# STATISTICAL ANALYSIS OF ROUNDED DATA: MEASUREMENT ERRORS VS ROUNDING ERRORS

**N.G. Ushakov**[1] **and V.G. Ushakov**[2]

Since data for statistical analysis are always given in a discretized form, observations contain not only measurement errors but also rounding errors which are determined by the discretization step. In this paper we consider situations where the rounding errors are considerable: they are comparable to or even greater (in average) than the measurement errors. It is shown that it can be reasonable to increase the measurement errors in order to reduce the error of the final result.

## 1. Introduction

Real data in statistics is always given in a discretized (rounded) form. The rounding errors can play an important role in many situations. Suppose, for example, that one needs to measure some quantity. To reduce the error one makes several measurements and then calculates the average. The observations are rounded. In this case, if the measurement errors are much less than the step of discretization then the (guaranteed) accuracy of the final result cannot be better than the discretization step, and the averaging is useless.

Rounding errors are especially serious and must not be ignored when the sample size is large. At present, big data sets are becoming more and more common due to the rapid development of computer technology, therefore there is a growing interest in statistical analysis of rounded data; see, for example, [1–4] and the references therein.

In the present paper we study (both theoretically and by Monte Carlo simulation) how the influence of the rounding errors can be reduced. It turns out that it is useful to add a suitable component to the measurement error. The uniqueness of this paper is that we do not suppose that the discretization step tends to zero. It is supposed to be constant, and, without loss of generality, we suppose that it is equal to 1. This means that we consider discretization as rounding to the nearest integer. So, let $x$ be a real number with the integer part $[x]$ and fractional part $\{x\}$. The rounded value (denote it by $x^*$) is equal to $[x]$ if $\{x\} < 1/2$ and $[x]+1$ if $\{x\} \geqslant 1/2$. Note that $x^* = [x+1/2]$.

Consider a random sample $X_1, \dots, X_n$ consisting of independent and identically distributed random variables with finite unknown expectation $\mu$. The expectation is estimated, but the $X$-s are not observed. Instead, we observe the rounded values $X_1^*, \dots, X_n^*$. We are interested in the properties of the estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(X_i)^*. \tag{1}$$

In the next section we obtain an upper bound for the distance between the almost sure limit of the estimator and the estimated expectation $\mu$. In Section 3, some simulation results will be presented for finite samples. Some sufficient conditions for the consistency of estimator (1) were obtained in [5].

If a sequence of random variables $T_1, T_2, \dots$ converges almost surely to a random varable $T$, then, as usual, we denote this $T_n \xrightarrow{a.s.} T$ and also write

$$T = \lim_{n\to\infty}(a.s.)T_n.$$

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway, e-mail: ushakov@math.ntnu.no

[2] Moscow State University, Moscow, Russia, e-mail: vgushakov@mail.ru

Proceedings of the XXXIII International Seminar on Stability Problems for Stochastic Models, Svetlogorsk, Russia, June 12–18, 2016.

1072-3374/18/2346-0770

## 2. The sample mean: an upper bound for the limiting error

We will use the following model and make the following assumptions. The observations $X_1, X_2, \dots$ are represented in the form

$$X_i = \mu + \lambda Y_i,\ i = 1, 2, \dots,$$

where $Y_1, Y_2, \dots$ are independent identically distributed random variables such that $\mathsf{E} Y_i = 0$. We can change the parameter $\lambda$. The aim is to reduce the distance between $n^{-1}\sum_{i=1}^{n}(X_i)^*$ and $\mu$.

Let $g(x)$ be a real-valued function defined on a finite or infinite interval $(a, b)$. The total variation of $g(x)$ is defined as

$$\mathrm{V}(g) = \sup \sum_{i=0}^{n-1} |g(x_{i+1}) - g(x_i)|,$$

where the supremum is taken over all $n$ and all sets $x_0, x_1, \dots, x_n$ such that $a < x_0 < x_1 < \dots < x_n < b$.

Denote

$$S_m = \sum_{k=1}^{\infty} \frac{1}{k^m},$$

where $m$ is a natural number. For example, $S_1 = \infty$, $S_2 = \pi^2/6$.

