AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 129–135

Exact Computation of Pearson Statistics Distribution and Some Experimental Results

Marina V. Filina and Andrew M. Zubkov
Steklov Mathematical Institute of RAS, Moscow, Russia

Abstract: A Markov chain based algorithms for exact and approximate computation of Pearson statistics distribution for multinomial scheme are described. Results of computational experiments reveal some new properties of the difference between this distribution and corresponding chi-square distribution.

Keywords: Approximation of Probability Distributions.

1 Introduction

The Pearson statistics for multinomial scheme and its modifications is used by different goodness-of-fit criteria. As a rule the selection of critical values for Pearson statistics is based on the convergence of its distribution to the χ^2 distribution with appropriate degrees of freedom as sample size tends to infinity. In practice sample sizes are bounded, and the question on the accuracy of such approximation (especially for distribution tails) arises naturally. Results of investigation of this problem was reported, in particular, in Holzman and Good (1986) where more than 250 examples of equiprobable multinomial scheme with $N \in [2, 160]$ outcomes and sample sizes $T \in [10, 80]$ were considered. To compute the distribution function of Pearson statistics Holzman and Good (1986) have used generating functions and Good, Gover, and Mitchell (1970) – Fast Fourier Transform. Computational method for decomposable statistics distribution (also based on generating functions) was proposed in Selivanov (2006). We propose to compute the Pearson statistics distribution by means of the Markov chain method suggested in Zubkov (1996, 2002); this method may be applied to distributions of decomposable statistics for multinomial and some other schemes also.

2 Method

Let $\nu_1, \ldots, \nu_N$ be frequencies of N outcomes with probabilities $p_1, \ldots, p_N$ in a multinomial sample of size T. Random variables of the form $\zeta = \sum_{j=1}^{N} f_j(\nu_j)$, where $f_1(x), \ldots, f_N(x)$ are given functions, are called *decomposable statistics*. In the case $f_j(x) = (x - Tp_j)^2/Tp_j$, $j = 1, \ldots, N$, we obtain the Pearson statistics

$$X^2_{N,T} = \sum_{j=1}^{N} \frac{(\nu_j - Tp_j)^2}{Tp_j}. \tag{1}$$

If the hypothetical probabilities $p_j = m_j/n_j$, $j = 1, \dots, N$, are rational then the formula for the Pearson statistics may be rewritten as

$$X^2_{N,T} = \frac{1}{T\hat{M}\hat{N}} \sum_{j=1}^{N} \frac{\hat{M}}{m_j} \frac{\hat{N}}{n_j} (n_j\nu_j - m_j T)^2, \tag{2}$$

where $\hat{M} = \mathrm{LCM}(m_1, \dots, m_N)$, $\hat{N} = \mathrm{LCM}(n_1, \dots, n_N)$.

If all hypothetical probabilities are equal ($p_1 = \cdots = p_N = 1/N$) then the formula for the Pearson statistics may be represented in another form as

$$X^2_{N,T} = \sum_{j=1}^{N} \frac{(\nu_j - T/N)^2}{T/N} = \frac{N}{T} \left(\sum_{j=1}^{N} \left(\nu_j - \left\langle \frac{T}{N} \right\rangle \right)^2 - N \left(\left\langle \frac{T}{N} \right\rangle - \frac{T}{N} \right)^2 \right), \tag{3}$$

where $\langle x \rangle = [x + 1/2]$ denotes the nearest integer to x. Formulas (2) and (3) reduce the computation of the Pearson statistics distribution to the one of integer-valued decomposable statistics. Exact distributions of integer-valued random variables may be stored as tables in a computer memory.

Further, the conditional distribution of the frequency ν_t on the set $\left\{ \sum_{j=1}^{t-1} \nu_j = u \right\}$ coincides with the binomial distribution $\mathrm{Bin}(T - u, p_t/P_t)$, $P_t \stackrel{\text{def}}{=} p_t + \cdots + p_N$; so the sequence $\kappa_0 = 0$, $\kappa_t \stackrel{\text{def}}{=} \sum_{j=1}^{t} \nu_j$, $t = 1, \dots, N$, may be considered as a time-inhomogeneous Markov chain with state space $\{0, 1, \dots, T\}$ and transition probabilities

$$p_t(v|u) = \mathbf{P}\{\kappa_t = v | \kappa_{t-1} = u\} = C^{v-u}_{T-u} \left(\frac{p_t}{P_t} \right)^{v-u} \left(1 - \frac{p_t}{P_t} \right)^{T-v}, \qquad 0 \le u \le v \le T. \tag{4}$$

