ have expectation $E_\theta X_1 = g(\theta)$ and variance $D_\theta X_1 = \sigma^2(\theta)$. The customary estimator for $g(\theta)$ based on n observation is

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i.(2.1)$$

This estimator is unbiased and consistent, and its variance is

$$R_n^*(\theta) = D_\theta\,T_n = \frac{\sigma^2(\theta)}{n}.(2.2)$$

If this estimator is based on the sample with random size, then we have (see Corollary 2.1)

$$R_n(\theta) = D_\theta\,T_{N_n} = \sigma^2(\theta)\,E\,N_n^{-1}.(2.3)$$

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

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - g(\theta))^2.(2.4)$$

This estimator is unbiased and consistent, and its variance is

$$\overline{R}^*_n(\theta) = D_\theta\, \overline{T}_n = \frac{\mu_4(\theta) - \sigma^4(\theta)}{n}, (2.5)$$

Where, $\mu_4(\theta) = (X_1 - g(\theta))^4$. For this estimator based on a sample with random size we have

$$\overline{R}_n(\theta) = D_\theta\, \overline{T}_{N_n} = (\mu_4(\theta) - \sigma^4(\theta))\, E\, N_n^{-1}.(2.6)$$

3. In the preceding example suppose that $g(\theta)$ is unknown and instead of (2.4) we consider any estimator of the form

$$\widetilde{T}_n \equiv \widetilde{T}_n(\gamma) = \frac{1}{n+\gamma} \sum_{i=1}^{n} (X_i - T_n)^2, \quad \gamma \in \mathbb{R}.(2.7)$$

with T_n defined in (2.1). If $\gamma \neq$ -1, this estimator is not unbiased but may have a less expected squared error than the unbiased estimator with γ = -1. One easily obtains (3.6) [6].

$$\widetilde{R}^*_n(\theta) = E_\theta\, (\widetilde{T}_n - \sigma^2(\theta))^2 = \frac{\sigma^4(\theta)}{n(n+\gamma)^2}[(n-1)((\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1)(n-1) + 2) + n(\gamma+1)^2]$$

and hence,

$$\widetilde{R}^*_n(\theta) = \sigma^4(\theta)\,[\frac{1}{n}(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1) + \frac{\gamma+1}{n^2}((\gamma+1) + 2 - 2(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1))] + O(n^{-3}).(2.8)$$

Using Theorem 2.1 we have

$$\widetilde{R}_n(\theta) = E_\theta\, (\widetilde{T}_{N_n} - \sigma^2(\theta))^2 = \sigma^4(\theta)[(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1)\, E\, N_n^{-1} + .$$

$$. + (\gamma+1)((\gamma+1) + 2 - 2(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1))E\, N_n^{-2}] + O(E\, N_n^{-3}).(2.9)$$

Deficiencies of some estimators based on samples with random size having the Poisson distribution

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 $EN_n = n$) 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 the Poisson-distributed sample size. Let M_n be the Poisson r.v. with parameter $n - 1$, $n \geq 2$, i.e.

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

Define the random sample size as $N_n = M_n + 1$. Then, obviously, $EN_n = n$ and

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

Expanding the exponent in the Taylor series, we easily obtain that

$$E\, N_n^{-1} = \frac{1}{n} + \frac{1}{n^2} + o(n^{-2}).(2.10)$$

The deficiency of T_{N_n} relative to T_n (see (2.1)) is given by (2.2), (2.3), (2.10) and (1.7) with $r = s = 1$, $a(\theta) = \sigma^2(\theta)$, $b(\theta) = 0$, $c(\theta) = \sigma^4(\theta)$, and hence, is equal to

$$d = 1.$$

Similarly, the deficiency of $\overline{T}_{N_n}$ relative to $\overline{T}_n$ (see (2.4)) is given by (2.5), (2.6), (2.10) and (1.7) with $r = s = 1$, $a(\theta) = c(\theta) = \mu_4(\theta) - \sigma^4(\theta)$, $b(\theta) = 0$, and hence, is equal to

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

Now consider the third example (see (2.7)). We have

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

Using the Bernoulli – L'H$\hat{o}$pital principle we obtain

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

and

$$E\, N_n^{-2} \sim \frac{1}{n^2}, \quad n \to \infty.(2.11)$$

Now the deficiency of $\widetilde{T}_{N_n}$ with respect to $\widetilde{T}_n$ (see (2.7)) is given by (2.8), (2.9), (2.11) and (1.7) with $r = s = 1$ and hence, is equal to

$$\tilde{\ }$$

whereas the deficiency of $\widetilde{T}_{N_n}(\gamma_1)$ with respect to $\widetilde{T}_{N_n}(\gamma_2)$ (see (2.7)) is given by (2.10), (2.11) and (1.7) with $r = s = 1$ and hence, is equal to

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

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

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

and

$$\tilde{d}_{\gamma_1,\gamma_2} = \frac{1}{2}(\gamma_1 - \gamma_2)(\gamma_1 + \gamma_2 - 2).$$

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 γ = 1 for which $\tilde{d}_{0,1} = 2$ and which therefore provides an additional saving 1 / 2 observations.

