

On the asymptotic deficiency of some statistical estimators based on samples with random size

Bening V.E.,*

Keywords

estimator; risk function; asymptotic deficiency; sample of random size; asymptotic expansions; Poisson distribution;

Abstract

In the paper we consider asymptotic deficiencies of some estimators constructed from samples with random sizes. The case of the Poisson distribution is investigated. Some examples are presented.

1 Introduction and summary

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 and/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

*Peoples' Friendship University of Russia (RUDN University), 6 Miklukho - Maklaya St., Moscow, 117198, Russian Federation; bening@yandex.ru

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 see, e. g., [1] – [4]. 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.

Suppose that $\delta_n^*(X_1, \dots, X_n)$ and $\delta_n(X_1, \dots, X_n)$ are two competing estimators of $g(\theta)$, $\theta \in \Theta$ based on n random 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 estimator $\delta_{m(n)}(X_1, \dots, X_{m(n)})$ to match the performance of $\delta_n^*(X_1, \dots, X_n)$. Asymptotic 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 behaviour. The first general investigation of $m(n) - n$ was carried out by Hodges and Lehmann ([5]). They name $m(n) - n$ the deficiency of δ_n with respect to δ_n^* and denote it as

$$d_n = m(n) - n. \quad (1.1)$$

Suppose that for $n \rightarrow \infty$, the ratio $n/m(n)$ tends to a limit b , the asymptotic relative efficiency of $\delta_n(X_1, \dots, X_n)$ with respect to $\delta_n^*(X_1, \dots, X_n)$. 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 be anything from $o(1)$ to $o(n)$, does provide important additional information.

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

The deficiency of δ_n relative to δ_n^* will then indicate how many observations one loses by insisting on δ_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

$$R_n(\theta) = \mathbf{E}_\theta (\delta_n - g(\theta))^2, \quad R_n^*(\theta) = \mathbf{E}_\theta (\delta_n^* - g(\theta))^2,$$

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

$$R_n^*(\theta) = R_{m(n)}(\theta). \quad (1.2)$$

In order to solve (1.2), $m(n)$ has to be treated as a continuous variable. This can be done in a satisfactory manner by defining $R_{m(n)}(\theta)$ for non - integral $m(n)$ as

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

(cf. [5]).

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

$$R_n^* = \frac{a(\theta)}{n^r} + \frac{b(\theta)}{n^{r+s}} + o(n^{-(r+s)}), \quad (1.3)$$

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

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 (1.1) – (1.4), it now easily follows that (see [5])

$$d_n(\theta) \equiv \frac{c(\theta) - b(\theta)}{r a(\theta)} n^{(1-s)} + o(n^{(1-s)}). \quad (1.5)$$

Hence

$$d(\theta) \equiv d = \begin{cases} \pm\infty, & 0 < s < 1, \\ \frac{c(\theta) - b(\theta)}{r a(\theta)}, & s = 1, \\ 0, & s > 1. \end{cases} \quad (1.6)$$

A useful property of deficiencies is the following (transitivity): if a third estimator $\bar{\delta}_n$ is given, for which the risk $\bar{R}_n(\theta)$ also has an expansion of the form (1.4), the deficiency d of $\bar{\delta}_n$ with respect to δ_n^* satisfies

$$d = d_1 + d_2,$$

where d_1 is the deficiency of $\bar{\delta}_n$ with respect to δ_n and d_2 is the deficiency of δ_n with respect to δ_n^* .

The situation where $s = 1$ seems to be the most interesting one. Hodges and Lehmann ([5]) demonstrate the use of deficiency in a number of simple examples for which this is the case (see also [6]). The present paper consists of a number of applications of the deficiency concept in problems of point estimation in the case when number of observations is random.

We use conventional notation: \mathbb{R} is the set of real numbers, \mathbb{N} is the set of natural numbers, $h(n) \sim f(n)$, $n \rightarrow \infty \iff \lim_{n \rightarrow \infty} h(n)/f(n) = 1$.

2 Estimators based on sample with random size

Consider random variables (r.v.'s) N_1, N_2, \dots and X_1, X_2, \dots , defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. By X_1, X_2, \dots, X_n we will mean random observations whereas the r.v. N_n will be regarded as the random sample size depending on the parameter $n \in \mathbb{N}$. For example, if the r.v. N_n has the geometric distribution

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

then

$$\mathbb{E} N_n = n,$$

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

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, \dots . Everywhere in what follows the r.v.'s X_1, X_2, \dots are assumed independent and identically distributed with distribution depending on $\theta \in \Theta \in \mathbb{R}$.