So the sequences

$$\zeta^*_0 = (0, 0), \qquad \zeta^*_t = \left(\sum_{j=1}^{t} \nu_j, \sum_{j=1}^{t} \left(\nu_j - \left\langle \frac{T}{N} \right\rangle \right)^2 \right), \qquad t = 1, \dots, N, \tag{5}$$

$$\zeta_0 = (0, 0), \qquad \zeta_t = \left(\sum_{j=1}^{t} \nu_j, \sum_{j=1}^{t} \frac{\hat{M}}{m_j} \frac{\hat{N}}{n_j} (n_j\nu_j - m_j T)^2 \right), \qquad t = 1, \dots, N, \tag{6}$$

(being additive functions of $\{\kappa_t\}$) are finite nonhomogeneous Markov chains $\zeta_t \stackrel{\text{def}}{=} (\zeta_{t,1}, \zeta_{t,2})$ and $\zeta^*_t \stackrel{\text{def}}{=} (\zeta^*_{t,1}, \zeta^*_{t,2})$ with transition probabilities

$$\begin{aligned} &\mathbf{P}\left\{ \zeta^*_t = \left(v, s + \left(v - u - \left\langle \frac{T}{N} \right\rangle \right)^2 \right) \middle| \, \zeta^*_{t-1} = (u, s) \right\} \\ &= \mathbf{P}\left\{ \zeta_t = \left(v, s + \frac{\hat{M}}{m_t} \frac{\hat{N}}{n_t} (n_t(v - u) - m_t T)^2 \right) \middle| \, \zeta_{t-1} = (u, s) \right\} = p_t(v|u) \end{aligned} \tag{7}$$

for $0 \le u \le v \le T$ and 0 in other cases. It is obvious that $\zeta^*_N = (T, \zeta^*_{N,2})$ a.s., and the distribution of $(N/T) \left(\zeta^*_{N,2} - N \left(\langle (T/N) \rangle - (T/N) \right)^2 \right)$ coincides with that of $X^2_{N,T}$ from

(3); analogously, $\zeta_N = (T, \zeta_{N,2})$ a.s., and the distribution of $1/(T\hat{M}\hat{N})\zeta_{N,2}$ coincides with that of $X^2_{N,T}$ from (2). It follows that we may find the distribution of $X^2_{N,T}$ by means of a recursive computation of the distributions of ζ^*_k (or ζ_k), $k = 1, \ldots, N$.

Note that probabilities $p_1, \ldots, p_N$ in the definition (4) of the chain κ_t transition probabilities need not be equal to the hypothetical probabilities in the formula for Pearson statistics. In other words, this approach may be applied to the computation of Pearson statistics distribution under alternative hypotheses also.

In the case of arbitrary probabilities $p_1, \ldots, p_N$ we may use the same ideas to compute *approximate* distribution function of $X^2_{N,T}$ by means of discretization. To this end it suffice to introduce functions $h_\varepsilon(x) = [1/2 + x/\varepsilon]$ (where $\varepsilon > 0$, $[y]$ denotes integer part of y) and consider the discrete Markov chain

$$\zeta^\varepsilon_0 = (0, 0), \quad \zeta^\varepsilon_t = \left(\sum_{j=1}^{t} \nu_j, \sum_{j=1}^{t} h_\varepsilon \frac{(\nu_j - Tp_j)^2}{Tp_j} \right), \quad t = 1, \ldots, N,$$

with transition probabilities

$$\mathbf{P}\left\{ \zeta^\varepsilon_t = \left(v, s + h_\varepsilon \frac{(v - u - Tp_t)^2}{Tp_t} \right) \middle| \, \zeta^\varepsilon_{t-1} = (u, s) \right\} = p(v|u),$$

where $p(v|u)$ may be defined by probabilities not necessarily coinciding with $p_1, \ldots, p_N$. Let, as earlier, $\zeta^\varepsilon_N = (T, \zeta^\varepsilon_{N,2})$. From the obvious estimate $|\varepsilon h_\varepsilon(x) - x| \le \varepsilon/2$ we have crude bounds

$$\mathbf{P}\left\{\varepsilon\zeta^\varepsilon_N \le x - N\varepsilon/2\right\} \le \mathbf{P}\left\{X^2_{N,T} \le x\right\} \le \mathbf{P}\left\{\varepsilon\zeta^\varepsilon_N \le x + N\varepsilon/2\right\}.$$

Reducing the value of ε we may find arbitrary good estimates for $\mathbf{P}\left\{X^2_{N,T} \le x\right\}$. The volume of memory used is inversely proportional to ε.

