ON THE BOUND OF PROXIMITY OF THE BINOMIAL DISTRIBUTION TO THE NORMAL ONE *

S.V. Nagaev
Sobolev Institute of Mathematics, Novosibirsk
e-mail: nagaev@math.nsc.ru

V.I. Chebotarev
Computing Center of FEB RAS, Khabarovsk
e-mail: chebotarev@as.khb.ru

April 2011

Bounds for the error of the Gaussian approximation for the binomial distribution are stated, depending from the probability of success and the number n of observations. As a consequence, the upper bound for the absolute constant in the Berry–Esseen inequality for identically distributed random variables, taking two values, is deduced which differs from asymptotical one slightly more than 0.01.

The following idea is realized in the work. We can obtain sharp bounds for sufficiently large n. The main purpose of the paper is to prove just these bounds. As to bounded number of observations, computations with the help of the computer must be produced. This part of investigations is developed by our pupils K.V. Mikhailov and A.S. Kondric.

Let $X, X_1, ..., X_n$ be a sequence of i.i.d. random variables with $\mathbf{E}X = 0$, $\beta_3 = \mathbf{E}|X|^3 < \infty$. Denote $b^2 = \mathbf{E}X^2$, $S_n = \sum_{i=1}^n X_i$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2}dt$.

A. Berry [1] and C.-G. Esseen [2] proved that

$$\Delta_n := \sup_x |\mathbf{P}(n^{-1/2}S_n < bx) - \Phi(x)| < C_0 \frac{\beta_3}{b^3\sqrt{n}},$$

where C_0 is an absolute constant.

The large amount of papers is devoted to the search of the optimal value of the constant C_0 (see, e.g. [3–12]). Esseen [13] showed that C_0 can not be less than $C_E = \frac{\sqrt{10}+3}{6\sqrt{2\pi}} = 0.409732\ldots$.

As to upper bounds for C_0 the best results in this direction were obtained in the recent papers by I.S. Tyurin [9, 10], $C_0 \le 0.4785$, and V. Yu. Korolev, I.G. Shevtzova [11, 12],

$$C_0 \le 0.4784. \tag{1}$$

In reality, sharper result

$$\Delta_n \le 0.33477 \frac{\beta_3 + 0.429b^3}{b^3\sqrt{n}}, \tag{2}$$

is obtained in [12] from which (1) easily follows.

*Siberian Branch of RAS (no. 30), Far Eastern Branch of RAS (09-II-SB-01-003, 09-I-BMS-02).

In the present paper we give the bound for Δ_n and C_0 in the particular case when X takes only two values. To formulate our results we introduce a lot of notations.

Thus, let $\mathbf{P}(X=a)=q$, $\mathbf{P}(X=d)=p$, where $p+q=1$, $a<0<d$, $\mathbf{E}X=0$. We assume for the brevity that $b^2=1$. Then

$$\beta_3=\frac{p^2+q^2}{\sqrt{pq}},\quad \mathbf{E}X^3=\frac{q-p}{\sqrt{pq}},\quad d-a=\frac{1}{\sqrt{pq}}. \tag{3}$$

Define the function $\mathcal{E}(p)$ by the equality

$$\mathcal{E}(p)=\frac{1}{\beta_3\sqrt{2\pi}}\Big(\frac{\mathbf{E}X^3}{6}+\frac{d-a}{2}\Big). \tag{4}$$

Note that the right-hand side of (4) first appeared in the paper by Esseen [13]. It is easily seen, using formulas (3), that

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

Denote

$$\sigma=\sigma(p,n)=\sqrt{npq},$$

$$\omega_3(p)=q-p,\quad \omega_4(p)=|q^3+p^3-3pq|,$$
$$\omega_5(p)=q^4-p^4,\quad \omega_6(p)=q^5+p^5+15(pq)^2.$$

Let

$$K_1(p,n)=\frac{\omega_3(p)}{4\sigma\sqrt{2\pi}(n-1)}\Big(1+\frac{1}{4(n-1)}\Big)+\frac{\omega_4(p)}{12\sigma^2\pi}\Big(\frac{n}{n-1}\Big)^2+$$
$$+\frac{\omega_5(p)}{40\sigma^3\sqrt{2\pi}}\Big(\frac{n}{n-1}\Big)^{5/2}+\frac{\omega_6(p)}{90\sigma^4\pi}\Big(\frac{n}{n-1}\Big)^3.$$

