AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 47–54

Tail Properties of Pearson Statistics Distributions

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

Abstract: By means of exact computation of Pearson statistics distributions we illustrate some differences between their tails and tails of corresponding chi-square distributions.

Keywords: Pearson Statistics, Tails of Distributions, Exact Calculation of Distributions.

1 Introduction

Let $\nu_1, \ldots, \nu_N$ be frequencies of all N outcomes of a multinomial scheme in a sample of size T. A usual goodness-of-fit test for the hypothesis that probabilities of outcomes are equal to $p_1, \ldots, p_N$ is based on the Pearson statistics

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

It is well-known that if this hypothesis is valid then the distribution of $X^2_{N,T}$ converges to the chi-square distribution with $N-1$ degrees of freedom as $T \to \infty$.

But in practice the values of T are bounded, and so the question on the accuracy of such approximation (especially for tails) arises naturally. Results of investigation of this problem for equiprobable multinomial schemes with $N \in [2, 160]$ outcomes and sample sizes $T \in [10, 80]$ was reported in Good, Gover, and Mitchell (1970), Holzman and Good (1986). Experimental results on the accuracy of approximation of Pearson statistics distribution function by chi-square distribution function in a central zone were presented in Filina and Zubkov (2008).

Here we investigate the differences between $\Pr\{X^2_{N,T} \le x\}$ and the distribution function $F_{N-1}(x)$ of chi-square distribution with $N-1$ degrees of freedom for large x. Results of our computations show that for finite sample sizes the tails of Pearson statistics distribution may be significantly heavier than the chi-square tails.