These examples illustrate the following statement.

Theorem 2.2. *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

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

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

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

The proof follows from Theorem 2.1, (1.6) and (1.7).

Deficiencies of some estimators based on samples with random size having the binomial distribution

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

(2.1), (2.4) and (2.7)) constructed from samples whose sizes are random and have the binomial distribution.

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

Lemma 2.1. *Let the r.v. B_n have the binomial distribution with the 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 $n \longrightarrow \infty$,

$$E\,N_n = n,\quad E\,N_n^{-1} = \frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3}),\quad E\,N_n^{-3/2} = \frac{1}{n^{3/2}} + O(n^{-5/2}),$$

$$R_n(\theta) = E_\theta\,(T_{N_n} - g(\theta))^2 = \sigma^2(\theta)(\frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3})),$$

Lemma 2.1 and relations (2.3), (2.6) and (2.9) yield the following result.

Theorem 2.3. *Let the r.v. B_n have the binomial distribution with the 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,*

$$R_n(\theta) = E_\theta\,(T_{N_n} - g(\theta))^2 = \sigma^2(\theta)(\frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3})),$$

$$\overline{R}_n(\theta) = E_\theta\,(\overline{T}_{N_n} - \sigma^2(\theta))^2 = (\mu_4(\theta) - \sigma^4(\theta))(\frac{1}{n} + \frac{m-1}{mn^2} + O(n^{-3})),$$

$$\widetilde{R}_n(\theta) = E_\theta\,(\widetilde{T}_{N_n} - \sigma^2(\theta))^2 = \sigma^4(\theta)\{\frac{1}{n}(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1) + .$$

$$. + \frac{1}{n^2}[(\gamma+1)^2 + 2(\frac{m-1}{m} - 2\gamma - 1)(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1)]\} + O(n^{-3}).$$

Corollary 2.2. *Under the conditions of Theorem* 2.3 *the asymptotic deficiencies 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 $\widetilde{T}_n$ has the form*

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

Deficiencies of some estimators based on samples with random 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

$$P\,(N_n = n - h_n) = P\,(N_n = n) = P\,(N_n = n + h_n) = \frac{1}{3}, (2.12)$$

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

$$\lim_{n\to\infty} \frac{h_n}{n} = 0, (2.13)$$

that is, $h_n = o(n)$ as $n \longrightarrow \infty$. It is easy to see that (2.12) and (2.13 imply that $N_n / n \longrightarrow 1$ in probability as $n \longrightarrow \infty$.

Lemma 2.2. *Let the r.v. N_n have distribution* (2.12) *under condition* (2.13). *Then $EN_n = n$ and, as $n \longrightarrow \infty$,*

$$E\,N_n^{-1} = \frac{1}{n} + \frac{2}{3n}(\frac{h_n}{n})^2 + O(\frac{1}{n}(\frac{h_n}{n})^4),\quad E\,N_n^{-3/2} = \frac{1}{n^{3/2}} + O(\frac{1}{n^{3/2}}(\frac{h_n}{n})^2),$$

$$E\,N_n^{-2} = \frac{1}{n^2} + O(\frac{1}{n^2}(\frac{h_n}{n})^2),\quad E\,N_n^{-5/2} = \frac{1}{n^{5/2}} + O(\frac{1}{n^{5/2}}(\frac{h_n}{n})^2),$$

$$E\,N_n^{-3} = \frac{1}{n^3} + O(\frac{1}{n^3}(\frac{h_n}{n})^2).$$

The proof follows from the easily verified equalities

$$E\,N_n^{-1} = \frac{3n^2 - h_n^2}{3n(n^2 - h_n^2)} =$$

$$E\,N_n^{-3/2} = \frac{1}{3n^{3/2}}(\frac{1}{(1 - h_n/n)^{3/2}} + 1 + \frac{1}{(1 + h_n/n)^{3/2}}) =$$

$$E\,N_n^{-3/2} = \frac{1}{3n^{3/2}}(\frac{1}{(1 - h_n/n)^{3/2}} + 1 + \frac{1}{(1 + h_n/n)^{3/2}}) =$$

$$= \frac{1}{n^{3/2}} + O(\frac{1}{n^{3/2}}(\frac{h_n}{n})^2),$$

$$E\,N_n^{-2} = \frac{1}{3n^2}(\frac{1}{(1 - h_n/n)^2} + 1 + \frac{1}{(1 + h_n/n)^2}) = \frac{1}{n^2} + O(\frac{1}{n^2}(\frac{h_n}{n})^2).$$

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

This Lemma and formulas (2.3), (2.6) and (2.9) directly imply the following statement.

