Bounds for the accuracy of invalid normal approximation

A. V. Dorofeeva[1], ***V. Yu. Korolev***[1,2,3], ***A. I. Zeifman***[2,3,4]

[1] Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia

[2] Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia

[3] Moscow Center for Fundamental and Applied Mathematics, Moscow, Russia

[4] Vologda State University, Vologda, Russia

In applied studies, the normal approximation is often used for the distribution of data with (at least assumed) additive structure. This tradition is based on the central limit theorem of probability theory which states that the distributions of sums of (independent) random variables satisfying certain conditions (say, the Lindeberg condition) converge to the normal law as the number of summands infinitely increases. However, it is practically impossible to check the conditions providing the validity of the central limit theorem when the observed sample size is limited. In particular, with moderate sample size, the histogram constructed from the sample from the Cauchy distribution whose tails are so heavy that even the mathematical expectation does not exist, is practically visually indistinguishable from the normal (Gaussian) density. Therefore it is very important to know what the real accuracy of the normal approximation is in the cases where it is used despite it is theoretically inapplicable. Moreover, in some situations related with computer simulation, if the distributions of separate summands in the sum belong to the domain of attraction of a stable law with characteristic exponent less than two, then the observed distance between the distribution of the normalized sum and the normal law first decreases as the number of summands grows and begins to increase only when the number of summands becomes large enough. In the present paper an attempt is undertaken to give some theoretical explanation to this effect. In Section 2 we introduce the notation, give necessary definitions and formulate some auxiliary results. In Section 3 the theorem is proved presenting the upper bound for the accuracy of the invalid normal approximation. In Section 4 the problem of evaluation of the threshold number of summands providing best possible accuracy of the invalid normal approximation is considered.

Throughout the paper we assume that all the random variables are defined on the same probability space $(\Omega, \mathfrak{F}, \mathsf{P})$. The mathematical expectation and variance with respect to the probability measure P will be denoted E and D, respectively. The symbol $\stackrel{d}{=}$ means the coincidence of distributions.

For $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be a homogeneous sample, that is, a set of independent identically distributed random variables with common distribution function $F(x) = \mathsf{P}(X_1 < x)$, $x \in \mathbb{R}$. For simplicity, without serious loss of generality we will assume that $F(x)$ is continuous.

Denote $S_n = X_1 + \ldots + X_n$. The indicator of a set (event) $A \in \mathfrak{F}$ will be denoted $\mathbb{I}_A = \mathbb{I}_A(\omega)$, $\omega \in \Omega$:

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

Consider $u > 0$ such that $0 < F(u) < 1$. It is obvious that $X_j = X_j\mathbb{I}_{\{|X_j| \le u\}} + X_j\mathbb{I}_{\{|X_j| > u\}}$. Then

$$S_n = \sum\nolimits_{j=1}^{n} X_j\mathbb{I}_{\{|X_j| \le u\}} + \sum\nolimits_{j=1}^{n} X_j\mathbb{I}_{\{|X_j| > u\}} \equiv S_n^{(\le u)} + S_n^{(> u)}.$$

We will follow the lines of approach described in [3]. The key point of this approach is the statement that will be formulated here as the following lemma.

LEMMA 1. *Let $u > 0$ so that $0 < F(u) < 1$. Then*

$$S_n^{(\le u)} \stackrel{d}{=} \sum\nolimits_{j=0}^{N_n(u)} X_j^{(\le u)} \tag{1}$$

and

$$S_n^{(>u)} \stackrel{d}{=} \sum\nolimits_{j=0}^{n-N_n(u)} X_j^{(>u)}, \tag{2}$$

where $N_n(u)$ is a random variable that has the binomial distribution with parameters n ("number of trials") and $p = p(u) = \mathsf{P}(|X_1| \le u) = F(u) - F(-u)$ (probability of "success"), the random variables $X_1^{(\le u)}, \ldots, X_n^{(\le u)}$ are independent and have one and the same distribution function

$$F^{(\le u)}(x) \equiv \mathsf{P}(X_1^{(\le u)} < x) = \mathsf{P}\big(X_1 \mathbb{I}_{\{|X_1| \le u\}} < x \big| \, |X_1| \le u\big) =$$

$$= \frac{\mathsf{P}(X_1 < x;\, |X_1| \le u)}{\mathsf{P}(|X_1| \le u)} = \begin{cases} 1, & x > u; \\ \dfrac{F(x) - F(-u)}{F(u) - F(-u)}, & |x| \le u; \\ 0, & x < -u, \end{cases} \tag{3}$$

the random variables $X_1^{(>u)}, \ldots, X_n^{(>u)}$ are independent and have one and the same distribution function

$$F^{(>u)}(x) \equiv \mathsf{P}(X_1^{(>u)} < x) = \mathsf{P}\big(X_1 \mathbb{I}_{\{|X_1| > u\}} < x \big| \, |X_1| > u\big) =$$

$$= \frac{\mathsf{P}(X_1 < x;\, |X_1| > u)}{\mathsf{P}(|X_1| > u)} = \begin{cases} \dfrac{F(x)}{F(-u) + 1 - F(u)}, & x < -u; \\ \dfrac{F(-u)}{F(-u) + 1 - F(u)}, & |x| \le u; \\ \dfrac{F(-u) + F(x) - F(u)}{F(-u) + 1 - F(u)}, & x > u. \end{cases} \tag{4}$$