This method was realized by several C++ programs. In particular, for equiprobable schemes the Pearson statistics distributions with the number of outcomes up to hundreds and with the number of trials up to thousands was computed. Computations on PC takes from seconds if number of outcomes and trials are less that 50 to minutes if these numbers are of the order of several hundreds. Time and memory requirements of a program realizing algorithm for rational probabilities depend heavily on their arithmetical structure. Time and memory requirements of a program for approximate computation are analogous to that of a program for equiprobable case.

3 Experimental Results

Computational experiments reveal some interesting features of the difference between exact Pearson statistics distributions and corresponding chi-square distributions.

The equiprobable case with $N = 10$, $T = 10$ is used in Figure 1 to explain the structure of all subsequent pictures. There are a piecewise-constant distribution function of exact Pearson statistics, a continuous distribution function of χ^2 distribution with 9 degrees of freedom, a discontinuous saw-like difference between two preceding functions and piecewise linear continuous function connecting mean values of the difference at

discontinuity points ("average difference") in the left part of Figure 1. In the following we consider plots of differences and average differences only (as in the right part of this figure).

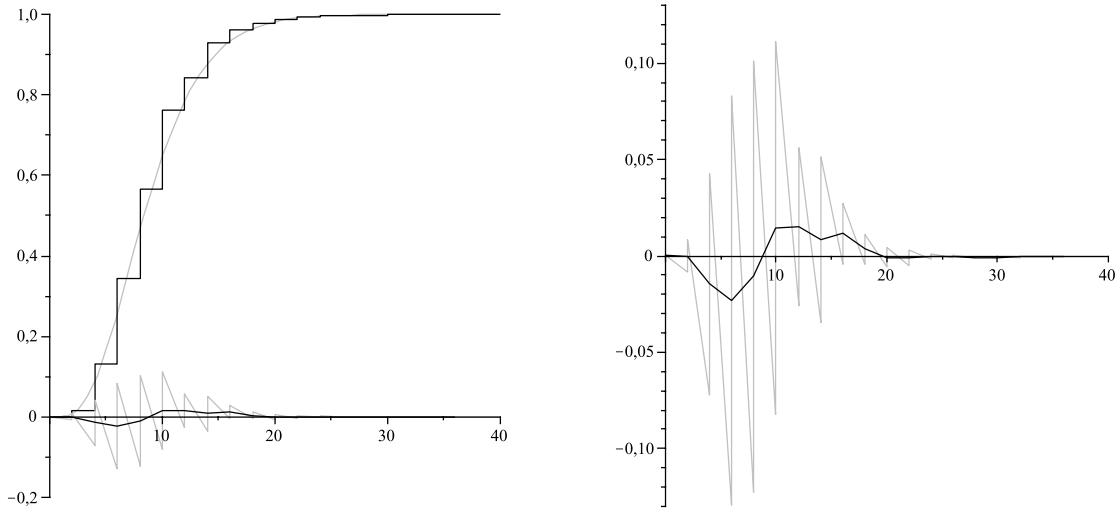

Figure 1: $N = 10$, $T = 10$, $p = p_1 = \cdots = p_{10} = 0.1$

In Figure 2 we plot differences and average differences for equiprobable case with $N = 10$ outcomes and $T = 100$, $T = 1000$ trials. Note that the shapes of plots are almost independent on T. The ranges of graphs are approximately inversely proportional to T. The shapes of average differences have a form of fading wave; the sign of average difference becomes negative on the right tail of distribution (after $x \approx 25$), but eventually it becomes positive again (due to the boundedness of the Pearson statistics distribution).