Theorem 2.3. *Let the r.v. N_n have distribution* (2.12) *under condition* (2.13). *Then,*

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

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

$$\widetilde{R}_n(\theta) = E_\theta(\widetilde{T}_{N_n} - \sigma^2(\theta))^2 = \sigma^4(\theta)\{\frac{1}{n}(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1) +$$

$$. + \frac{1}{n^2}[2 + (\gamma+1)(\gamma + 1 - 2(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1))] + \frac{2h_n^2}{3n^3}(\frac{\mu_4(\theta)}{\sigma^4(\theta)} - 1)\} + o(n^{-2}).$$

Corollary 2.3. *Let the conditions of Theorem* 2.3 *hold and*

$$\frac{h_n^2}{n} \to h > 0,\quad n \to \infty.$$

Then 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 $\widetilde{T}_n$ has the form

$$d = \frac{2h}{3}.$$

It is worth noting that in Corollary 2.3 h can be arbitrarily large. Therefore the *finite* asymptotic deficiency d considered in Corollary 2.3 can be arbitrarily large. This is in full correspondence with the conclusion of Section 2.1.

Asymptotic Deficiency and Quantiles

For $n \geq 1$ let $T_n = T_n(X_1, ..., X_n)$ be a statistic, that is, a measurable function of the r.v.'s $X_1, ..., X_n$. The asymptotic quantile of order α, α є (0, 1) (the α – quantile) of statistic T_n is the value $c_\alpha^*(n)$ for which

$$P(\sqrt{n}\,T_n \geqslant c_\alpha^*(n)) = \alpha + o(n^{-1}),\quad n \to \infty. (3.1)$$

Using Taylor's formula one has

Lemma 3.1. *Suppose that the distribution function of $\sqrt{n}\,T_n$ satisfies (uniformly in $x \in \mathbb{R}$) the relation*

$$P(\sqrt{n}\,T_n < x) = G(x) + \frac{1}{\sqrt{n}}g_1(x) + \frac{1}{n}g_2(x) + o(n^{-1}),$$

Where, $G(x)$, $g_1(x)$, $g_2(x)$ are sufficiently smooth functions. Then,

$$c_\alpha^*(n) = c_\alpha - \frac{g_1(c_\alpha)}{\sqrt{n}\, G^{(1)}(c_\alpha)} -$$

$$-\frac{1}{n}\Big(\frac{G^{(2)}(c_\alpha)g_1^2(c_\alpha)}{2(G^{(1)}(c_\alpha))^3} + \frac{G^{(1)}(c_\alpha)g_2(c_\alpha) - g_1(c_\alpha)g_1^{(1)}(c_\alpha)}{(G^{(1)}(c_\alpha))^2}\Big) + o(n^{-1}),$$

Where, $G(c_\alpha) = 1$ - α.

Corollary 3.1. *Let $\delta_n \to 0$, $n \to \infty$. Then under the conditions of Lemma 3.1 uniformly in $x \in \mathbb{R}$*

$$P(\sqrt{n}\, T_n < x + \delta_n) =$$

$$= P(\sqrt{n}\, T_n < x) + \delta_n\, G^{(1)}(x) + \frac{\delta_n^2}{2} G^{(2)}(x) + \frac{\delta_n}{\sqrt{n}} g_1^{(1)}(x) + o(\max(\delta_n^2, \frac{\delta_n}{\sqrt{n}}, n^{-1})).$$

Now consider a statistic $S_n = S_n(X_1, ..., X_n)$ other than T_n having α – quantile $c_\alpha(n)$

$$P(\sqrt{n}\, S_n \geqslant c_\alpha(n)) = \alpha + o(n^{-1}), \quad n \to \infty.(3.2)$$

Suppose that:

$$P(\sqrt{n}\, S_n < x) = G(x) + \frac{1}{\sqrt{n}} g_1(x) + \frac{1}{n}\bar{g}_2(x) + o(n^{-1}),(3.3)$$

