ABOUT THE CALCULATION OF THE ABSOLUTE CONSTANT IN THE BERRY – ESSEEN INEQUALITY FOR TWO-POINT DISTRIBUTIONS

Anatolii Zolotukhin*, Sergei Nagaev**, Vladimir Chebotarev***

*Tula State University, Tula
**Sobolev Institute of Mathematics, Novosibirsk,
***Computing Center, FEB RAS, Khabarovsk

1. The Berry – Esseen inequality. Let X, X_1, X_2, $\ldots$, X_n be i.i.d. random variables with a finite third moment,

$$\mathbf{E}X = 0, \qquad \mathbf{E}X^2 = 1.$$

Denote

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\, dt, \qquad \beta_3 = \mathbf{E}|X|^3.$$

The Berry – Esseen inequality (1941–42): for all $x \in \mathbb{R}$ and $n = 1, 2, \ldots$

$$\left| \mathbf{P}\Big(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j < x \Big) - \Phi(x) \right| \le \frac{C_0 \beta_3}{\sqrt{n}},$$

where C_0 is an absolute constant.

2. Bounds of the constant C_0. The first upper bounds of the constant C_0 were obtained by the following mathematicians: C.-G. Esseen (1942), H. Bergström (1949) и K. Takano (1951).

In 1956 C.-G. Esseen [1: Esseen, 1956] found a two-point distribution, for which

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\beta_3} \sup_{x\in\mathbb{R}} \left| \mathbf{P}\Big(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j < x \Big) - \Phi(x) \right| = C_E, \tag{1}$$

where

$$C_E := \frac{3+\sqrt{10}}{6\sqrt{2\pi}} = 0.40973218\ldots\ .$$

Consequently, $C_0 \geq C_E$. The result of Esseen served as an argument for the congecture

$$C_0 = C_E, \tag{2}$$

V.M. Zolotarev made in 1966 [2: Zolotarev, 1966]. The question if this conjecture is correct remains open up to now.

The following mathematicians obtained upper bounds for C_0 after the result of Esseen: Zolotarev V.M. (1966, 1967), Zahl S. (1966), van Beek P. (1972), Prawitz H. (1975), Shiganov I.S. (1982), Chistyakov G.P. (2001), Nagaev S.V., Chebotarev V.I. (2004), Korolev V.Yu., Shevtsova I.G. (2005, 2009), Shevtsova I.G. (2006, 2013), Tyurin I.S. (2009, 2010).

The best upper bound, known to date, belongs to Shevtsova: $C_0 \leq 0.469$.

3. Bounds for the constant C_{02}. The work of Esseen [1] also served as an impetus to the study the problem of the constant C_0 in the special case of two-point distributions. In this case we shall write C_{02} instead of C_0.

In 2007 Hipp C., Mattner L. [4: Hipp&Mattner, 2007] obtained the following inequality in the case of symmetric distribution of the Bernoulli random variables,

$$C_{02} \leq \frac{1}{\sqrt{2\pi}} \quad \text{(symmetric case)}.$$

In 2016 Schulz J. [5: Schulz, 2016] proved that if the symmetry condition is violated, then the following equality holds,

$$C_{02} = C_E \quad \text{(asymmetric case)}.$$

4. About bounds of the constant C_{02} (2010-2011). It should be noted that back in 2011 Nagaev and Chebotarev [6: Nagaev&Chebotarev, 2011] (Theorem 1.1 and Corollary 1.1) obtained a new bound of the error of the normal approximation for the binomial distribution.
The main part of the majorant of the error found there contains as a factor the function

$$\mathcal{E}(p) = \frac{2-p}{3\sqrt{2\pi}\,[p^2+(1-p)^2]},$$

where p is the parameter of the initial random variable, $0 \le p \le 0.5$. The remainder part of the majorant is positive, decreases in n and tends to zero faster of the main part when $n \to \infty$.

The maximum value of the function $\mathcal{E}(p)$ is equal to the constant C_E.