For every $n \geq 1$, by $T_n = T_n(X_1, \dots, X_n)$ denote a statistic, i.e., a real-valued measurable function of X_1, \dots, X_n . For each $n \geq 1$, we define a r.v. T_{N_n} by setting $T_{N_n}(\omega) \equiv T_{N_n(\omega)}(X_1(\omega), \dots, X_{N_n(\omega)}(\omega))$, $\omega \in \Omega$.

Everywhere in what follows it will be assumed that $\mathbf{E}N_n = n$, that is, the expected sample size equals the sample size for the case where it is non-random.

Theorem 2.1.

1. If $\delta_n = \delta_n(X_1, \dots, X_n)$ is any unbiased estimator of $g(\theta)$, that is, it satisfies

$$\mathbf{E}_\theta \delta_n = g(\theta), \quad \theta \in \Theta,$$

and $\delta_{N_n} \equiv \delta_{N_n}(X_1, \dots, X_{N_n})$, then

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

2. Suppose that numbers $a(\theta)$, $b(\theta)$ and $C(\theta) > 0$, $\alpha > 0$, $r > 0$, $s > 0$ exist such that

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

where

$$R_n^*(\theta) = \mathbf{E}_\theta (\delta_n^*(X_1, \dots, X_n) - g(\theta))^2,$$

then

$$\left| R_n(\theta) - a(\theta) \mathbf{E} N_n^{-r} - b(\theta) \mathbf{E} N_n^{-r-s} \right| \leq C(\theta) \mathbf{E} N_n^{-r-s-\alpha},$$

where

$$R_n(\theta) = \mathbf{E}_\theta (\delta_{N_n}^*(X_1, \dots, X_{N_n}) - g(\theta))^2.$$

Proof. The proof follows from the total probability formula

1.

$$\begin{aligned} \mathbf{E}_\theta \delta_{N_n} &= \sum_{k=1}^{\infty} \mathbf{E}_\theta \delta_k \mathbf{P}(N_n = k) = \\ &= \sum_{k=1}^{\infty} g(\theta) \mathbf{P}(N_n = k) = g(\theta) \sum_{k=1}^{\infty} \mathbf{P}(N_n = k) = g(\theta), \quad \theta \in \Theta. \end{aligned}$$

2.

$$\begin{aligned} &\left| R_n(\theta) - a(\theta) \mathbf{E} N_n^{-r} - b(\theta) \mathbf{E} N_n^{-r-s} \right| = \\ &= \left| \sum_{k=1}^{\infty} \mathbf{E}_\theta (\delta_k^* - g(\theta))^2 \mathbf{P}(N_n = k) - a(\theta) \sum_{k=1}^{\infty} \frac{1}{k^r} \mathbf{P}(N_n = k) - \right. \\ &\quad \left. - b(\theta) \sum_{k=1}^{\infty} \frac{1}{k^{r+s}} \mathbf{P}(N_n = k) \right| = \end{aligned}$$

$$\begin{aligned}
 &= \left| \sum_{k=1}^{\infty} \left(\mathbf{E}_{\theta} (\delta_k^* - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}} \right) \mathbf{P} (N_n = k) \right| \leq \\
 &\leq \sum_{k=1}^{\infty} \left| \mathbf{E}_{\theta} (\delta_k^* - g(\theta))^2 - \frac{a(\theta)}{k^r} - \frac{b(\theta)}{k^{r+s}} \right| \mathbf{P} (N_n = k) \leq \\
 &\leq \sum_{k=1}^{\infty} \frac{C(\theta)}{k^{r+s+\alpha}} \mathbf{P} (N_n = k) = \\
 &= C(\theta) \mathbf{E} N_n^{-r-s-\alpha}.
 \end{aligned}$$

□

Corollary 2.1.

Suppose that numbers $a(\theta)$, $b(\theta)$ and $r > 0$, $s > 0$ exist such that

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

where

$$R_n^*(\theta) = \mathbf{E}_{\theta} (\delta_n^*(X_1, \dots, X_n) - g(\theta))^2,$$

then

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

where

$$R_n(\theta) = \mathbf{E}_{\theta} (\delta_{N_n}^*(X_1, \dots, X_{N_n}) - g(\theta))^2.$$

Let observations X_1, \dots, X_n have expectation

$$\mathbf{E}_{\theta} X_1 = g(\theta)$$

and variance

$$\mathbf{D}_{\theta} X_1 = \sigma^2(\theta).$$

The customary estimator for $g(\theta)$ based on n observation is

$$\delta_n = \frac{1}{n} \sum_{i=1}^n X_i. \tag{2.1}$$

This estimator is unbiased and consistent, and its variance is

$$R_n^*(\theta) = \mathbf{D}_{\theta} \delta_n = \frac{\sigma^2(\theta)}{n}. \tag{2.2}$$