Where, $G(x), g_1(x), \bar{g}_2(x)$ are some smooth functions. Define the sequence of positive integers $\{m(n) = n + d + o(1),\ d \in \mathbb{R},\quad n = 1, 2, \ldots\}$ by the relation (d is the asymptotic deficiency)

$$P(\sqrt{n}\, S_{m(n)} \geqslant c_\alpha^*(m(n))) = \alpha + o(n^{-1}), \quad n \to \infty.(3.4)$$

Theorem 3.1. *Under the conditions of Lemma 3.1 and (3.3) the asymptotic deficiency d equals*

$$d = \frac{2(g_2(c_\alpha) - \bar{g}_2(c_\alpha))}{G^{(1)}(c_\alpha)\, c_\alpha} + o(1).$$

Proof. It follows from (3.1) and Lemma 3.1 that

$$c_\alpha(n) = c_\alpha - \frac{g_1(c_\alpha)}{\sqrt{n}\, G^{(1)}(c_\alpha)} -$$

$$-\frac{1}{n}\Big(\frac{G^{(2)}(c_\alpha)g_1^2(c_\alpha)}{2(G^{(1)}(c_\alpha))^3} + \frac{G^{(1)}(c_\alpha)\bar{g}_2(c_\alpha) - g_1(c_\alpha)g_1^{(1)}(c_\alpha)}{(G^{(1)}(c_\alpha))^2}\Big) + o(n^{-1})(3.5)$$

and

$$\delta_n \equiv \sqrt{\frac{m(n)}{n}}\, c_\alpha^*(m(n)) - c_\alpha(m(n)) = \frac{d}{2n} c_\alpha - \frac{1}{n} \frac{(g_2(c_\alpha) - \bar{g}_2(c_\alpha))}{G^{(1)}(c_\alpha)} + o(n^{-1}).(3.6)$$

Moreover (3.4) implies

$$\alpha + o(n^{-1}) = P(\sqrt{n}\, S_{m(n)} \geqslant c_\alpha^*(m(n))) =$$

$$= P(\sqrt{m(n)}\, S_{m(n)} \geqslant c_\alpha(m(n)) + \delta_n).(3.7)$$

Using Corollary 3.1 we obtain

$$\alpha + o(n^{-1}) = P(\sqrt{m(n)}\, S_{m(n)} \geqslant c_\alpha(m(n))) - \delta_n G^{(1)}(c_\alpha) + o(n^{-1}).$$

Then (3.2) and (3.6) imply

$$d = \frac{2(g_2(c_\alpha) - \bar{g}_2(c_\alpha))}{G^{(1)}(c_\alpha)\, c_\alpha} + o(1). \quad \square$$

Now we apply these results to our exapmle.

Let $X_1, X_2, \ldots$ be i.i.d.r.v.'s with

$$E\, X_1 = 0,\ E\, X_1^2 = 1, E\,|X_1|^{k+\delta} < \infty,\ k \geqslant 3, k \in \mathbb{N}, \delta > 0.(3.8)$$

Define

$$T_n = \frac{1}{n}(X_1 + \ldots + X_n).(3.9)$$

Suppose that the distribution of X_1 satisfies the Cramer condition (C)

$$\limsup_{|t|\to\infty} |E\, \exp\{itX_1\}| < 1.(3.10)$$

Under the conditions (3.8) and (3.10) (see Theorem 6.3.2) we have [8]

$$\sup_x |P(\sqrt{n}\, T_n < x) - \Phi(x) - \sum_{i=1}^{k-2} n^{-i/2} Q_i(x)| \leq \frac{C_{k,\delta}}{n^{(k-2+\delta)/2}}, C_{k,\delta} > 0, \quad n \in \mathbb{N},(3.11)$$

where the functions $Q_1(x), \ldots, Q_{k-2}(x)$ are defined in [8]

$$Q_1(x) = -(x^2 - 1)\, \varphi(x) \frac{E\, X_1^3}{6},$$

$$Q_2(x) = -(x^3 - 3x)\, \varphi(x) \frac{E\, X_1^4 - 3}{24} - (x^5 - 10x^3 + 15x)\, \varphi(x) \frac{(E\, X_1^3)^2}{72}.(3.12)$$

Carrying out the type of computation outlined above we arrive at the following simplified version of Lemma 1.1 (see (3.11)).