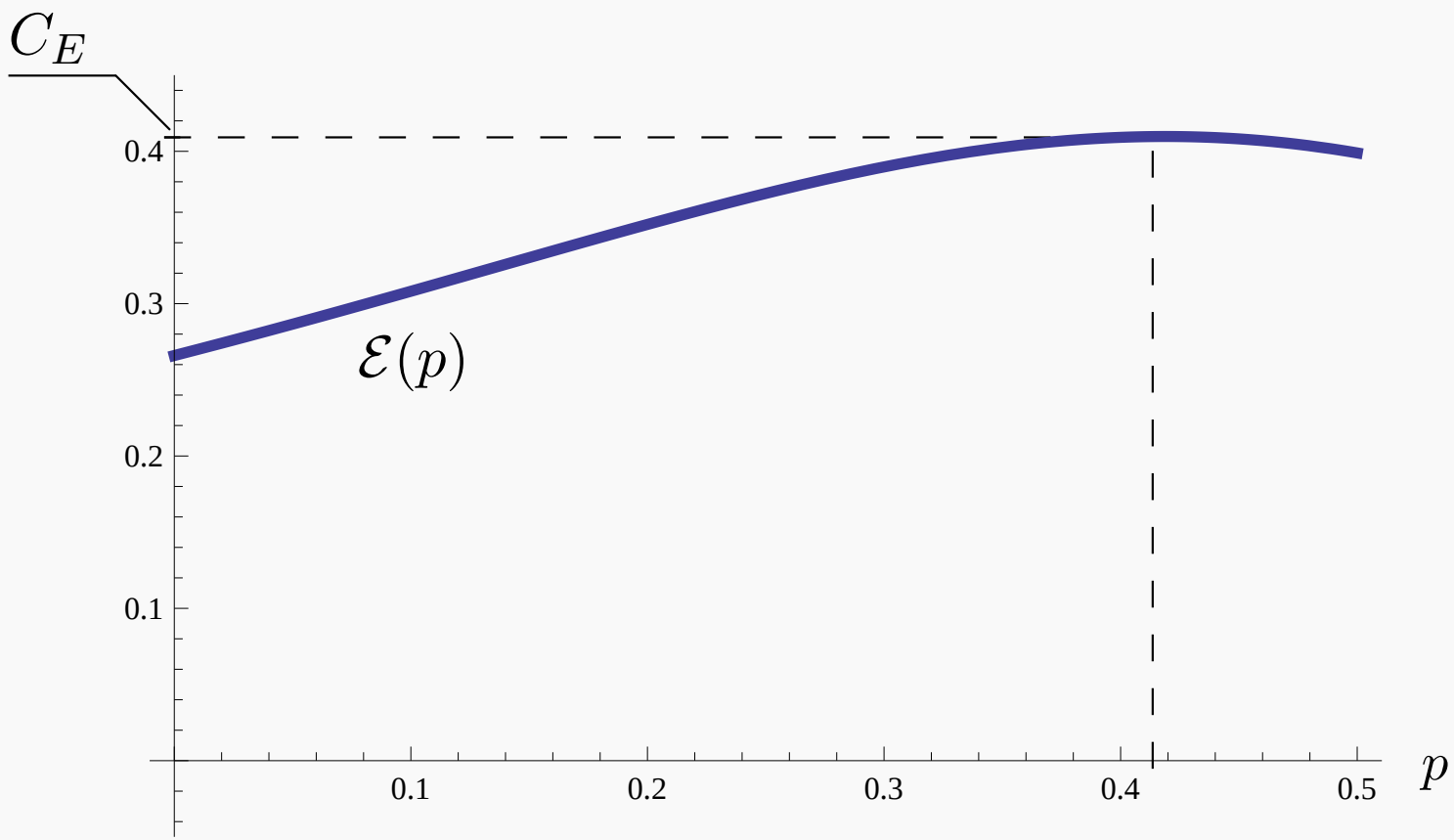

Figure 1. The graph of the function $\mathcal{E}(p)$

Let X, X_1, X_2, $\ldots$, X_n be i.i.d. random variables with the following distribution:

$$\mathbf{P}(X=1)=p, \quad \mathbf{P}(X=0)=q=1-p.$$

In what follows we shall use the following notations:

$$F_{n,p}(x)=\mathbf{P}\Big(\sum_{i=1}^{n} X_i < x\Big), \qquad G_{n,p}(x)=\Phi\Big(\frac{x-np}{\sqrt{npq}}\Big),$$