If this estimator is based on sample with random size, we have (see Corollary 1.1)

$$R_n(\theta) = \mathbf{D}_{\theta} \delta_{N_n}(X_1, \dots, X_{N_n}) = \sigma^2(\theta) \mathbf{E} N_n^{-1}. \tag{2.3}$$

If $g(\theta)$ is given, we consider the estimator for $\sigma^2(\theta)$ in the form

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

This estimator is unbiased and consistent, and its variance is

$$\bar{R}_n^*(\theta) = D_\theta \bar{\delta}_n = \frac{\mu_4(\theta) - \sigma^4(\theta)}{n}, \quad \mu_4(\theta) = E_\theta (X_1 - g(\theta))^4. \quad (2.5)$$

For this estimator with random size one has

$$\bar{R}_n(\theta) = D_\theta \bar{\delta}_{N_n}(X_1, \dots, X_n) = (\mu_4(\theta) - \sigma^4(\theta)) E N_n^{-1}. \quad (2.6)$$

Suppose now that $g(\theta)$ is unknown but that instead of (2.4) we are willing to consider any estimator of the form (see (2.1))

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

If $\gamma \neq -1$, this will not be unbiased but may have a smaller expected squared error than the unbiased estimator with $\gamma = -1$.

One easily finds that (see [1], (3.6) and [2])

$$\begin{aligned} \tilde{R}_n^*(\theta) &= E_\theta (\tilde{\delta}_n(X_1, \dots, X_n) - \sigma^2(\theta))^2 = \\ &= \frac{\sigma^4(\theta)}{n(n + \gamma)^2} \left((n - 1) ((\mu_4(\theta)/\sigma^4(\theta) - 1) (n - 1) + 2) + n(\gamma + 1)^2 \right) \end{aligned} \quad (2.8)$$

and hence

$$\begin{aligned} \tilde{R}_n^*(\theta) &= \sigma^4(\theta) \left(\frac{(\mu_4(\theta)/\sigma^4(\theta) - 1)}{n} + \right. \\ &\left. + \frac{(\gamma + 1)^2 - 2(\mu_4(\theta)/\sigma^4(\theta) - 1) + 2 - 2\gamma(\mu_4(\theta)/\sigma^4(\theta) - 1)}{n^2} \right) + O(n^{-3}). \end{aligned} \quad (2.9)$$

Using Theorem 1.1, we have

$$\begin{aligned} \tilde{R}_n(\theta) &= E_\theta (\tilde{\delta}_{N_n}(X_1, \dots, X_{N_n}) - \sigma^2(\theta))^2 = \\ &= \sigma^4(\theta) \left((\mu_4(\theta)/\sigma^4(\theta) - 1) E N_n^{-1} + \right. \\ &\left. + ((\gamma + 1)^2 - 2(\mu_4(\theta)/\sigma^4(\theta) - 1) + 2 - 2\gamma(\mu_4(\theta)/\sigma^4(\theta) - 1)) E N_n^{-2} \right) + O(E N_n^{-3}). \end{aligned} \quad (2.10)$$

□

3 Deficiencies of some estimators based on samples with random 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 $E N_n = n$) are evaluated, we actually compare the expected size $m(n)$ of a random sample with n by virtue of the quantity $d_n = m(n) - n$ and its limit value.

We now apply the results of section 2 to the three examples given in this section. Let M_n be the Poisson r.v. with parameter $n - 1$, $n \geq 2$, i.e.

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

Define the random size as

$$N_n = M_n + 1,$$

then

$$\mathbf{E} N_n = n$$

and

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

Then

$$\mathbf{E} N_n^{-1} = \frac{1}{n} + \frac{1}{n^2} + o(n^{-2}). \tag{3.1}$$

The deficiency of δ_{N_n} relative to δ_n (see (2.1)) is given by (2.2), (2.3), (3.1) and (1.6) with $r = s = 1, a(\theta) = \sigma^2(\theta), b(\theta) = 0, c(\theta) = \sigma^4(\theta)$, and hence is equal to

$$d = 1. \tag{3.2}$$

Similarly, the deficiency of $\bar{\delta}_{N_n}$ relative to $\bar{\delta}_n$ (see (2.4)) is given by (2.5), (2.6), (3.1) and (1.6) with $r = s = 1, a(\theta) = c(\theta) = \mu_4(\theta) - \sigma^4(\theta), b(\theta) = 0$, and hence is equal to

$$\bar{d} = 1. \tag{3.3}$$

Consider now third example (see (2.7)). We have

$$\begin{aligned} \mathbf{E} N_n^{-2} &= e^{1-n} \sum_{k=0}^{\infty} \frac{(n-1)^k}{(k+1)^2 k!} = \frac{e^{1-n}}{n-1} \sum_{k=1}^{\infty} \frac{(n-1)^k}{k k!} = \\ &= \frac{e^{1-n}}{n-1} \int_0^{n-1} \frac{e^x - 1}{x} dx. \end{aligned}$$