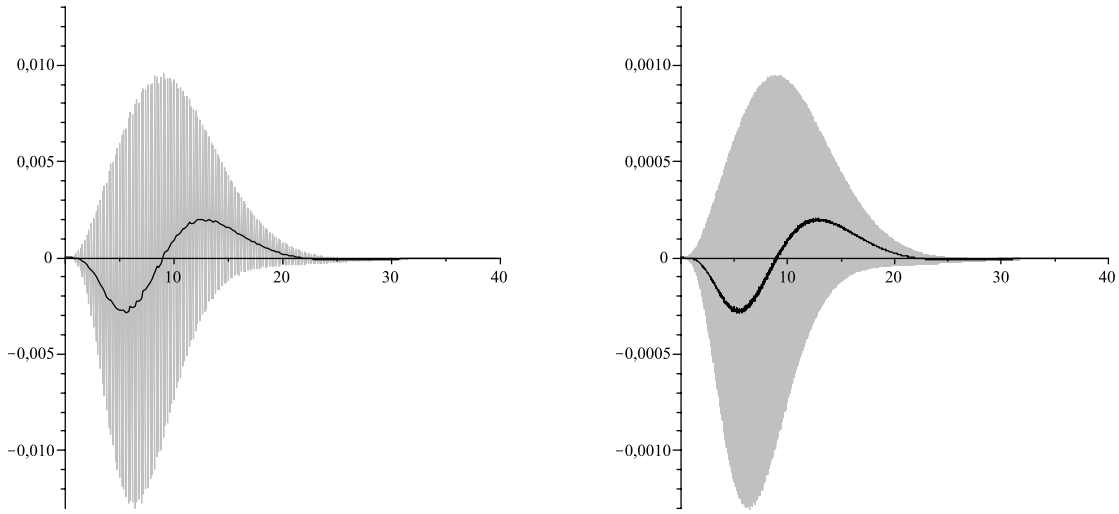

Figure 2: Equiprobable cases, $N = 10$, $T = 100$ (left) and $T = 1000$ (right)

Plots of the differences for equiprobable cases with constant values of ratios $T/N = 10$ and $T/N = 2$ are shown in Figures 3 and 4, respectively. We can see that the shapes of

the differences slightly depend on N, that the shape of the average differences are more stable and that the ranges of the graphs are approximately inversely proportional to the square root of the number of outcomes. So large number of outcomes may compensate an insufficient number of trials even when the quotient T/N is as small as 2 (Figure 4).

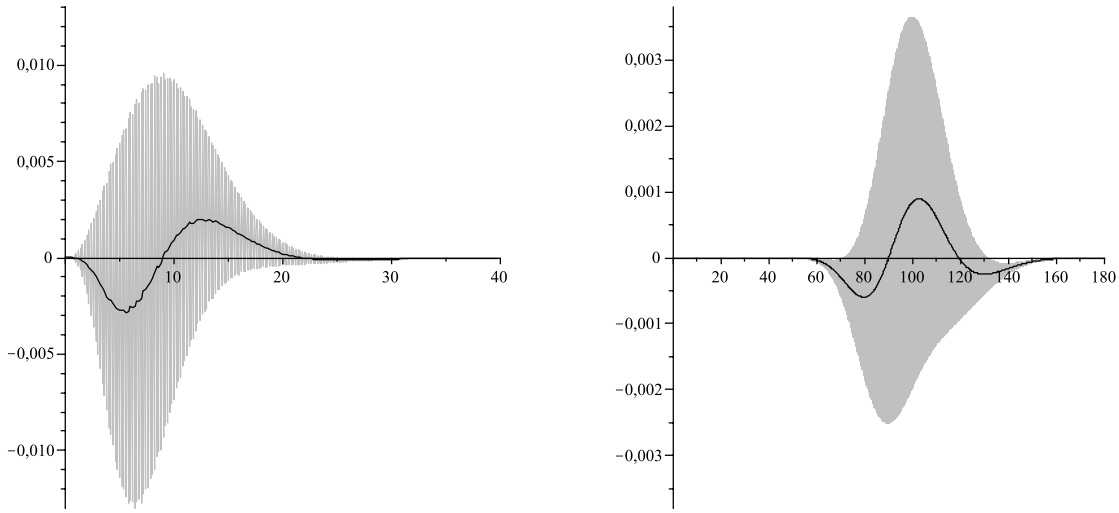

Figure 3: Equiprobable cases, $T = 10N$, $N = 10$ (left), $N = 100$ (right)

