On the approximation error of the tails of the binomial distribution by these of the Poisson law

S. V. Nagaev[1]***, V. I. Chebotarev***[2]

[1]Sobolev Institute of Mathematics, Novosibirsk, Russia; nagaev@math.nsc.ru
[2]Computing Center, Far Eastern Branch of the Russian Academy of Sciences, Khabarovsk, Russia; vladimir.ch@ccfebras.ru

1. Results. The subject of this study is upper and lower bounds for probabilities of the type $\mathbf{P}\big(\sum\limits_{i=1}^{n} X_i \geq nx\big)$, where $X_1, \ldots, X_n$ are independent identically distributed Bernoulli random variables. In other words, we estimate tail probabilities for the binomial distribution. To this end we use the Poisson approximation.

In what follows, we use the next notations: F is the distribution function of the Bernoulli random variable with parameter p, $0 < p \leq \frac{1}{2}$, $F_{n,p} = F^{*n}$ the n-fold convolution of F. We assume x to satisfy the following condition,

$$0 < p < x < 1. \tag{1}$$

Denote by $\Pi_\lambda(t)$ the distribution function of Poisson law with a parameter $\lambda > 0$, $\pi_\lambda(j) = \frac{\lambda^j}{j!}\, e^{-\lambda}$. If the variable x approaches 0, it is natural to take Π_λ with $\lambda = np$ as the approximating distribution for $F_{n,p}$. Just this distribution is used in Theorem 2. But first we need another approximating Poisson distribution with the mean $\lambda_1 = \lambda_1(n,p,x) = \frac{np(1-x)}{1-p}$, depending not only on the parameters n and p, but on the variable x from (1). We shall call this distribution by *the variable Poisson distribution.*

Let us formulate the first statement about the connection between the behaviors of tails $1 - F_{n,p}(nx)$ and $1 - \Pi_{\lambda_1}(nx)$. First introduce the function $A(x,n,p) = \left(\frac{1-x}{q}\right)^{-n} e^{-\frac{n(x-p)}{q}}$. Hereinafter $q = 1 - p$. We have

$$A(x,n,p) = \left(1 - \frac{x-p}{q}\right)^{-n} e^{-\frac{n(x-p)}{q}} = e^{-n\left[\ln\left(1-\frac{x-p}{q}\right) + \frac{x-p}{q}\right]} = e^{-n[\ln(1-u)+u]},$$

where $u = \frac{x-p}{q}$. Since $\big(-\ln(1-u) - u\big) = \sum_{k=2}^{\infty} \frac{u^k}{k} =: \Lambda_2(u)$, the following equality is true,

$$A(x,n,p) = e^{n\Lambda_2(u)}.$$

Note that the series $\Lambda_2(u)$ converges since by condition (1), we have $0 < \frac{x-p}{q} < 1$.

Let $0 < t < 1$. We will need the function $H(t,p) = t \ln \frac{t}{p} + (1-t) \ln \frac{1-t}{1-p}$, the so-called relative entropy or Kullback – Leibler distance between two two-point distributions $(t, 1-t)$ and $(p, 1-p)$ concentrated at the same pair of points.

Proposition. *If condition (1) is fulfilled, then*

$$1 - F_{n,p}(nx) = \big[1 - \Pi_{\lambda_1}(nx)\big]\, A(x,n,p) + R_1 = \big[1 - \Pi_{\lambda_1}(nx)\big]\, e^{n\Lambda_2(u)} + R_1,$$

where

$$|R_1| \le 2e^{-nH(x,p)} \max_{y \ge nx} \Big| F_{n,x}(y) - \Pi_{nx}(y) \Big| \le 2xe^{-nH(x,p)}. \tag{2}$$

Remark that the second inequality in (2) follows from Barbour and Hall [1]. Indeed, let $X_1, \ldots, X_n$ be independent Bernoulli random variables. We denote $S_n = \sum\limits_{j=1}^{n} X_j$, F_{S_n} the distribution of the sum S_n, $p_j = \mathbf{P}(X_j = 1)$, $\lambda = \sum\limits_{j=1}^{n} p_j$, Π_λ is the Poisson distribution with parameter λ. In Barbour and Hall [1], the following estimate for the total variation distance $d_{TV}(F_{S_n}, \Pi_\lambda)$ between F_{S_n} and Π_λ is obtained.

Theorem (Barbour and Hall [1, Theorem 1]). *The following inequality is valid,*

$$d_{TV}(F_{S_n}, \Pi_\lambda) \le (1 - e^{-\lambda})\, \frac{1}{\lambda} \sum_{j=1}^{n} p_j^2. \tag{3}$$

In the particular case when

$$p_1 = p_2 = \ldots = p_n = p, \tag{4}$$

we have $\lambda = np$. Then it follows from (3) that $d_{TV}(F_{S_n}, \Pi_\lambda) \le (1 - e^{-\lambda})\, p$, whence

$$d_{TV}(F_{S_n}, \Pi_\lambda) \le p. \tag{5}$$

In the case (4) we will use the notation

$$d_{n,p} = d_K(F_{S_n}, \Pi_\lambda),$$

where $d_K(F_{S_n}, \Pi_\lambda)$ is the Kolmogorov distance between the distributions F_{S_n} and Π_λ. Since $d_{n,p} \le d_{TV}(F_{S_n}, \Pi_\lambda)$, it follows from (5) that

$$d_{n,p} \le p. \tag{6}$$

Inequality (6) with $p = x$ entails the second inequality in (2).