Lemma 3.2. *Let the conditions (3.8) – (3.10) with k = 3 be satisfied and $c_\alpha^*(n)$ be defined by (3.9), then*

$$c_\alpha^*(n) = u_\alpha + \frac{E\, X_1^3}{6\sqrt{n}}(u_\alpha^2 - 1) +$$

$$+ \frac{1}{12n}\Big(\frac{E^2\, X_1^3}{3}(5u_\alpha - 2u_\alpha^3) + \frac{E\, X_1^4 - 3}{2}(u_\alpha^3 - 3u_\alpha)\Big) + o(n^{-1}),$$

where $u_\alpha = \Phi^{-1}(1 - \alpha)$ denotes the upper α – point of the standard normal distribution.

Now let $Y_1, Y_2, \ldots$ be i.i.d.r.v.'s and

$$E\, Y_1 = 0,\ E\, Y_1^2 = 1, E\,|Y_1|^{4+\delta} < \infty,\ \delta > 0.(3.13)$$

Define

$$S_n = \frac{1}{n}(Y_1 + \ldots + Y_n).(3.14)$$

Suppose that

$$E\, Y_1^3 = E\, X_1^3,(3.15)$$

And

$$\limsup_{|t|\to\infty} |E\, \exp\{itY_1\}| < 1.(3.16)$$

Applying Theorem 3.1 we obtain

Lemma 3.3. *Under the above conditions of Lemma 3.2 and (3.13) - (3.16) the asymptotic deficiency d (see (3.4)) equals*

$$d = \frac{(E\, X_1^4 - E\, Y_1^4)(3 - u_\alpha^2)}{12} + o(1).$$

Samples with Random Sizes

Consider random variables $N_1, N_2, \ldots$ è $X_1, X_2, \ldots$, defined on the same probability space (Ω, A, P). The r.v.'s $X_1, X_2, \ldots, X_n$ will be treated as observations with n being a non-random sample size, whereas the r.v.'s N_n will be treated as random sample size depending on the parameter $n \in \mathbb{N}$. For example, if the r.v. N_n has the geometric distribution

$$P(N_n = k) = \frac{1}{n}(1-\frac{1}{n})^{k-1}, k \in \mathbb{N},$$

then

$$E\ N_n = n, (4.1)$$

that is, the r.v. N_n is parametrized by its expectation *n*.

Assume that for each $n \geq 1$ the r.v. N_n takes only natural values, that is, $N_n \in \mathbb{N}$ and are independent of the sequence $X_1, X_2, \ldots$. Everywhere in what follows consider the r.v.'s $X_1, X_2, \ldots$ to be independent and identically distributed. By $H_n = H_n(X_1, \ldots, X_n)$ denote a statistic, that is, real measurable function of observations $X_1, \ldots, X_n$. For each $n \geq 1$ define tne statistic H_{N_n} constructed from the sample of random size, that is

$$H_{N_n}(\omega) \equiv H_{N_n(\omega)}(X_1(\omega), \ldots, X_{N_n(\omega)}(\omega)),\ \omega \in \Omega.$$

Now assume that the d.f. of the non-normalized statistic H_n admits an asymptotic expansion described by the following condition.

Condition A. *There exist constants* $k \in \mathbb{N}, k \geqslant 2$, $\alpha_{in} \in \mathbb{R}, i = 1, \ldots, k, \beta_n > 0$, $C_k > 0$, *a differentiable d.f. G*(*x*) *and measurable functions* $g_j(x)$, *j = 1,…,k such that*

$$\beta_n \to 0,\ \max_{1 \leqslant i \leqslant k} |\alpha_{in}| \to 0,\ \ n \to \infty,$$

$$\sup_x |P(H_n < x) - G(x) - \sum_{i=1}^{k} \alpha_{in}\ g_i(x)| \leq C_k\ \beta_n\ \ \ n \in \mathbb{N}.$$

Lemma 4.1. *If the condition A holds, then*

$$\sup_x |P(H_{N_n} < x) - G(x) - \sum_{i=1}^{k} E\ \alpha_{iN_n}\ g_i(x)| \leq C_k\ E\ \beta_{N_n}.$$

The proof is a simple exercise on the application of the formula of total probability.