Moreover, the random variable N_n is independent of $X_1^{(\le u)}, \ldots, X_n^{(\le u)}$ and $X_1^{(>u)}, \ldots, X_n^{(>u)}$. For definiteness, if $N_n(u) = 0$, then the sum $S_n^{(\le u)}$ is set equal to zero and if $N_n(u) = n$, then the sum $S_n^{(>u)}$ is set equal to zero.

LEMMA 2. *Let $A \in \mathfrak{F}$, $B \in \mathfrak{F}$. Then $\mathsf{P}(AB) \ge \mathsf{P}(A) - \mathsf{P}(\overline{B})$.*

The uniform (Kolmogorov) distance between the distribution functions F_ξ and F_η of random variables ξ and η will be denoted $\rho(F_\xi,\, F_\eta)$, $\rho(F_\xi,\, F_\eta) = \sup_x |F_\xi(x) - F_\eta(x)|$. The normal distribution function with expectation $a \in \mathbb{R}$ and variance $\sigma^2 > 0$ will be denoted $\Phi_{a,\sigma}(x)$,

$$\Phi_{a,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\Big\{ - \frac{(z-a)^2}{2\sigma^2} \Big\} dz = \Phi_{0,1}\Big(\frac{x-a}{\sigma}\Big) = \Phi_{0,\sigma}(x-a), \quad x \in \mathbb{R}.$$

LEMMA 3. *For any $a \in \mathbb{R}$, $\sigma > 0$, $b \in \mathbb{R}$*

$$\rho(\Phi_{a+b,\,\sigma},\, \Phi_{a,\,\sigma}) = 2\Phi_{0,\sigma}\big(\tfrac{|b|}{2}\big) - 1.$$

LEMMA 4. *For $n \in \mathbb{N}$ let $\xi_1, \ldots, \xi_n$ be random variables, $a_1, \ldots, a_n$ be positive numbers such that $a_1 + \ldots + a_n = 1$. Then for any $x > 0$*

$$\mathsf{P}\Big(\Big|\sum\nolimits_{j=1}^{n} \xi_j\Big| \ge x\Big) \le \sum\nolimits_{j=1}^{n} \mathsf{P}(|\xi_j| \ge a_j x).$$

If, in addition, the random variables $\xi_1, \dots, \xi_n$ are identically distributed, then

$$\mathsf{P}\Big(\Big|\sum\nolimits_{j=1}^{n} \xi_j\Big| \ge x\Big) \le n\mathsf{P}\big(|\xi_1| \ge \tfrac{x}{n}\big).$$

LEMMA 5. *For $n \in \mathbb{N}$ let $\xi_1, \dots, \xi_n$ be random variables such that $\mathsf{E}|\xi_j|^\delta < \infty$ for some $\delta > 0$, $j = 1, \dots, n$. Denote $\theta_n = \xi_1 + \ldots + \xi_n$.*
(i) *If $0 < \delta \le 1$, then*

$$\mathsf{E}|\theta_n|^\delta \le \sum\nolimits_{j=1}^{n} \mathsf{E}|\xi_j|^\delta.$$

(ii) *If $1 \le \delta \le 2$, the random variables $\xi_1, \dots, \xi_n$ are independent and $\mathsf{E}\xi_j = 0$, $j = 1, \dots, n$, then*

$$\mathsf{E}|\theta_n|^\delta \le \Big(2 - \frac{1}{n}\Big) \sum\nolimits_{j=1}^{n} \mathsf{E}|\xi_j|^\delta.$$

Now turn to main results. Consider the upper bound for the uniform distance between the distribution of the normalized sum

$$S_n^* = \frac{1}{\sqrt{n}} \sum\nolimits_{j=1}^{n} X_j$$

and the normal law with some expectation $a \in \mathbb{R}$ and variance $\sigma^2 > 0$. The choice of concrete values of a and σ^2 will be discussed later.