$$\Delta_n(p)=\sup_{x\in\mathbb{R}}\big|F_{n,p}(x)-G_{n,p}(x)\big|, \qquad \varrho(p)=\frac{\mathbf{E}|X-p|^3}{(\mathbf{D}X)^{3/2}}=\frac{p^2+q^2}{\sqrt{pq}},$$

$$K_n(p)=\frac{\Delta_n(p)\sqrt{n}}{\varrho(p)}.$$

The quantity

$$\sup_{n\geq 1}\sup_{p\in(0,0.5]} K_n(p)$$

is the absolute constant in the Berry – Esseen inequality if to consider it not for all distributions with finite absolute third moment but only for the class of two-point distributions.

In this paper we solve the problem of computing the quantity

$$\max_{1\leq n\leq 500000}\max_{p_j\in S} K_n(p_j),$$

where S is a grid on $(0,\, 0.5]$. Using this numerical result we prove the following bound,

$$\max_{1\leq n\leq 500000}\max_{p\in(0,\,0.5]} K_n(p) < C_E.$$

Then, applying Corollary 1.1 [6], we obtain the inequality

$$C_{02} \leq 0.4099539 = C_E + \varepsilon,$$

where $\varepsilon < 0.000222$.

Let us formulate Corollary 1.1 from [6: Nagaev&Chebotarev, 2011].

Corollary 1.1 [6]. *Let the following conditions are fullfiled:* $\frac{4}{n} \le p \le 0.5$ *and* $n \ge 200$. *Then there exists a function* $E(p,n)$ *possessing the following properties:*
1) $K_n(p) \le E(p,n)$,
2) for each fixed $p \in (0, 0.5]$ *the sequence* $E(p,n)$ *converges to* $\mathcal{E}(p)$, *monotonically decreasing in* $n \ge \max\left\{200, \frac{4}{p}\right\}$.

The function $E(p,n)$ is defined with the help of Theorem 1.1 [6], and has a rather cumbersome form.

An illustration to Corollary 1.1 [6] is the Figure 2. It shows the mutual arrangement of the functions: $K_n(p)\big|_{n=50}$, $E(p,n)$ for $n = 200$ and $n = 800$, and also $\mathcal{E}(p)$.