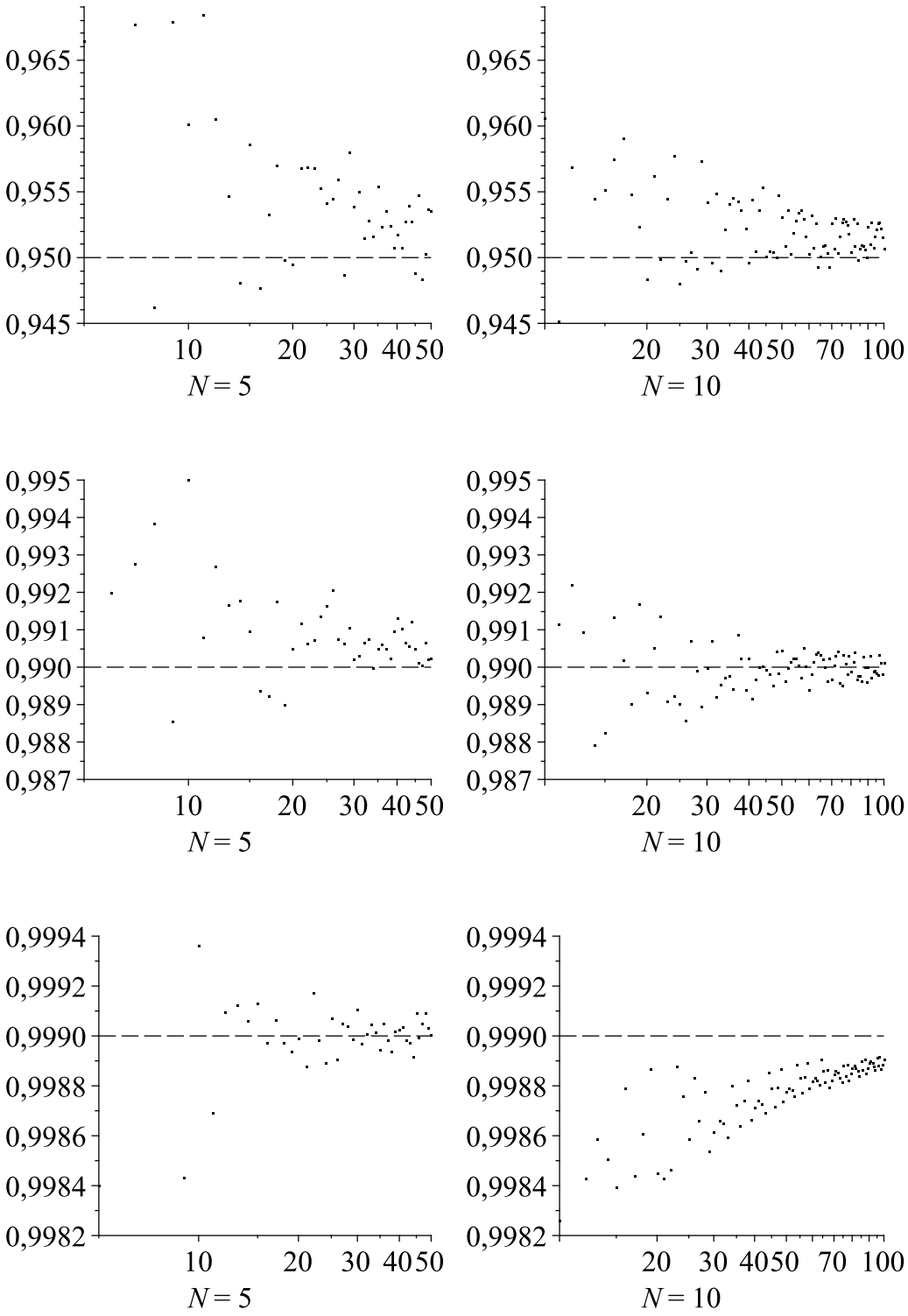

Figure 1: Equiprobable cases

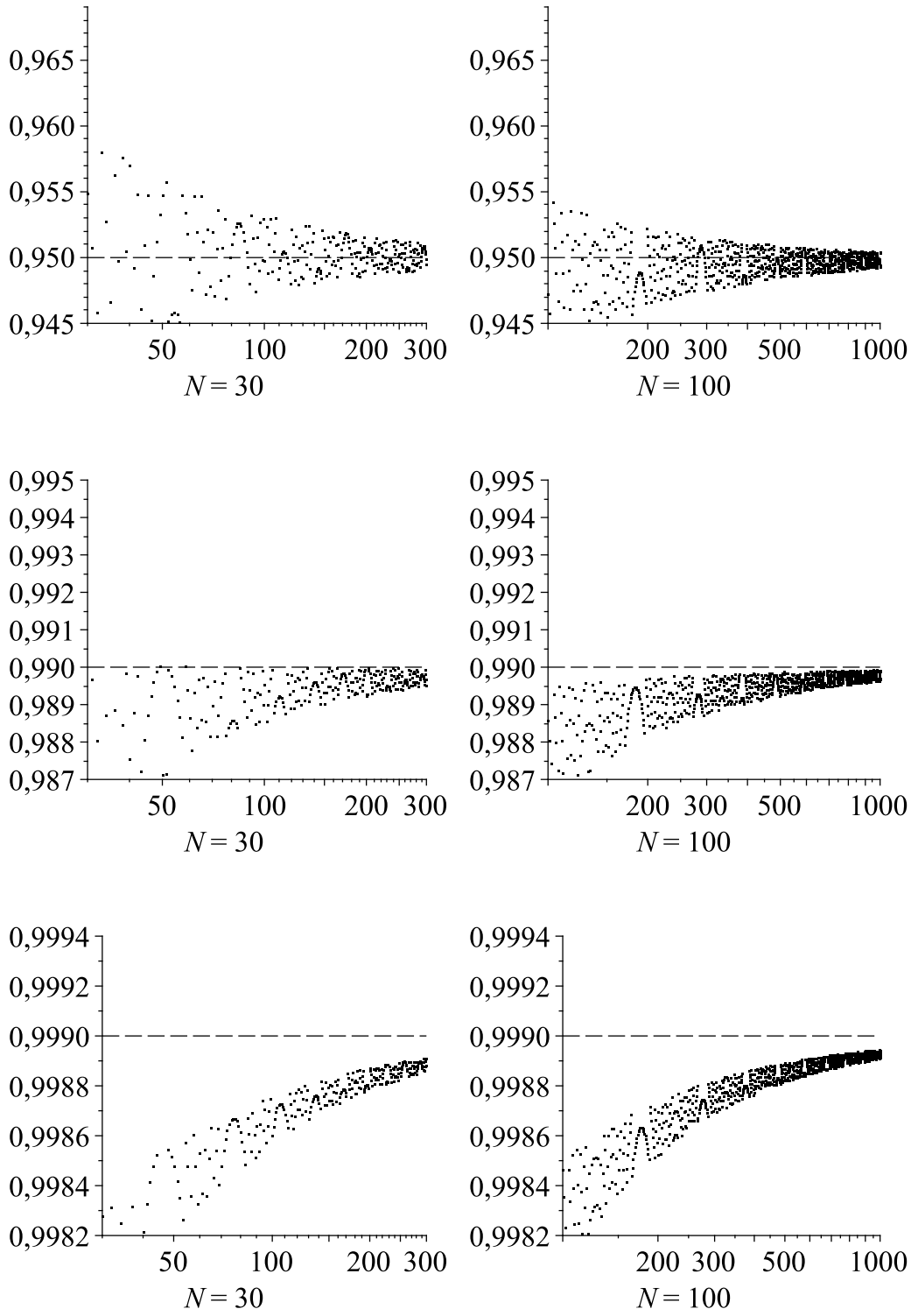

Figure 2: Equiprobable cases