From what has been said it follows that

$$S_n^* \stackrel{d}{=} \frac{S_n^{(\le u)}}{\sqrt{n}} + \frac{S_n^{(>u)}}{\sqrt{n}}.$$

For brevity and convenience, we will use the notation

$$\zeta_n = \frac{S_n^{(\le u)}}{\sqrt{n}}, \quad \eta_n = \frac{S_n^{(>u)}}{\sqrt{n}}.$$

THEOREM 1. *Let $u > 0$ be arbitrary. Then for any $a \in \mathbb{R}$ and $\sigma > 0$ we have*

$$\rho(F_{\zeta_n+\eta_n},\, \Phi_{a,\sigma}) \le \rho(F_{\zeta_n},\, \Phi_{a,\sigma}) + n\big(F(-u) + 1 - F(u)\big). \tag{5}$$

The proof of this result is based on Lemmas 1-4.

Theorem 1 can be easily extended to the case of non-identically distributed summands.

In practice, the values of the parameters a and σ can be chosen by the following reasoning. It is easy to verify (say, by the consideration of characteristic functions) that

$$S_n^{(\le u)} \stackrel{d}{=} \sum\nolimits_{j=1}^{n} \widetilde{X}_j^{(\le u)},$$

where $\widetilde{X}_1^{(\le u)}, \dots, \widetilde{X}_n^{(\le u)}$ are independent identically distributed random variables,

$$\widetilde{X}_j^{(\le u)} = \begin{cases} X_j^{(\le u)} & \text{with probability } F(u) - F(-u); \\ 0 & \text{with probability } F(-u) + 1 - F(u). \end{cases}$$

Then in accordance with (3), the parameter a can be defined as

$$a = a(u) = n\mathsf{E}\widetilde{X}_1^{(\le u)} = n[F(u) - F(-u)]\mathsf{E}X_1^{(\le u)},$$

and the parameter σ^2 can be defined as

$$\sigma^2 = \sigma^2(u) = \mathsf{D}\widetilde{X}_1^{(\le u)} = [F(u) - F(-u)]\mathsf{D}X_1^{(\le u)} + [F(-u) + 1 - F(u)]\big(\mathsf{E}X_1^{(\le u)}\big)^2.$$

With these values of a and σ the first term on the right-hand side of (5) will tend to zero by the central limit theorem as $n \to \infty$, and can be estimated by the standard techniques, say, by the Berry–Esseen inequality for binomial random sums, see [5, 4].

As regards the second term on the right-hand side of (5), with large u, $p = F(u) - F(-u)$ close to one and moderate (but large enough) n the term η_n may be small due to that the sum $S_n^{(>u)}$ contains very few summands. Moreover, in the case of light tails, putting $u = u_n$ so that $n[F(-u) + 1 - F(u_n)] \to 0$ as $n \to \infty$, it is possible to make sure that the right-hand side of (5) can be made arbitrarily small by the choice of arbitrarily large n so that the limit distribution for the normalized sum S_n^* will be normal, see the details in Section 4.

Under some additional conditions, at the expense of introducing additional parameter, the dependence of the second term of the bound given in Theorem 1 on n can be made better.

For $c \in (0, 2]$ let $h(c) = \mathbb{I}_{(1,2]}(c)$.

THEOREM 2. *Assume that the distribution function $F(x)$ belongs to the domain of attraction of a stable law with characteristic exponent $\alpha \in (0, 2)$. If, moreover, $\alpha \ge 1$, then additionally assume that F is symmetric (that is, $F(-x) = 1 - F(x)$ for $x > 0$). Then for any $u > 0$ $\epsilon > 0$ and $\delta \in (0, \alpha)$ we have*

$$\rho(F_{\zeta_n+\eta_n},\, \Phi_{a,\,\sigma}) \le \rho(F_{\zeta_n},\, \Phi_{a,\,\sigma}) + \big[2\Phi_{0,\sigma}\big(\tfrac{\epsilon}{2}\big) - 1\big] + 2^{h(\delta)}\epsilon^{-\delta}n^{1-\delta/2}\big(F(-u) + 1 - F(u)\big)\mathsf{E}\big|X_1^{(>u)}\big|^{\delta}. \quad (14)$$

The proof is based on Theorem 1 and Lemma 5.

We see that in (14) the exponent of n is less than that in (5). However, in (14) an additional parameter ϵ appeared. The second term on the right-hand side of (14) can be made arbitrarily small by the appropriate choice of ϵ. With n and ϵ fixed, the third term on the right-hand side of (14) can be made arbitrarily small by the choice of u large enough.

Actually Theorems 1 and 2 are simple variants of a so-called pre-limit theorem, see [2].