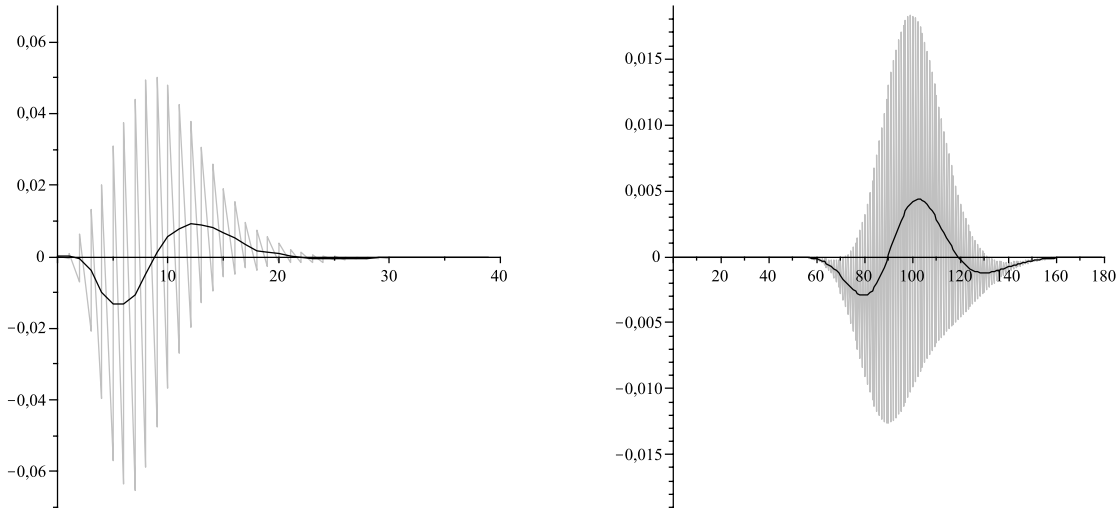

Figure 4: Equiprobable cases, $T = 2N$, $N = 10$ (left), $N = 100$ (right)

As long as the distribution on the set of outcomes becomes more non-uniform the shape of plots of differences approaches the shape of plots of average differences, namely, the shape of fading wave with two minima and one maximum (see Figures 5, 6). Values of these extremes are comparable with the extremum values of average differences for corresponding equiprobable cases.

The reason of shrinking the differences to the average differences when the distribution of the outcomes goes away from a uniform one may be explained (on the heuristic level) as follows.

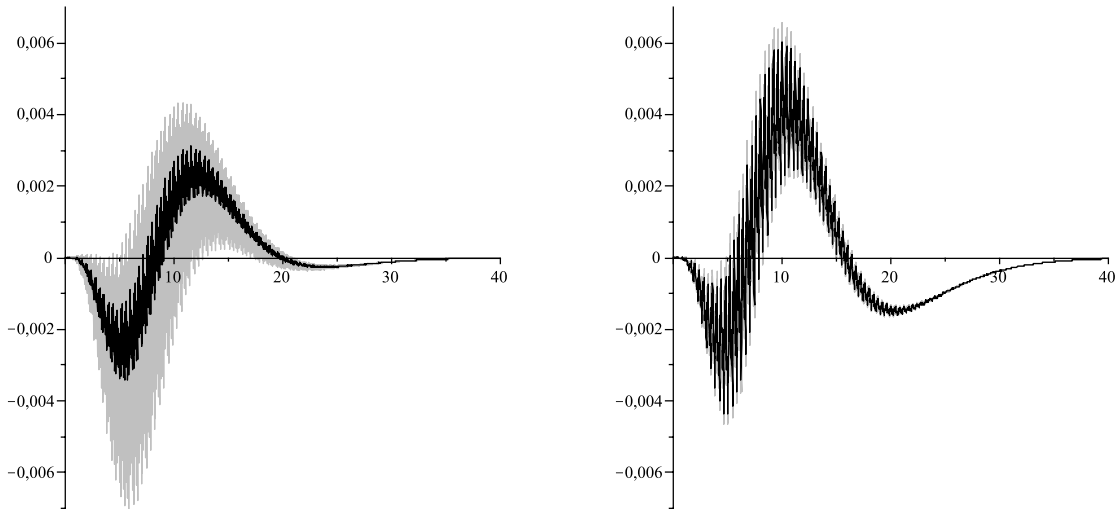

Figure 5: $N = 10$, $T = 100$, $p = (0.05, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.15)$ (left), $p = (0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.1, 0.15, 0.2, 0.25)$ (right)