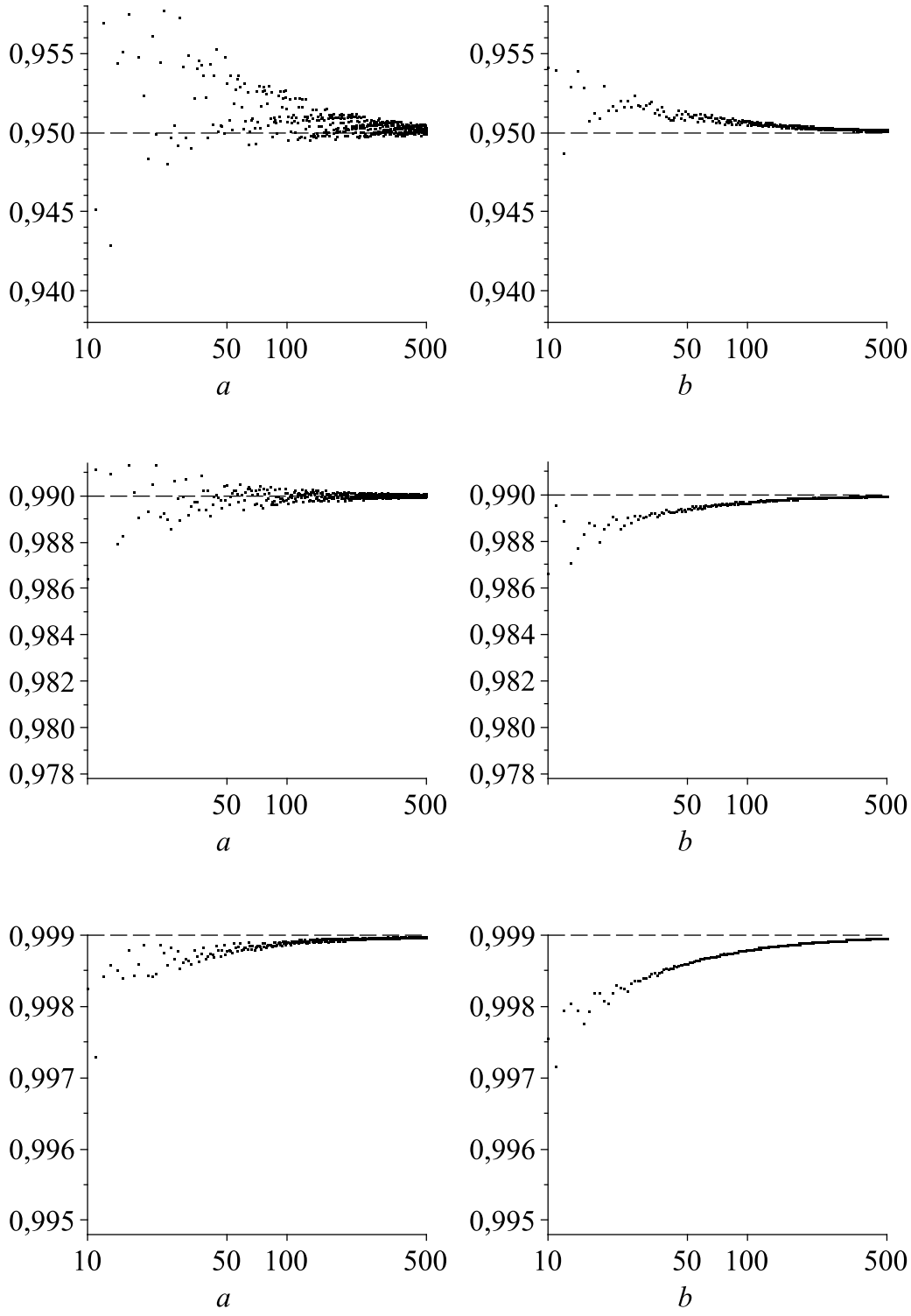

Figure 3: Equiprobable and non-equiprobable cases: $N = 10$:
a) $\mathbf{p} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$,
b) $\mathbf{p} = (0.13, 0.13, 0.13, 0.13, 0.13, 0.07, 0.07, 0.07, 0.07, 0.07)$.