Now consider the problem of determination of n_0 such that for n growing from 1 to n_0 the distance $\rho(F_{\zeta_n+\eta_n},\, \Phi_{a,\sigma})$ decreases and for $n > n_0$ this distance increases. Assume that the first summand on the right-hand side of (5) with $a = a(u)$ and $\sigma^2 = \sigma^2(u)$ is estimated by the Berry–Esseen inequality with some $\gamma \in (0, 1]$:

$$\rho(F_{\zeta_n},\, \Phi_{a(u),\sigma(u)}) \le \frac{C(\gamma)\widetilde{L}_{2+\gamma}^{(\le u)}}{n^{\gamma/2}}, \quad (6)$$

where $\widetilde{L}_{2+\gamma}^{(\le u)}$ is the Lyapunov fraction of order $2 + \gamma$,

$$\widetilde{L}_{2+\gamma}^{(\le u)} = \frac{\mathsf{E}\big|\widetilde{X}_1^{(\le u)} - \mathsf{E}\widetilde{X}_1^{(\le u)}\big|^{2+\gamma}}{\big(\mathsf{D}\widetilde{X}_1^{(\le u)}\big)^{1+\gamma/2}},$$

$C(\gamma) > 0$ is the absolute constant, for example, $C(1) \le 0.4690$ [7]. It is easy to verify that if $c > 0$, $d > 0$, then

$$\arg\min_{z>0}\Big(\frac{c}{z^{\gamma/2}} + dz\Big) = \Big(\frac{\gamma c}{2d}\Big)^{\frac{2}{2+\gamma}}.$$

Putting $z = n$, $c = C(\gamma)\widetilde{L}_{2+\gamma}^{(\le u)}$, $d = F(-u) + 1 - F(u)$, we see that the minimum of the upper bound for $\rho(F_{\zeta_n+\eta_n},\, \Phi_{a(u),\sigma(u)})$ is attained either at n_γ which is the integer part of

$$m_\gamma = \left[\frac{\gamma C(\gamma)\widetilde{L}_{2+\gamma}^{(\le u)}}{2\big(F(-u) + 1 - F(u)\big)}\right]^{\frac{2}{2+\gamma}},$$

or at $n_\gamma + 1$.

Substituting (6) with $n = n_\gamma$ in (5) we arrive at the following result.

THEOREM 3. *For any $u > 0$, $\gamma \in (0, 1]$*

$$\min_n \rho(F_{\zeta_n+\eta_n}, \Phi_{a(u),\sigma(u)}) \le (2+\gamma) \cdot \left[\frac{C(\gamma)\widetilde{L}_{2+\gamma}^{(\le u)}}{2^{\gamma/2}\gamma}\right]^{\frac{2}{2+\gamma}} \cdot \big(F(-u) + 1 - F(u)\big)^{\frac{\gamma}{2+\gamma}}.$$

If $\mathsf{E}|X_1|^{2+\gamma} < \infty$, then

$$\lim_{u\to\infty} \widetilde{L}_{2+\gamma}^{(\le u)} = L_{2+\gamma} \equiv \frac{\mathsf{E}|X_1 - \mathsf{E}X_1|^{2+\gamma}}{(\mathsf{D}X_1)^{1+\gamma/2}} < \infty$$

so that, according to Theorem 3,

$$\lim_{u\to\infty} \min_n \rho(F_{\zeta_n+\eta_n}, \Phi_{a(u),\sigma(u)}) = 0.$$

In the case $\gamma = 0$, instead of (6) the bounds obtained in [4] can be used to obtain results similar to Theorem 3.

References

[1] *B. von Bahr, C.-G. Esseen.* Inequalities for the rth absolute moment of a sum of random variables, $1 \le r \le 2$ // Annals of Mathematical Statistics, 1965. Vol 36. No. 1. P. 299–303.

[2] *L. B. Klebanov, S. T. Rachev, G. J. Szekely.* Pre-limit theorems and their applications // Acta Applicandae Mathematicae, 1999. Vol 58. P. 159–174.

[3] *V. Yu. Korolev.* On the distribution of the ratio of the sum of sample elements exceeding a certain threshold to the sum of all sample elements. I // Informatics and Its Applications, 2020. Vol. 14. No. 3. P. 26–34.

[4] *V. Yu. Korolev, A. V. Dorofeeva.* Bounds of the accuracy of the normal approximation to the distributions of random sums under relaxed moment conditions // Lithuanian Mathematical Journal, 2017. Vol. 57. No. 1. P. 38–58.

[5] *V. Yu. Korolev, I. G. Shevtsova.* An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums // Scandinavian Actuarial Journal, 2012. No. 2. P. 81–105.

[6] *V. V. Petrov.* Sums of independent random variables. – Moscow: Nauka, 1972.

[7] *I. G. Shevtsova.* On the absolute constants in the Berry–Esseen-type inequalities // Doklady Mathematics, 2014. Vol. 456. No. 6. P. 650–654.