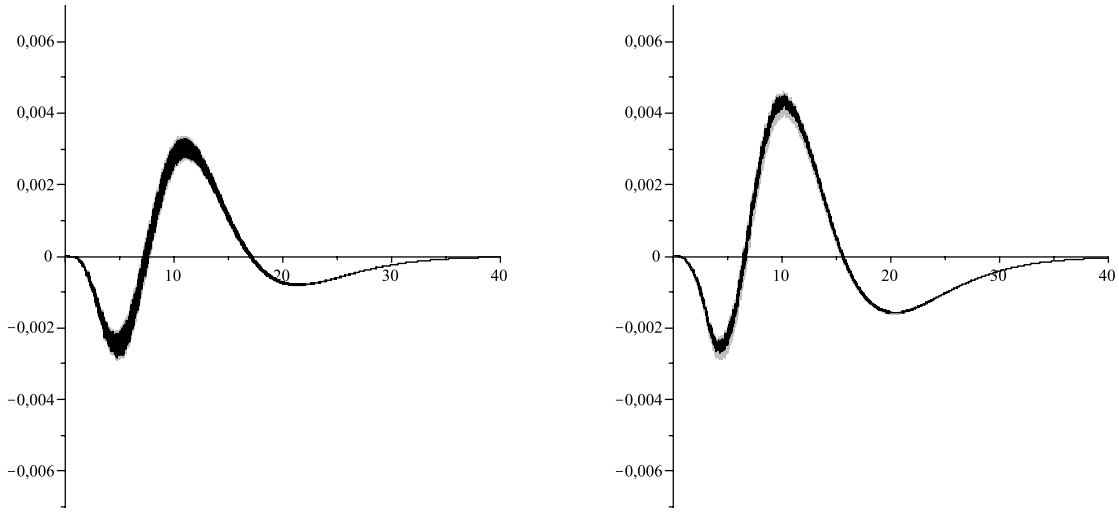

Figure 6: $N = 10$, $T = 100$, $p = (0.04, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.14, 0.14, 0.14)$ (left), $p = (0.02, 0.04, 0.06, 0.08, 0.1, 0.1, 0.12, 0.14, 0.16, 0.18)$ (right)

According to (1) the support of the Pearson statistics distribution is contained in the set of nonnegative numbers having representation $\sum_{j=1}^{N} k_j^2/Tp_j - T$, where $k_1, \ldots, k_N$ are nonnegative integer numbers. If the probabilities $p_1 = c_1/d_1, \ldots, p_N = c_N/d_N > 0$ are rational numbers and $M = \text{LCM}(c_1, \ldots, c_N)$ then this set is contained in the arithmetical progression $\{k/TM, k = 0, 1, \ldots\}$. The distribution function $F_{N,T}(x) = \mathbf{P}\{X_{N,T}^2 < x\}$ is approximated by the distribution function of a χ^2 distribution with $N - 1$ degrees of freedom, so the smaller the value TM the greater the weights of atoms of distribution $X_{N,T}^2$, i.e. jumps of piecewise-constant distribution function $F_{N,T}(x)$ (in particular, the largest atoms appear in the equiprobable case). Therefore, the accuracy of approximation $F_{N,T}(x)$ by distribution function of χ^2 statistics with $N - 1$ degrees of freedom cannot be

good for small values of TM.

We hope to find quantitative theoretical explanation of effects described.

Acknowledgements

This research was partly supported by the Leading Scientific Schools Support Fund of the President of Russia Federation, grant 4129.2006.1, and by the Program of RAS "Modern Problems of Theoretical Mathematics".

References

Good, I. J., Gover, T. N., and Mitchell, G. J. (1970). Exact distributions for χ^2 and for likelihood-ratio statistic for the equiprobable multinomial distribution. *Journal of the American Statistical Association*, *65*, 267-283.

Holzman, G. I., and Good, I. J. (1986). The Poisson and chi-squared approximation as compared with the true upper-tail probability of Pearson's χ^2 for equiprobable multinomials. *Journal of Statistical Planning and Inference*, *13*, 283-295.

Selivanov, B. I. (2006). On the exact computation of decomposable statistics distributions for polynomial scheme (in russian). *Diskretnaya matematika*, *18*, 85-94.

Zubkov, A. M. (1996). Recurrent formulae for distributions of functions of discrete random variables (in russian). *Obozr. prikl. prom. matem.*, *3*, 567-573.

Zubkov, A. M. (2002). Computational methods for distributions of sums of random variables (in russian). In *Trudy po diskretnoi matematike* (Vol. 5, p. 51-60). Moscow: Fismatlit.

Corresponding Authors' Addresses:

Andrew M. Zubkov
Department of Discrete Mathematics
Steklov Mathematical Institute
Gubkina Str. 8
119991 Moscow
Russia

E-mail: MFilina@mi.ras.ru and zubkov@mi.ras.ru