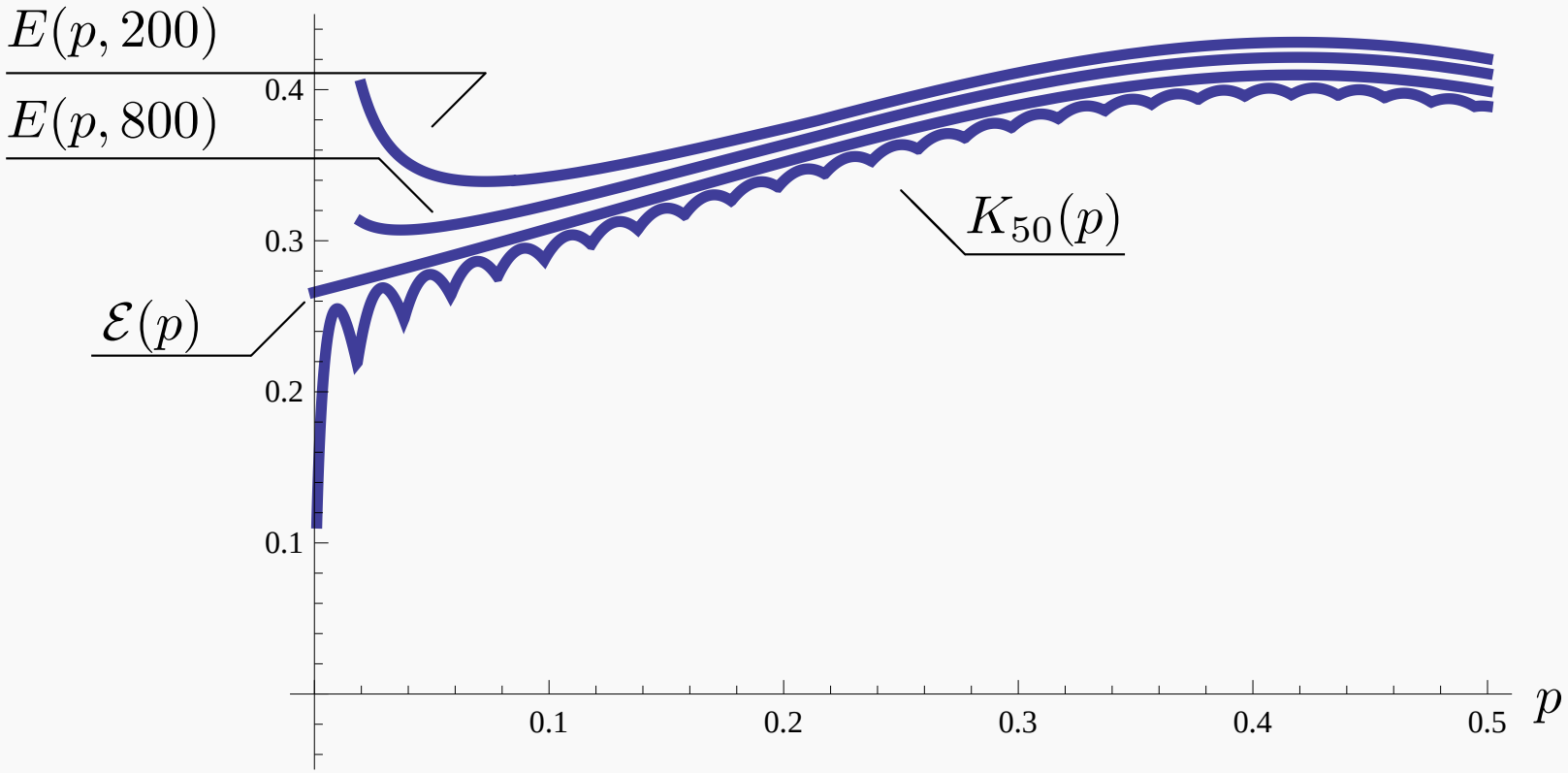

Figure 2. Graphs of the functions: $K_{50}(p)$, $E(200,p)$, $E(800,p)$, $\mathcal{E}(p)$

By virtue of Corollary 1.1 [6] the sequence $E(p, n)$, decreasing, convereges to $\mathcal{E}(p)$ in each point $p \in (0, 0.5)$. On the other hand, our calculations allows to make the assumption that for all $p \in (\delta, 0.5)$ (for some $\delta > 0$) the sequence $K_n(p)$ converges to $\mathcal{E}(p)$ as well, remaining less of its limit. An illustration to this is Figure 2.

Since the maximum value of the function $\mathcal{E}(p)$ is equal to the constant C_E, then Corollary 1.1 [6] allows to find upper bounds for C_{02} that are arbitrarily close to C_E, under condition that n is large enough.

We shall write $C_{02}(\underline{N})$, instead of C_{02}, if it is assumed that $n \geq N$, and $C_{02}(\overline{N})$, if $n \leq N$.

Corollary 1.1 [6] enables us to find such a N_ε that the following bound holds,

$$C_{02}(\underline{N_\varepsilon}) \leq C_E + \varepsilon.$$

In particular, it was proved in [6], that if $\varepsilon = 0.4215 - C_E = 0.0117678\ldots$, then $N_\varepsilon = 200$, i. e.

$$C_{02}(\underline{200}) \leq 0.4215.$$

In addition to this analytical result, calculations were made in 2010 in preprint [7: Kondrik&Mikhaylov&Nagaev&Chebotarev,2010], which show that

$$C_{02}(\overline{200}) < C_E.$$

Thus, in 2011 г. the following bound was known: for every $n \geq 1$,

$$C_{02} \leq 0.4215 = C_E + \varepsilon, \quad \varepsilon < 0.01177.$$

Calculations were carried out on a personal computer. However, for sufficiently small $\varepsilon > 0$, the number N_ε can be so large that more powerful computing equipment will be required.

We note that in obtaining upper bounds of C_0, beginning with the estimates of V.M. Zolotarev [2] and [3: Zolotarev, 1967], an important role was played by computer technology.