**Theorem.** *Let $Y$-s be absolutely continuous with probability density function $f(x)$. If $f(x)$ is $m$ times differentiable, $m \geqslant 0$, then*

$$\left| \lim_{n\to\infty} (a.s.) \frac{1}{n} \sum_{i=1}^{n} (X_i)^* - \mu \right| \leqslant \frac{S_{m+2}\mathrm{V}(f^{(m)})}{2^{m+1}\pi^{m+2}\lambda^{m+1}} \tag{2}$$

*($f^{(0)}$ is defined as $f$).*

**Proof.** Denote the characteristic function of the variables $Y$-s by $\varphi(t)$. According to Theorem 2 of [6],

$$|\varphi(t)| \leqslant \frac{\mathrm{V}(f^{(m)})}{|t|^{m+1}} \tag{3}$$

for all real $t$. Consider $\mathsf{E}(X_i)^*$. We have

$$\mathsf{E}(X_i)^* = \mathsf{E}\left[X_i + \frac{1}{2}\right] = \mathsf{E}\left(X_i + \frac{1}{2}\right) - \mathsf{E}\left\{X_i + \frac{1}{2}\right\} = \mu + \left(\frac{1}{2} - \mathsf{E}\left\{X_i + \frac{1}{2}\right\}\right);$$

therefore

$$|\mathsf{E}(X_i)^* - \mu| = \left| \frac{1}{2} - \mathsf{E}\left\{X_i + \frac{1}{2}\right\} \right|. \tag{4}$$

Let $g(x)$ and $\psi(t)$ be the density and the characteristic function of $X_i + 1/2$. Then

$$\mathsf{E}\left\{X_i + \frac{1}{2}\right\} = \int_0^1 x \sum_{n=-\infty}^{\infty} g(x+n)dx.$$

Due to the Poisson summation formula (see Feller (1971), p. 632, Formula (5.9) with $\zeta = 0$, $\lambda = \pi$),

$$\sum_{n=-\infty}^{\infty} g(x+n) = \sum_{k=-\infty}^{\infty} \psi(2\pi k)e^{-ix2\pi k};$$

therefore

$$\mathsf{E}\left\{X_i + \frac{1}{2}\right\} = \sum_{k=-\infty}^{\infty} \left(\psi(2\pi k) \int_0^1 x e^{-ix2\pi k}\right) = \frac{1}{2} + \sum_{k\neq 0} \left(\psi(2\pi k)\frac{i}{2\pi k}\right) = \frac{1}{2} - \sum_{k=1}^{\infty} \frac{\Im\psi(2\pi k)}{\pi k} \tag{5}$$

(we took into account that the real part of any characteristic function is even and the imaginary part is odd). But

$$\psi(t) = e^{it(\mu+1/2)}\varphi(\lambda t). \tag{6}$$

From (3)–(6) we obtain

$$|\mathsf{E}(X_i)^* - \mu| \leqslant \sum_{k=1}^{\infty} \frac{\Im\psi(2\pi k)}{\pi k} \leqslant \sum_{k=1}^{\infty} \frac{\psi(2\pi k)}{\pi k} = \sum_{k=1}^{\infty} \frac{\varphi(2\pi k\lambda)}{\pi k} \leqslant$$

$$\leqslant \sum_{k=1}^{\infty} \frac{\mathrm{V}(f^{(m)})}{\lambda^{(m+1)}(2\pi k)^{m+1}\pi k} = \frac{S_{m+2}\mathrm{V}(f^{(m)})}{2^{m+1}\pi^{m+2}\lambda^{m+1}}.$$

The result follows now from the classical strong law of large numbers.

In Table 1 we present the values of the right-hand side of inequality (2) when the distribution of $\lambda Y_i$ is normal with zero mean and variance $\sigma^2$, and when $m = 0$. The numbers are given for different values of $\sigma$. The table shows that the limiting error of the estimator is much less than the discretization step, if the variance of the measurement errors is large enough.

**Table 1.** The right-hand side of inequality (2) when the distribution of $\lambda Y_i$ is normal with zero mean and variance $\sigma^2$, and when $m = 0$. The numbers are given for different values of $\sigma$.

| $\sigma = 0$ | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.4$ | $\sigma = 0.6$ | $\sigma = 0.8$ | $\sigma = 1$ | $\sigma = 2$ | $\sigma = 3$ | $\sigma = 4$ |
|---|---|---|---|---|---|---|---|---|---|
| $\infty$ | 0.66490 | 0.33245 | 0.16623 | 0.11082 | 0.08311 | 0.06649 | 0.03325 | 0.02216 | 0.01662 |

## 3. Finite samples: simulation study

In the previous section we studied the limit behavior of the sample mean of rounded data. Now we consider the problem for finite samples, using the Monte Carlo simulation. The main aim is to study the dependence of the quality of the estimator on the variance $\sigma^2$ of the measurement errors.