Let $X_1, X_2, \ldots$ be i.i.d.r.v.'s and

$$E\ X_1 = 0,\ E\ X_1^2 = 1, E\,|X_1|^{k+\delta} < \infty,\ k \geqslant 3, k \in \mathbb{N}, \delta > 0. (4.2)$$

Define for each $n \in \mathbb{N}$

$$H_n = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n). (4.3)$$

Suppose that the distribution of X_1 satisfies the Cramer condition (*C*)

$$\limsup_{|t| \to \infty} |E\ \exp\{itX_1\}| < 1. (4.4)$$

Taking into accopnt (4.2), (4.4) and Theorem 6.3.2 [8] we obtain

$$\sup_x |P(H_n < x) - \Phi(x) - \sum_{i=1}^{k-2} n^{-i/2}\ Q_i(x)| \leq \frac{C_{k,\delta}}{n^{(k-2+\delta)/2}}, C_{k,\delta} > 0,\ \ n \in \mathbb{N}, (4.5)$$

Where, [8]

$$Q_1(x) = -(x^2 - 1)\ \varphi(x)\,\frac{E\ X_1^3}{6},$$

$$Q_2(x) = -(x^3 - 3x)\ \varphi(x)\,\frac{E\ X_1^4 - 3}{24} - (x^5 - 10x^3 + 15x)\ \varphi(x)\,\frac{(E\ X_1^3)^2}{72}. (4.6)$$

Using (4.5) and Lemma 4.1, one has

Lemma 4.2. *Let the conditions (4.2) - (4.4) be satisfied, then*

$$\sup_x |P(H_{N_n} < x) - \Phi(x) - \sum_{i=1}^{k-2} E\ N_n^{-i/2}\ Q_i(x)| \leq C_{k,\delta}\ E\ N_n^{-(k-2+\delta)/2}.$$

After these preliminaries (see (4.5) and Lemma 4.2), the following Lemma can be formulated.

Lemma 4.3. *Suppose that the conditions (4.2) - (4.4) hold with k* = 4, *δ* > 0 *and there exist a, b such that*

$$E\ N_n = n,\ \ E\ N_n^{-1/2} = \frac{1}{\sqrt{n}} + \frac{a}{n} + o(n^{-1}),\ \ a \in \mathbb{R},$$
$$E\ N_n^{-1} = \frac{b}{n} + o(n^{-1}),\ \ E\ N_n^{-(2+\delta)/2} = o(n^{-1}),\ \ b \in \mathbb{R},$$

Then,

$$\sup_x |P(H_n < x) - \Phi(x) - \frac{Q_1(x)}{\sqrt{n}} - \frac{Q_2(x)}{n}| = o(n^{-1})$$

and

$$\sup_x |P(H_{N_n} < x) - \Phi(x) - \frac{Q_1(x)}{\sqrt{n}} - \frac{bQ_2(x) + aQ_1(x)}{n}| = o(n^{-1}).$$

For $n \geq 1$ let $H_n = H_n(X_1, \ldots, X_n)$ be a statistic, that is, a measurable function of the r.v.'s $X_1, \ldots, X_n$. The asymptotic quantile of order $h_\alpha^*(n)$ *α*, *α* ϵ (0, 1) (the α – quantile) of statistic H_n is the value for which

$$P(\sqrt{n}\ H_n \geqslant h_\alpha^*(n)) = \alpha + o(n^{-1}),\ \ \ n \to \infty. (4.7)$$

and we consider α – quantile of statistic H_{N_n}. That is the value $h_\alpha(n)$ for which

$$P(H_{N_n} \geqslant h_\alpha(n)) = \alpha + o(n^{-1}),\ \ \ n \to \infty. (4.8)$$

Taking into account (4.5), (4.6) and Lemma 3.1 we obtain

Lemma 4.4. *Suppose that the conditions (4.2) - (4.4) hold with k* = 4, *δ* > 0, *then under the conditions of Lemma 4.3 α – quantiles* $h_\alpha^*(n)$ *and* $h_\alpha(n)$ *admit the following asymptotic expansions*

$$h_\alpha^*(n) = u_\alpha + \frac{E\ X_1^3}{6\sqrt{n}}(u_\alpha^2 - 1) +$$
$$+ \frac{1}{12n}(\frac{E^2\ X_1^3}{3}(5u_\alpha - 2u_\alpha^3) + \frac{E\ X_1^4 - 3}{2}(u_\alpha^3 - 3u_\alpha)) + o(n^{-1}),$$

$$h_\alpha(n) = u_\alpha + \frac{E\ X_1^3}{6\sqrt{n}}(u_\alpha^2 - 1) +$$
$$+ \frac{1}{12n}(\frac{E^2\ X_1^3}{3}(5u_\alpha - 2u_\alpha^3) + \frac{b(E\ X_1^4 - 3)}{2}(u_\alpha^3 - 3u_\alpha) + 2a\ E\ X_1^3\ (u_\alpha^2 - 1)) + o(n^{-1}),$$

where $\Phi(u_\alpha) = 1 - \alpha$.