5. On the problem of calculating the constant C_{02}. Let $\varepsilon = 0.4099539 - C_E = 0.0002217\ldots$. In virtue of Corollary 1.1 [6] we have $N_\varepsilon = 500000$ and

$$C_{02}(\underline{N_\varepsilon}) \le 0.4099539.$$

To prove the validity of the bound

$$C_{02} \le 0.4099539 \tag{3}$$

(for all $n \ge 1$), it remains to verify that $C_{02}(\overline{500000}) \le 0.4099539$. In fact, as calculations and additional analysis have shown, the stronger inequality holds,

$$C_{02}(\overline{500000}) < C_E.$$

Calculations were carried out on the supercomputer Blue Gene/P, Computing Center, Faculty of Computational Mathematics and Cybernetics, Moscow State University.

6. Main result. Briefly on the proof. One can show that under fixed n and p the supremum $\sup\limits_{x\in\mathbb{R}} |F_{n,p}(x) - G_{n,p}(x)|$ is attained in some discontinuity point of the function $F_{n,p}(x)$. We consider distribution functions which are continuous from the left. Consequently,

$$\Delta_n(p) = \max_{0\le i\le n} \Delta_{n,i}(p), \tag{4}$$

where $\Delta_{n,i}(p) = \max\left\{\left|F_{n,p}(i) - G_{n,p}(i)\right|, \left|F_{n,p}(i+1) - G_{n,p}(i)\right|\right\}$, i are integers.

Note that we can vary the parameter p in a narrower interval than $[0, 0.5]$, which is separated from zero, namely, in

$$I := [0.1689, 0.5].$$

The basis for this conclusion is the following statement.

Lemma 1. *If* $0 < p \le 0.1689$, *then*

$$K_n(p) < 0.4096 \tag{5}$$

fo every $n \ge 1$.

Remark. Lemma 1 is proved with the help of the modified Berry – Esseen with numerical constants found in [8: Korolev&Shevtsova, 2010]. Note that the following bound was obtained in [9, Theorem 8.2: Nagaev&Chebotarev, 2009],

$$K_n(p) \leq 0.37128$$

for $0 < p < 0.03$ and $np > 6$. Its proof is based on the error estimate of the Poisson approximation for the binomial distribution when p is close to zero. We also use the bound

$$\Delta_n(p) \leq 0.541, \quad n \geq 1, \quad 0 \leq p \leq 1,$$

which is a consequence of the universal estimate of the uniform distance between the distribution of the normalized sum of i.i.d. random variables and the standard normal distribution, refined in [10: Chebotarev&Kondrik&Mikhaylov, 2007].

Denote by S the uniform grid on I with the step $h = 10^{-12}$.

The result of calculations. *For all* $1 \leq n \leq 500000$,

$$\max_{p_j \in S} K_n(p_j) < 0.40973214. \tag{6}$$

The counting algorithm is a triple cycle: the cycle with respect to the parameter i (see (4)) is nested in the cycle with respect to the parameter p, which in turn is nested in the cycle with respect to the parameter n.

With increasing n, the computation time increased rapidly. For instance, for $2000 \leq n \leq 2100$, the computations took more than 3 hours on the computer with the processor Core2Due E6400. For $2101 \leq n \leq 500000$ the calculations were performed on the supercomputer Blue Gene/P.

It follows from [11, Corollary 7: Nagaev&Chebotarev&Zolotukhin, 2016] that for $n > 200$ in the cycle with respect to i one can take into account not all i from 0 up to n, but only that satisfy the inequality

$$np - (\nu + 1)\sqrt{npq} \leq i \leq np + \nu\sqrt{npq},$$

where $\nu = \sqrt{3 + \sqrt{6}}$. This led to a significant reduction in computation time. For example, for the interval $10000 \leq n \leq 11024$ the computation time was about 3 minutes (without account for the waiting a queue). Note that the calculation on the supercomputer lasted 7 hours for $490000 \leq n \leq 500000$. The program is written in the programming language C+MPI and registered [12: Zolotukhin, 2015].

The following statement allows to make the conclusion about values of $K_n(p)$, when $p \notin S$.