Then, using L'Hôpital principle we obtain

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

and

$$\mathbf{E} N_n^{-2} \sim \frac{1}{n^2}, \quad n \rightarrow \infty. \tag{3.4}$$

Now the deficiency of $\tilde{\delta}_{N_n}$ relative to $\tilde{\delta}_n$ (see (2.7)) is given by (2.9), (2.10), (3.4) and (1.6) with $r = s = 1$ and hence is equal to

$$\tilde{d} = 1 \tag{3.5}$$

and the deficiency of $\tilde{\delta}_{N_n}^{(\gamma_1)}$ relative to $\tilde{\delta}_{N_n}^{(\gamma_2)}$ (see (2.7)) is given by (3.1), (3.4) and (1.6) with $r = s = 1$ and hence is equal to

$$\tilde{d}_{\gamma_1, \gamma_2} = (\gamma_1 - \gamma_2) \left(\frac{\gamma_1 + \gamma_2 + 2}{\mu_4(\theta)/\sigma^4(\theta) - 1} - 2 \right). \tag{3.6}$$

The classical $\tilde{\delta}_{N_n}^{(0)}$ is thus better than $\tilde{\delta}_{N_n}^{(-1)}$ when

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

with the situation reversed when

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

When X_1 is normal, in particular,

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

and

$$\tilde{d}_{\gamma_1, \gamma_2} = \frac{1}{2} (\gamma_1 - \gamma_2) (\gamma_1 + \gamma_2 - 2). \quad (3.7)$$

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

These examples illustrate the following:

Theorem 3.1.

Suppose that numbers $a(\theta)$, $b(\theta)$ and k_1 , k_2 exist such that

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

and

$$\mathbf{E} N_n^{-1} = \frac{1}{n} + \frac{k_1}{n^2} + o(n^{-2}),$$

$$\mathbf{E} N_n^{-2} = \frac{k_2}{n^2} + o(n^{-2}),$$

$$\mathbf{E} N_n^{-3} = o(n^{-2}),$$

then the asymptotic deficiency of $\delta_{N_n}(X_1, \dots, X_{N_n})$ with respect to $\delta_n(X_1, \dots, X_n)$ is equal to

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

For a proof, see Theorem 2.1 and (1.5), (1.6).

Acknowledgement

The publication was financially supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 02.a03.21.0008).

References

1. *Bening V. E., Korolev V. Yu.* On an application of the Student distribution in the theory of probability and mathematical statistics, Theory of Probability and Its Applications, 2005, Vol. 49, No. 3, pp. 377–391.
2. *Bening V. E., Korolev V. Yu.* *Generalized Poisson Models and their Applications in Insurance and Finance*, W. de Gruyter, (Berlin, Germany) ISBN 978 - 3 - 11 - 093601 - 8, 2012, 433 p.

3. Gnedenko B. V., Korolev V. Yu. *Random Summation. Limit Theorems and Applications*, Boca Raton: CRC Press, 1996, 267 p.
4. Bening V. E., Korolev V. Yu. Some statistical problems related to the Laplace distribution, *Informatics and Its Applications*, 2008, Vol. 2, No. 2, pp. 19 – 34.
5. Hodges J. L., Lehmann E. L. Deficiency, *Ann. Math. Statist*, 1970, V.41, No. 5, pp. 783 – 801.
6. Bening V. E. *Asymptotic Theory of Testing Statistical Hypotheses: Efficient Statistics, Optimality, Power Loss, and Deficiency*, W. de Gruyter, (Berlin, Germany) ISBN 978 - 3 - 11 - 093599 - 8, 2011, 277 p.
7. Cramér H. *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946, 631 p.

Author biographies

VLADIMIR BENING is Doctor of Science in physics and mathematics; professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University; leading scientist, Institute of Applied Mathematics and Telecommunications, Peoples' Friendship University of Russia (RUDN University), senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. His email is bening@yandex.ru.