Define the sequence of positive integers $\{m(n) = n + d^* + o(1),\ d^* \in \mathbb{R},\ \ \ n = 1, 2, \ldots\}$ by the relation (*d* is the asymptotic deficiency)

$$P(H_{N_{m(n)}} \geqslant \sqrt{m(n)/n}\ h_\alpha^*(m(n))) = \alpha + o(n^{-1}),\ \ n \to \infty, (4.9)$$

Now we have in analogy to Theorem 3.1

Theorem 4.5. *Suppose that*

$$E\ N_n = n,\ \ E\ N_n^{-1/2} = \frac{1}{\sqrt{n}} + \frac{a}{n} + o(n^{-1}),\ \ a \in \mathbb{R},$$

$$E\ N_n^{-1} = \frac{b}{n} + o(n^{-1}),\ \ E\ N_n^{-(2+\delta)/2} = o(n^{-1}),\ \ b \in \mathbb{R}$$

and

$$\sup_x |P(H_n < x) - G(x) - \frac{g_1(x)}{\sqrt{n}} - \frac{g_2(x)}{n}| \leqslant \frac{C}{n^{(2+\delta)/2}},\ \ \delta > 0,$$

then the asymptotic deficiency d^* (see. (4.9)) satisfies

$$d^* = \frac{2(g_2(c_\alpha)\,(1-b) - a\ g_1(c_\alpha))}{G^{(1)}(c_\alpha)\ c_\alpha} + o(1),$$

where $G(c_\alpha) = 1 - \alpha$.

The result of these steps is the following Lemma.

Lemma 4.6. *If the conditions of Lemma 4.3 are satisfied, we have (see. (3.12))*

$$d^* = \frac{2((1-b)\,Q_2(u_\alpha) - a\,Q_1(u_\alpha))}{\varphi(u_\alpha)\,u_\alpha} + o(1).$$

If

$$E\,X_1^3 = 0,$$

Then,

$$d^* = \frac{(1-b)\,(3-u_\alpha^2)\,(E\,X_1^4 - 3)}{12} + o(1).$$

Discussion

The case of the samples with random size having a three-point symmetric distribution

In the previous section the results of section 3 were used to solve the main problem of this section. Here we briefly discuss another application of these results (see Lemma 4.2 and Theorem 4.5). Let N_n have a three-point distribution with parameter h_n

$$N_n : \begin{matrix} n-h_n, & n, & n+h_n \\ \frac{1}{3} & \frac{1}{3}, & \frac{1}{3}, \end{matrix} \qquad (5.1)$$

where $h_n < n$ and

$$\lim_{n\to\infty} \frac{h_n}{n} = 0.(5.2)$$

Lemma 5.1. *Let $\{h_n\}$ be a sequence of positive real numbers with $h_n < n$ and assume that (5.1) and (5.2) hold. Then,*

$$E\,N_n = n,$$

$$E\,N_n^{-1/2} = \frac{1}{\sqrt{n}} - \frac{1}{4\sqrt{n}}(\frac{h_n}{n})^2 + O(\frac{1}{\sqrt{n}}(\frac{h_n}{n})^3),$$

$$E\,N_n^{-1} = \frac{1}{n} + \frac{2}{3n}(\frac{h_n}{n})^2 + O(\frac{1}{n}(\frac{h_n}{n})^4),$$

$$E\,N_n^{-3/2} = \frac{1}{n^{3/2}} + O(\frac{1}{n^{3/2}}(\frac{h_n}{n})^2), \quad n \to \infty.$$

Proof: Here we only sketch the proof. We have:

$$E\,N_n^{-1} = \frac{3n^2 - h_n^2}{3n(n^2 - h_n^2)} =$$

$$= \frac{1}{n}(1 - \frac{h_n^2}{3n})\,(1 + \frac{h_n^2}{n^2} + O(\frac{h_n^4}{n^4})) =$$

$$= \frac{1}{n} + \frac{2}{3n}(\frac{h_n}{n})^2 + O(\frac{1}{n}(\frac{h_n}{n})^4),$$

$$E\,N_n^{-3/2} = \frac{1}{3n^{3/2}}(\frac{1}{(1-h_n/n)^{3/2}} + 1 + \frac{1}{(1+h_n/n)^{3/2}}) =$$

$$= \frac{1}{n^{3/2}} + O(\frac{1}{n^{3/2}}(\frac{h_n}{n})^2).$$

The proof for the other cases are similar and left to the reader.

Carrying out the type of computation outlined above we arrive at the following simplified version of Lemma 4.1.