For each $\sigma$ we simulate $N$ random samples $X_{1j}, \ldots, X_{nj}$, $j = 1, \ldots, N$, of size $n$ from the normal distribution with the expectation $\mu$ and the standard deviation $\sigma$. The observations are rounded, and the expectation $\mu$ is estimated by the sample mean of the rounded data, i.e., by

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}(X_{ij})^*.$$

Then the average distance is calculated between estimates and the estimated expectation, i.e., values

$$d(\sigma, n) = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{1}{n}\sum_{i=1}^{n}(X_{ij})^* - \mu\right|.$$

These values are presented in Table 2. We took $\mu = 0.4$, $N = 1000$.

The results presented in Table 2 show that if the measurement errors are small ($\sigma = 0, 0.1, 0.2$), then the accuracy of the estimator practically does not depend on the sample size and it is relatively low. Moreover, the less the standard deviation of the measurement errors, the less the accuracy of the estimator. As the standard deviation grows, the accuracy begins to depend on the sample size. For each sample size $n$, $d(\sigma, n)$ first decreases in $\sigma$ and then increases.

It is useful to compare Table 2 with Table 1. For “small” sample sizes ($10^2$, $10^3$, $10^4$), the observed average errors are less than the upper bound for the limiting error for small values of $\sigma$ and greater for large values of $\sigma$. For large sample sizes ($10^5$, $10^6$), the upper bound is essentially greater than the observed values for all $\sigma$-s. Perhaps this means that for the normal distribution inequality (2) can be essentially improved.

**Table 2.** Simulation results: the mean distance between estimates and the expectation for different distributions of the additional errors, maximum with respect to $\mu$.

| $\sigma$ | $n = 10^2$ | $n = 10^3$ | $n = 10^4$ | $n = 10^5$ | $n = 10^6$ |
|---|---|---|---|---|---|
| 0 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| 0.1 | 0.24171 | 0.24135 | 0.24124 | 0.24133 | 0.24136 |
| 0.2 | 0.09436 | 0.09254 | 0.09128 | 0.09152 | 0.09146 |
| 0.4 | 0.04147 | 0.01462 | 0.00794 | 0.00792 | 0.00797 |
| 0.6 | 0.05214 | 0.01633 | 0.00542 | 0.00165 | 0.00053 |
| 0.8 | 0.06545 | 0.02177 | 0.00698 | 0.00210 | 0.00067 |
| 1 | 0.08237 | 0.02604 | 0.00837 | 0.00254 | 0.00084 |
| 2 | 0.16131 | 0.05097 | 0.01576 | 0.00504 | 0.00161 |
| 3 | 0.23262 | 0.07834 | 0.02274 | 0.00771 | 0.00248 |
| 4 | 0.30467 | 0.10610 | 0.03250 | 0.01036 | 0.00311 |

## Acknowledgments

This research was supported by the Russian Scientific Foundation, project 14–11–00364.

## REFERENCES

1. W.M.Li and Z.D.Bai, “Rounded data analysis based on multi-layer ranked set sampling,” *Acta Math. Sin. (Engl. Ser.)*, **27**, 2507–2518 (2011).
2. H.Schneeweiss, J.Komlos, and A.S.Ahmad, “Symmetric and asymmetric rounding: a review and some new results,” *AStA Adv. Stat. Anal.*, **94**, 247–271 (2010).
3. B.Wang and W.Wertelecki, “Density estimation for data with rounding errors,” *Comput. Stat. Data Anal.*, **65**, 4–12 (2013).
4. N.Zhao and Z.Bai, “Analysis of rounded data in mixture normal model,” *Stat. Papers*, **53**, 895–914 (2012).
5. V.G.Ushakov and N.G.Ushakov, “On averaging of rounded data,” *Inf. Appl.*, **9**, 116–119 (2015).
6. V.G.Ushakov and N.G.Ushakov, “Some inequalities for characteristic functions of densities with bounded variation,” *Moscow Univ. Comput. Math. Cybern.*, **23**, 45–52 (2000).