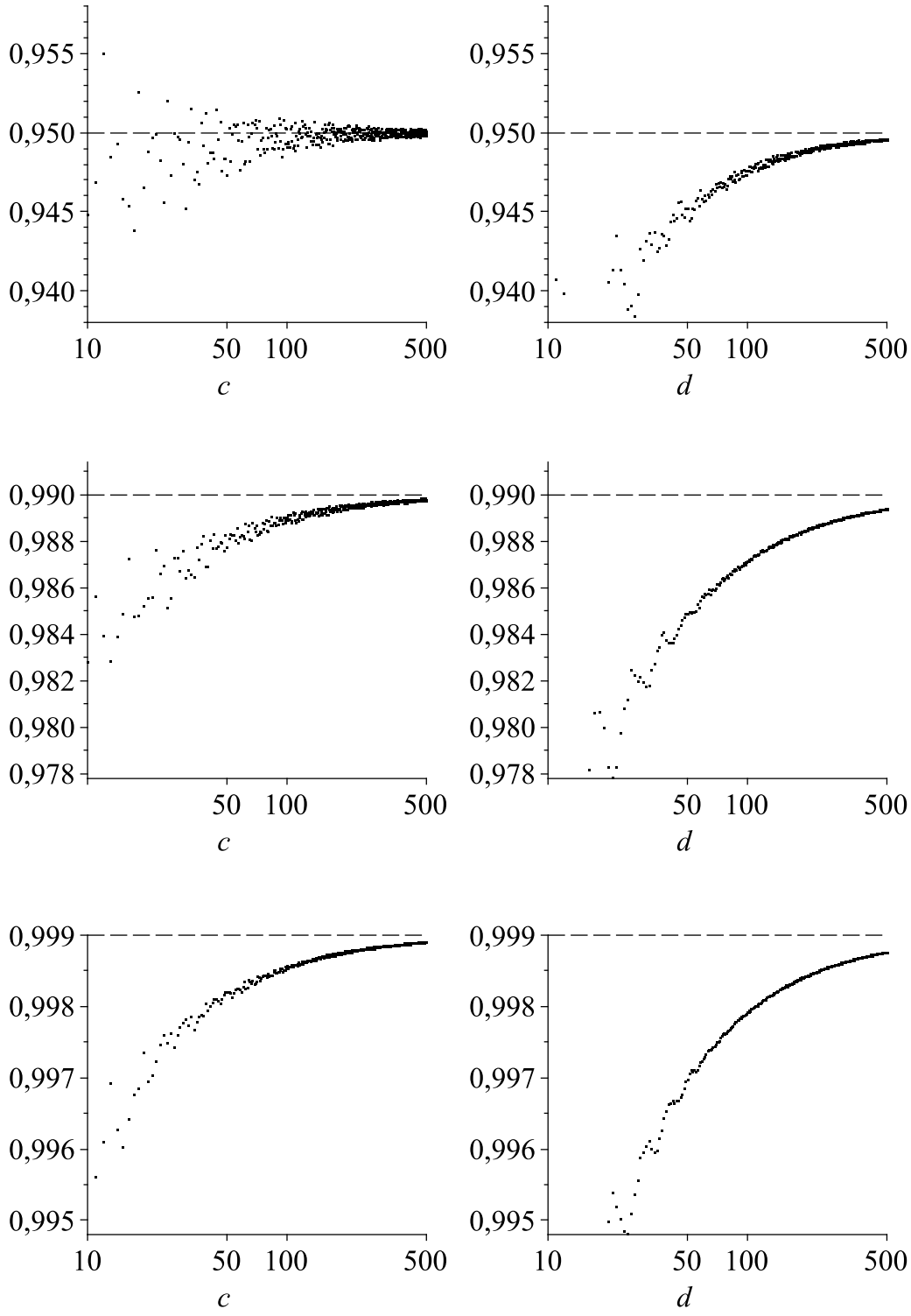

Figure 4: Non-equiprobable cases, $N = 10$:
c) $\mathbf{p} = (0.15, 0.15, 0.15, 0.15, 0.15, 0.05, 0.05, 0.05, 0.05, 0.05)$,
d) $\mathbf{p} = (0.17, 0.17, 0.17, 0.17, 0.17, 0.03, 0.03, 0.03, 0.03, 0.03)$.

2 Numerical Results

Typical results of our computations are shown in Figures 1 to 4. The horizontal axes in these graphs are used for the values of the sample size T (in a logarithmic scale), the vertical axes for the values of the distribution functions. Three graphs in each column correspond to the values of x such that $F_{N-1}(x)$ equals to 0.95, 0.99, and 0.999 (these numbers are marked by dotted lines). The values of N are shown under the graphs in Figures 1 and 2 (for equiprobable outcomes) and in the legends to Figures 3 and 4 (for non-equiprobable outcomes).

The points on all figures have coordinates $(T, \Pr\{X^2_{N,T} \le x\})$, so the distance between the point and the dotted line equals to the error of approximation of exact value $\Pr\{X^2_{N,T} \le x\}$ by its asymptotic value $F_{N-1}(x)$. The point is under the dotted line, iff the tail $\Pr\{X^2_{N,T} > x\}$ of the Pearson statistics distribution is heavier than the tail $1 - F_{N-1}(x)$ of the chi-square distribution. In all examples the values of $X^2_{N,T}$ are rational, so there are no atoms at 0.95, 0.99, and 0.999 quantiles of $F_{N-1}(x)$.

Figures 1 to 4 illustrate the following phenomena:

- For fixed x the absolute values of difference $\Delta_{N,T}(x) = \Pr\{X^2_{N,T} \le x\} - F_{N-1}(x)$ decrease to 0 as $T \to \infty$ approximately as $C(x)T^{-1}$.
- The Relative error of approximation $|\Delta_{N,T}(x)/(1-F_{N-1}(x))|$ increases as $F_{N-1}(x)$ grows from 0.95 to 0.999 (and further, at least up to 0.9999999; the computation errors are smaller than 10^{-8}). If the outcomes are equiprobable than the bound of this relative error approximately depends on T/N; the relative error becomes greater as the differences between the probabilities of the outcomes grow.

The width of the strip containing points in our graphs has the same order as the probabilities of the atoms of the $X^2_{N,T}$ distribution near the corresponding value of x. So, these widths are maximal for equiprobable outcomes and are decreasing when probabilities of outcomes become more different.

"Parabolic structures" in Figures 1 and 2 possibly are the results of the interplay of two processes. In case of equiprobable outcomes $\Pr\{X^2_{N,T} \le x\}$ is the sum of probabilities of the form $\Pr\{(\nu_1 = n_1, \ldots, \nu_N = n_N\}$ over the integer points $(n_1, \ldots, n_N)$ lying in the intersection of the hyperplane $n_1 + \cdots + n_N = T$ and the ball $n_1^2 + \cdots + n_N^2 \le (x+T)T/N$. With the increasing of T the probabilities of points and number of integer points in the intersection are changing in a regular interrelated ways.