Further, denote

$$\zeta(p)=\Big(\frac{\omega(p)}{6}\Big)^{2/3},\quad e(n,p)=\exp\Big\{\frac{1}{24\sigma^{2/3}\zeta^2(p)}\Big\},\quad e_5=0.0277905,$$
$$\widetilde{\omega}_5(p)=p^4+q^4+5!e_5(pq)^{3/2},\quad A_k(n)=\Big(\frac{n}{n-2}\Big)^{k/2}\,\frac{n-1}{n},$$
$$V_6(p)=\omega_3^2(p),\quad V_7(p)=\omega_3(p)\,\omega_4(p),\quad V_8(p)=\frac{2\,\widetilde{\omega}_5(p)\,\omega_3(p)}{5!\,3!}+\frac{\omega_4^2(p)}{(4!)^2},$$
$$V_9(p)=\widetilde{\omega}_5(p)\,\omega_4(p),\quad V_{10}(p)=\frac{2^6\cdot 3}{(5!)^2}\,\widetilde{\omega}_5^2(p),$$

$$\gamma_6=\frac{1}{9},\quad \gamma_7=\frac{5\sqrt{2\pi}}{96},\quad \gamma_8=24,\quad \gamma_9=\frac{7\sqrt{2\pi}}{4!\,16},\quad \gamma_{10}=\frac{2^6\cdot 3}{(5!)^2},$$
$$\widetilde{\gamma}_6=\frac{2}{3},\quad \widetilde{\gamma}_7=\frac{7}{8},\quad \widetilde{\gamma}_8=\frac{10}{9},\quad \widetilde{\gamma}_9=\frac{11}{8},\quad \widetilde{\gamma}_{10}=\frac{5}{3}.$$

Let

$$K_2(p,n) = \frac{1}{\pi\sigma}\sum_{j=1}^{5}\frac{\gamma_{j+5}\,A_{j+5}(n)\,V_{j+5}(p)}{\sigma^j}\left[1+\frac{\widetilde{\gamma}_{j+5}\,e(n,p)\,n}{\sigma^2\,(n-2)}\right].$$

Finally, put

$$K_3(p,n) = \frac{1}{\pi}\Bigg\{\frac{1}{12\sigma^2}+\Big(\frac{1}{36}+\frac{\mu}{8}\Big)\frac{1}{\sigma^4}+\Big(\frac{e^{A_1/6}}{36}+\frac{\mu}{8}\Big)\frac{1}{\sigma^6}+\frac{5\mu\,e^{A_2/6}}{24\sigma^8}+$$
$$+\frac{1}{3}\,e^{-\sigma\sqrt{A_1}+A_1/6}+(\pi-2)\,\mu\,e^{-\sigma\sqrt{A_2}+A_2/6}+\frac{1}{4}\,e^{-\sigma\sqrt{A_3}+A_3/6}\,\ln\Big(\frac{\pi^4\sigma^2}{4A_3}\Big)+$$
$$+\exp\Big\{-\frac{\sigma^{2/3}}{2\zeta(p)}\Big\}\left[\frac{2\zeta(p)}{\sigma^{2/3}}+e^{A_3/6}\frac{1+\chi(p,n)}{24\,\zeta(p)\,\sigma^{4/3}}\right]\Bigg\},$$

where

$$A_1 = 5.40466, \quad A_2 = 7.52058, \quad A_3 = 5.2335, \quad \mu = \frac{3\pi^2-16}{\pi^4},$$
$$\chi(p,n) = \begin{cases} \frac{2\zeta(p)}{\sigma^{2/3}} & \text{if } 0<p<0.085,\\ 0 & \text{if } 0.085\le p\le 0.5. \end{cases}$$

Denote

$$R_0(p,n) = \frac{\sqrt{n}}{\beta_3(p)}\sum_{j=1}^{3}K_j(p,n).$$

Denote also the values Δ_n and β_3 for given p by $\Delta_n(p)$ and $\beta_3(p)$ respectively.