Lemma 5.2. *Suppose that (4.2) - (4.4) (k = 4 and* $0 < \delta \le 1$*), (5.1) and (5.2) are satisfied. Then*

$$\sup_x |\,P(H_{N_n} < x) - \Phi(x) - \frac{1}{\sqrt{n}}(1 - \frac{h_n^2}{4n^2})\,Q_1(x) - .$$

$$. - \frac{1}{n}(1 + \frac{2h_n^2}{3n^2})\,Q_2(x)\,| = O(\frac{1}{n^{(2+\delta)/2}}(\frac{h_n}{n})^{(4+2\delta)/3}).$$

Corollary 5.2. *Under the conditions of Lemma 5.2 we have for $h_n = n^{3/4}$ (uniformly in $x \in \mathbb{R}$)*

$$P(H_{N_n} < x) = \Phi(x) + \frac{1}{\sqrt{n}}Q_1(x) + \frac{1}{n}(Q_2(x) - \frac{1}{4}Q_1(x)) + o(n^{-1}).$$

The result of these Lemmas is the following Theorem.

Theorem 5.3. *If the conditions of Corollary 5.2 are satisfied, we have (see (4.7), (4.8) and (4.9))*

$$h_\alpha^*(n) = u_\alpha + \frac{E\,X_1^3}{6\sqrt{n}}(u_\alpha^2 - 1) +$$

$$+ \frac{1}{12n}(\frac{E^2\,X_1^3}{3}(5u_\alpha - 2u_\alpha^3) + \frac{E\,X_1^4 - 3}{2}(u_\alpha^3 - 3u_\alpha)) + o(n^{-1}),$$

$$h_\alpha(n) = u_\alpha + \frac{E\,X_1^3}{6\sqrt{n}}(u_\alpha^2 - 1) +$$

$$+ \frac{1}{12n}(\frac{E^2\,X_1^3}{3}(5u_\alpha - 2u_\alpha^3) + \frac{E\,X_1^4 - 3}{2}(u_\alpha^3 - 3u_\alpha) - \frac{1}{2}E\,X_1^3\,(u_\alpha^2 - 1)) + o(n^{-1}),$$

where $\Phi(u_\alpha) = 1 - \alpha$ and

$$d^* = \frac{Q_1(u_\alpha)}{2\varphi(u_\alpha)\,u_\alpha} + o(1) = \frac{(1-u_\alpha^2)\,E\,X_1^3}{12\,u_\alpha} + o(1).$$

Conclusion

In the paper we consider asymptotic deficiencies of some estimators based on the samples with random sizes. It can be illustrative characteristic of a possible loss of the accuracy of statistical inference if a random-size-sample is erroneously regarded as a sample with non-random size. Some basic results dealing with some properties of estimators based on the samples with random sizes are also presented.

Acknowledgement

The research is supported by the Russian Foundation for Basic Research, Project 18-07-00252.

2010 Mathematical Subject Classification: Primary 60F05, Secondary 62E20, 91B30, 91B70

References

1. Bening VE. Korolev VY. On an application of the student distribution in the theory of probability and mathematical statistics. Theory Probab Its Appl. 2005; 49(3):377-91.

2. Bening VE. Korolev VY. Some statistical problems related to the Laplace distribution. Informatics and Its Applications. 2008; 2(2):19-34.

3. Bening VE. Korolev VY. Generalized Poisson models and their applications in insurance and finance. Berlin, Germany: W. de Gruyter. 2012; p:433.

4. Gnedenko BV. On estimation of unknown parameters from a random number of independent observations. Trans. Razmadze Tbilisi Math. Inst. 1989; 92:146-50.

5. Gnedenko BV. Korolev VY. Random summation. Limit

theorems and applications. (1st edn) Boca Raton: CRC Press. 1996; p:267.

6. Hodges JL, Lehmann EL. Deficiency. Ann. Math. Statist. 1970; 41(5):783-801.
7. Bening VE. Asymptotic theory of testing statistical hypotheses efficient statistics, optimality, power loss, and deficiency. (2nd edn) -Berlin, Germany: W. de Gruyter. 2011; p:277.
8. Petrov VV. Sums of independent random variables. Springer-Verlag, Berlin. 1975; p:414.
9. Bening VE. Korolev VY, SavushkinVA, Zeifman AI. On the deficiency of some estimators constructed from samples with random sizes. AIP Conference Proceedings, New York, USA. 2015; 1648(1).
10. Cramér H. Mathematical methods of statistics. Princeton University Press, Princeton, New Jersey, USA. 1946. pp:631.

***Correspondence to:**

V.E. Bening
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Russia
Tel: +7 495 939-31-21
E-mail: bening@yandex.ru