2.1 Methods

In calculations of the exact distributions of $X^2_{N,T}$ we use algorithms described in Zubkov (1996, 2002), Filina and Zubkov (2008). These algorithms use the representation of polynomial distribution $\mathrm{Poly}(T; p_1, \ldots, p_N)$ as a distribution of some time-inhomogeneous Markov chain after N steps. The state space of this chain depends on T and on the probabilities $p_1, \ldots, p_N$. For equiprobable cases such chains have minimal state space $O(T)$ and complexity of algorithms are minimal also (of order $O(NT^2)$).

To compute the values of $\Pr\{X^2_{N,T} \le x\}$ for non-equiprobable cases shown in Figures 3 and 4 we use an additional idea. It will be described in a simplified form to avoid too cumbersome formulas.

Let $\nu_{n,j}(t)$ be the frequency of j-th outcome in a sample of size t from the equiprobable polynomial scheme with n outcomes and

$$f_{n,t}(k) = \Pr\{\nu_{n,1}^2(t) + \cdots + \nu_{n,n}^2(t) = k\}, \quad F_{n,t}(k) = \Pr\{\nu_{n,1}^2(t) + \cdots + \nu_{n,n}^2(t) \le k\}.$$

Then considering $\{\nu_{n,j}(k)\}_{j=1}^{n}$ and $\{\nu_{N-n,j}(T-k)\}_{j=1}^{N-n}$ as independent sets of frequencies (also in the case $n = N-n$, $k = T-k$) and supposing that $\mathbf{p} = (p_1, \ldots, p_N) = (q_1, \ldots, q_1, q_2, \ldots, q_2)$ such that $nq_1 + (N-n)q_2 = 1$ we have

$$\Pr\{X_{N,T}^2 \le x\} = \Pr\left\{\sum_{i=1}^{N} \frac{(\nu_i - Tp_i)^2}{Tp_i} \le x\right\} = \Pr\left\{\sum_{i=1}^{N} \frac{\nu_i^2}{Tp_i} \le x + T\right\} \tag{1}$$

$$= \sum_{m=0}^{T} \Pr\left\{\sum_{i=1}^{n} \nu_i = m\right\} \Pr\left\{\frac{1}{Tq_1}\sum_{i=1}^{n} \nu_{n,i}^2(m) + \frac{1}{Tq_2}\sum_{j=1}^{N-n} \nu_{N-n,j}^2(T-m) \le x + T\right\}$$

$$= \sum_{m=0}^{T} C_T^m (nq_1)^m ((N-n)q_2)^{T-m} \sum_{k\ge 0} f_{n,m}(k) F_{N-n,T-m}\left(Tq_2\left(x + T - \frac{k}{Tq_1}\right)\right).$$

So to compute $\Pr\{X_{N,T}^2 \le x\}$ for all $T \le T_{\max}$ and any q_1, q_2 it is sufficient to compute tables of values of $f_{n,m}(k)$ for all $m \le T_{\max}$, $k \le Tq_1(x+T)$ and of $F_{N-n,m}(k)$ for all $m \le T_{\max}$, $k \le Tq_2(x+T)$ and then use the formula (1).

An analogous approach may be used:

- for cases when the number of different values of p_i is somewhat larger than 2,
- for cases of arbitrary distributions when N is not very large (because the number C_{k+n-1}^k of all possible values of $\{\nu_{n,j}(k)\}_{j=1}^{n}$ is not very large for small values of n and k).

2.2 Concluding Remarks

1. To choose the right critical level for tests based on the Pearson statistics and limited sample size a detailed analysis of the exact distribution of this statistics is desirable.
2. It will be interesting to find theoretical justifications of the phenomena discovered.

References

Filina, M. V., and Zubkov, A. M. (2008). Exact computation of Pearson statistics distribution and some experimental results. *Austrian Journal of Statistics*, *37*, 129-135.

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.

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.

Authors address:

Marina V. Filina and Andrew M. Zubkov
Department of Discrete Mathematics
Steklov Mathematical Institute
Gubkina Str. 8
119991 Moscow, Russia

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