Theorem 1. *If*

$$\frac{4}{n}\le p\le 0.5, \quad n\ge 200, \tag{5}$$

then

$$\frac{\sqrt{n}}{\beta_3(p)}\,\Delta_n(p)\le \mathcal{E}(p)+R_0(p,n),$$

and for every fixed p ($0<p\le 0.5$) *the sequence* $R_0(p,n)$ *tends to 0 decreasing in* $n\ge\max\big\{200,\frac{4}{p}\big\}$.

Above stated formulas for $K_i(p,n)$, $i=\overline{1,3}$, by which $R_0(p,n)$ is expressed, are very complicated. Of course, we can estimate $K_i(p,n)$ from above by simpler expressions, but doing so we loose much in exactness.

Proving Theorem 1, we applied the smoothing method as in almost all papers devoted to estimating a constant in the Berry–Esseen inequality. However, in difference with the traditional, after paper [4] by S. Zahl, smoothing by means of signed measures, we apply, with this purpose, the uniform distribution on the interval $\big(-\frac{1}{2\sqrt{pq}},\frac{1}{2\sqrt{pq}}\big)$.

Denote $p_0=\frac{4-\sqrt{10}}{2}=0.418861\ldots$. One can show that $\mathcal{E}(p)$ increases for $0<p<p_0$ and decreases for $p_0<p\le 0.5$, i.e. p_0 is the point of maximum of $\mathcal{E}(p)$, and $\mathcal{E}(p_0)=\frac{\sqrt{10}+3}{6\sqrt{2\pi}}\equiv C_E$.

By virtue of Theorem 1, for $n \geq 200$ the problem is reduced to finding $M \equiv \max_{p\in[0.02,0.5]} E(p,200)$. This is practically impossible to realize without using a computer in view of extreme complication of the function $E(p,n)$. Two ways are applied for solving the problem, which give the results, differing one from the other not more than by $8 \cdot 10^{-5}$.

The first way is that computations of $E(p,200)$ are produced in eleven values of p only. The function $E(p,200)$ is estimated above in each of ten intervals, formed by the selected points. Monotonicity in these intervals of all 23 functions, defining $E(p,n)$, is used here. As a result we obtain the bound $M < 0.421498$, which is formulated below as Corollary.

Creation of a code for computing M using a lattice with the step 10^{-4} in $[0.02, 0.5]$ is the alternative way. The bound $M < 0.421421$ is obtained by this method.

Note that the advantage of the first way is the considerably lesser volume of computations.

Figure 1 illustrates behaviour of $E(p,n)$ and $\mathcal{E}(p)$.

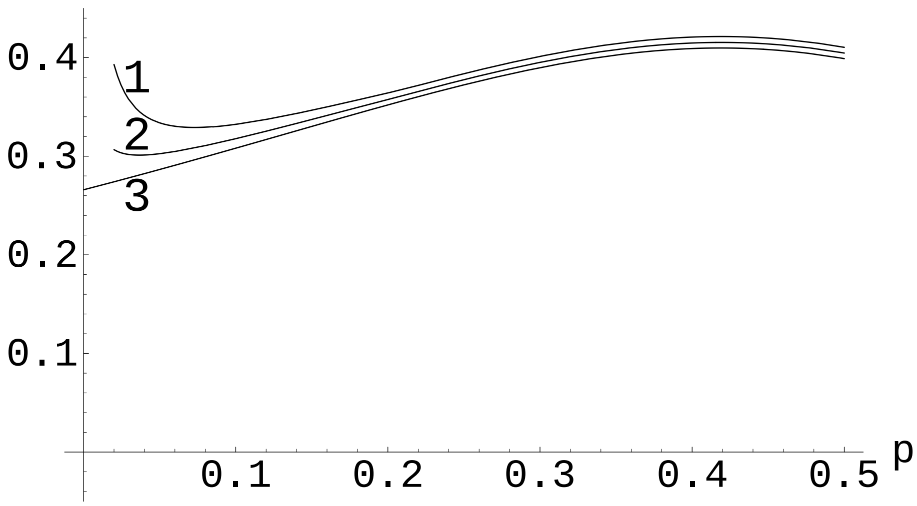

Fig. 1. 1 – graph of $E(p,200)$, $p\in[0.02,0.5]$; 2 – graph of $E(p,800)$, $p\in[0.02,0.5]$; 3 – graph of $\mathcal{E}(p)$, $p\in[0,0.5]$