The following theorem gives one more form of the dependence of the tails of the binomial distribution on the tails $1 - \Pi_{\lambda_1}(nx)$ of the variable Poisson distribution. It is a consequence of Proposition, but by no means trivial, and requires the proof of a number of additional statements.

Theorem 1. *If condition (1) is fulfilled, then*

$$\begin{aligned} 1 - F_{n,p}(nx) = \big[1 - \Pi_{\lambda_1}(nx)\big]\, A(x,n,p)\, (1 + r_1) = \\ = \big[1 - \Pi_{\lambda_1}(nx)\big]\, e^{n\Lambda_2(u)}\, (1 + r_1(x)), \end{aligned}$$

where $u = \frac{x-p}{q}$, $|r_1(x)| \le c_1\sqrt{nx^3}$, $c_1 = 2e^{1/12}\sqrt{2\pi} = 5.4489\ldots$.

Let us define $M(k;\lambda) := \frac{1-\Pi_\lambda(k)}{\pi_\lambda(k)}$, an analogue of the Mills ratio. Using this ratio, one can deduce the following statement from Theorem 1.

Theorem 2. *If condition (1) is fulfilled, then the following equality holds,*

$$\frac{1-F_{n,p}(nx)}{1-\Pi_{np}(nx)} = \frac{M(nx;\lambda_1)}{M(nx;np)}\, e^{-nq\Lambda_3\left(\frac{x-p}{q}\right)}(1+r_1(x)),$$

where $\Lambda_3(u) = \sum\limits_{k=2}^{\infty} \frac{u^k}{k(k-1)}$, $r_1(x)$ *is the function from Theorem 1.*

In turn, Theorem 2 implies the following corollary.

Corollary 1. *Let condition (1) be fulfilled and* $c_1\sqrt{nx^3} < 1$. *Then*

$$\frac{1-F_{n,p}(nx)}{1-\Pi_{np}(nx)} = e^{-\frac{n(x-p)^2}{2}}\left[1+\theta\Big(5.74\sqrt{nx^3}+\frac{p}{x}\Big)\right], \quad |\theta| < 1.$$

2. Some conjectures. Numerical experiments that we carried out led us to some conjectures. Our first conjecture is as follows: for every $2 \le k \le n$,

$$\max_{n\ge 1} d_{n,1/k} = d_{k,1/k}.$$

The next conjecture concerns existence and the value of the limit of $d_{k,1/k}$, when $k \to \infty$. Calculations lead to the assumption

$$d_{k,1/k} \equiv \max_{0\le j\le k} |F_{k,1/k}(j+)-\Pi_1(j+)| = |F_{k,1/k}(0+)-\Pi_1(0+)| \equiv e^{-1}-(1-1/k)^k.$$

It is easily seen that the sequence $kd(k,1/k)$ decreases and

$$kd(k,1/k) = k\Big(e^{-1}-e^{k\ln(1-1/k)}\Big) = e^{-1}k\Big(1-e^{-\frac{1}{2k}+O(k^{-2})}\Big) \underset{k\to\infty}{\longrightarrow} \frac{1}{2e}.$$

The constant c_0 in the inequality $\sup_{n,p}\frac{d_{n,p}}{p} \le c_0$ cannot be less $kd(k,1/k)\big|_{k=2} = 2(e^{-1}-\frac{1}{4}) = 0.2357\ldots$. Moreover, we can assume that the following equality holds,

$$c_0 = 2(e^{-1}-1/4). \tag{7}$$

Note that accordingly to Zacharovas and Hwang [2, P. 113] the inequality $d_K(F_n,\Pi_\lambda) < \frac{0.36}{\lambda}\sum_{j=1}^n p_j^2$ was obtained in Daley and Vere-Jones [3] (although we did not find this result in [3]). This means that in the case of independent identically distributed Bernoulli random variables the following bound holds, $c_0 \le 0.36$.

If the equality (7) was proved, then it would be an improvement of the inequality $c_0 \le 0.36$. Moreover, then one can write $d_{n,p} \le 2(e^{-1}-1/4)p$

instead of (6). And this, in turn, would lead to more accurate estimates in Proposition, Theorems $1, 2$ and Corollary 1 (since the constants in them would become smaller).

References

1. A. D. Barbour, P. Hall, On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95** (1984) 473-480.

2. V. Zacharovas, H-K. Hwang, A Charlier – Parseval approach to Poisson approximation and its applications. *Lithuanian mathematical journal* **50**:1 (2010) 88-119; arXiv:0810.4756v1, [math.PR]. (2008) 27 Oct.

3. D. J. Daley, D. Vere-Jones, *An introduction to the theory of point processes. Vol. II. General Theory and Structure. Second edition*, Springer, New York. 2008.