Theorem 1. *If $p \in I$, and p' is the node of the grid S, which is closest to p, then for all $1 \le n \le 500000$,*

$$|K_n(p) - K_n(p')| \le 1.1 \cdot 10^{-8}.$$

It follows from Theorem 1, Lemma 1 and (6) that for all $1 \le n \le 500000$ and $p \in (0, 0.5]$ the following bound holds,

$$K_n(p) < 0.40973216.$$

It is easy to justify that this inequality is also true for $p \in (0.5, 1)$. From here and Corollary 1.1 [6] we obtain

Theorem 2. *The bound* (3) *holds, i.e.*

$$C_{02} \le 0.4099539 = C_E + \varepsilon,$$

where $\varepsilon < 0.000222$.

REFERENCES

1. [1]*Esseen C.-G.* A moment inequality with an application to the central limit theorem // Scand. Aktuarietidskr. J. — 1956. — Vol. 3–4. — P. 1–170.
2. [2] *Zolotarev V.M.* An absolute estimate of the remainder term in Central Limit Theorem // Teor. Veroyatn. i Primen. — 1966. — V. 11, No. 1. — P. 108–119. (In Russian)
3. [3] *Zolotarev V. M.* A sharpening of the inequality of Berry – Esseen // Z. Wahrscheinlichkeitstheor. verw. Geb. — 1967. — Vol. 8. — P. 332–342.
4. [4] *Hipp C., Mattner L.* On the normal approximation to symmetric binomial distribution // Teor. Veroyatn. i Primen. — 2007. — V. 52, No. 3. — P. 610–617.
5. [5] *Schulz J.* The optimal Berry – Esseen constant in the binomial case. – Dissertation. Universität Trier, 2016.

9. [6] *Nagaev S. V., Chebotarev V. I.* On the bound of proximity of the binomial distribution to the normal one // Theory Probab. Appl. — 2012. —Vol. 56, No. 2, 213–239. Original Russian Text@ Teor. Veroyatn. i Primen. — 2011. — 56, No 2, P. 248–278.
10. [7] *Kondrik A. S., Mikhaylov K. V., Nagaev S. V., Chebotarev V. I.* On the bound of closeness of the binomial distribution to the normal one for a limited number of observations. — Research Report No. 2010/160 . Khabarovsk: Computing Centre FEB RAS, 2010. 19 pp. (In Russian)
11. [8] *Korolev V. Yu., Shevtsova I. G.* An improvement of the Berry – Esseen inequality with applications to Poisson and mixed Poisson random sums // Survey of Applied and Industrial Mathematics. — 2010. — V. 17, No. 1. — P. 25–56. (In Russian)
7. [9] *Nagaev S. V., Chebotarev V. I.* On the bound of closeness of the bianomial distribution to the normal one. — Research Report No 2009/142 . Khabarovsk: Computing Centre FEB RAS, 2009, 47 pp. (In Russian)
11. [10] *Chebotarev V.I., Kondrik A.S., Mikhaylov K.V.* On an extreme two-point distribution // http://arxiv.org/abs/0710.3456, 18 Oct

11. [11] *S. V. Nagaev, V. I. Chebotarev, A. Ya. Zolotukhin.* On a non-uniform bound of the remainder term in Central Limit Theorem for Bernoulli random variables // Journal of Mathematical Sciences. — 2016. — Vol. 214, no. 1. — P. 83–100.
11. [12] *Zolotukhin A. Ya.* Program for calculating the estimate of the main term in the Central Limit Theorem for the Bernoulli distribution for a limited number of observations. – Certificate of the state registration of the computer program No. 2015617151, 01.06.2015. (In Russian)
6. [13] *Korolev V. Yu., Shevtsova I. G.* On the upper bound of the absolute constant in the Berry – Esseen inequality // Theory of Probability and Its Applications. —2010. —V. 54, No. 4, 638–658. Original Russian Text@ Teor. Veroyatn. i Primen. — 2009. — V. 53, No 4, P. 671–695.
8. [14] *Shevtsova I. G.* Optimization of the structure of moment estimates of the accuracy of normal approximation for distributions of sums of independent random variables. – Thesis for the degree of Doctor of Physical and Mathematical Sciences. Moscow: MSU, 2013. (In Russian)