Corollary. *For n and p, satisfying* (5),

$$\sup_{0.02\leq p<0.5} \frac{\sqrt{n}}{\beta_3(p)}\,\Delta_n(p) < 0.4215. \tag{6}$$

On the other hand, K.V. Mikhailov and A.S. Kondric found in [14] that

$$\max_{1\leq n\leq 200}\ \sup_{0.02\leq p\leq 0.5} \frac{\sqrt{n}}{\beta_3(p)}\,\Delta_n(p) < 0.4096. \tag{7}$$

Now, let $0 < p < 0.02$. In this case, it follows from bound (2) that

$$\frac{\sqrt{n}}{\beta_3(p)}\,\Delta_n(p) < 0.356. \tag{8}$$

Indeed, $\beta_3(p)$ decreases with increasing p. Therefore, for $p < 0.02$

$$\Delta_n(p) < \frac{\beta_3(p)}{\sqrt{n}}\Big(0.33477 + 0.14362\beta_3^{-1}(0.02)\Big) < 0.3557\,\frac{\beta_3(p)}{\sqrt{n}}.$$

Combining bounds (6) – (8) we obtain

Theorem 2. *For every* $0 < p \leq 0.5$

$$\Delta_n(p) \leq 0.4215\,\frac{\beta_3(p)}{\sqrt{n}}. \qquad (9)$$

We see that the constant in the right–hand side of inequality (9) differs from C_E approximately by 0.0118. It gives grounds to expect that the least constant in the Berry–Esseen inequality for i.i.d. random variables, taking two values, equals, in reality, C_E.

References

[1] Berry A.C. The accuracy of the Gaussian approximation to the sum of independent variates // Trans. Amer. Math. Soc. 1941. Vol. 49, p. 122–126.

[2] Esseen C.-G. On the Liapounnoff limit error in the theory of probability // Ark. Mat. Astron. Fys. 1942. Vol. 28A. P. 1–19.

[3] Zolotarev V.M. An absolute estimate of the remainder term in the central limit theorem // Teor. Verojatnost. i Primenen. 1966. Vol. 11, No. 1. P. 108–119.

[4] Zahl S. Bounds for the central limit theorem error // SIAM J. Appl. Math. 1966. V. 14, No. 6. P. 1225–1245.

[5] Van Beek P. An application of Fourier methods to the problem of sharpening the Berry–Esseen inequality // Z. Wahrsch. verw. Geb. 1972. Bd. 23. P. 187–196.

[6] Shiganov I.S. On a refinement of the upper constant in the remainder term of the central limit theorem. In: Stability problems for stochastic models. Proceedings of the seminar. M.: VNIICI, 1982. P. 109–115 (in Russian).

[7] Prawitz H. On the Remainder in the Central Limit Theorem. Part I. Onedimentional Independent Variables with Absolute Moments of Third Order // Scand. Aktuarial J. 1975. No. 3. P. 145–156.

[8] Korolev V.Yu., Shevtsova I.G. On the upper estimate of the absolute constant in the Berry–Esseen inequality // Teor. Verojatnost. i Primenen. 2009. Vol. 53, No. 4. P. 671–695 (in Russian).

[9] Tyurin I. New estimates of the convergence rate in the Lyapunov theorem. ArXiv: 0912.0726v1, 2009.

[10] Tyurin I.S. A refinement of the upper bounds of the constants in the Lyapunov theorem // Uspekhi matem. nauk. 2010. Vol. 65, No. 3. P. 201-202 (in Russian).

[11] Korolev V.Yu., Shevtsova I.G. An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums. ArXiv: 0912.2795v2, 2009.

[12] Korolev V.Yu., Shevtsova I.G. A refinement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums // Obozrenie prikl. i prom. matem. 2010. Vol.17, No. 1. P. 25–56 (in Russian).

[13] Esseen C.-G. A moment inequality with an application to the central limit theorem // Scand. Aktuarietidskr. J. 1956. Vol. 39. P. 160–170.

[14] 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 2010/160 . Khabarovsk: Computing Centre FEB RAS, 2010.