SOME REMARKS ON THE NONCENTRAL PEARSON STATISTICS DISTRIBUTIONS

M. V. Filina[1], A. M. Zubkov[2]
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: [1]`mfilina@mi.ras.ru`, [2]`zubkov@mi.ras.ru`

Abstract

By means of numerical algorithms we investigate the exact distributions of the Pearson statistics under alternatives and possibility to use the noncentral chi-square or normal distributions as approximations.

1 Introduction

Let $\nu_1, \ldots, \nu_N$ be frequencies of N outcomes of a multinomial scheme in a sample of size T. A most popular goodness-of-fit test for the hypothesis H_p: "probabilities of outcomes are positive and equal to $p_1, \ldots, p_N$" is based on the Pearson statistics

$$X^2_{N,T} = \sum_{i=1}^{N} \frac{(\nu_i - Tp_i)^2}{Tp_i} \,. \tag{1}$$

If the hypothesis H_p is valid, then the distribution of $X^2_{N,T}$ converges (as $T \to \infty$) to the chi-square distribution with $N-1$ degrees of freedom having mean $N-1$ and variance $2(N-1)$. It is well-known that if the hypothesis is not valid, then in the triangular scheme with $T \to \infty$ and true probabilities of outcomes having the form $\pi_1 = p_1 + \frac{a_1}{\sqrt{T}}, \ldots, \pi_N = p_n + \frac{a_N}{\sqrt{T}}$ ($a_1, \ldots, a_N$ are fixed and $a_1 + \ldots + a_N = 0$) the distribution of the Pearson statistics $X^2_{N,T}$ converges to the noncentral chi-square distribution with noncentrality parameter $\lambda = \sum_{i=1}^{N} \frac{a_k^2}{p_k} = \mathbf{E}X^2_{N,T} - (N-1)$. If $T \to \infty$ and true probabilities of outcomes $\pi_1, \ldots, \pi_N$ are fixed, $\sum_{i=1}^{N}(\pi_i - p_i)^2 > 0$, then the Pearson statis1tics $X^2_{N,T}$ is asymptotically normal (see [1]) with mean

$$\mathbf{E}X^2_{N,T} = N - 1 + (T-1)\sum_{i=1}^{N} \frac{(\pi_i - p_i)^2}{p_i} + \sum_{i=1}^{N} \frac{\pi_i - p_i}{p_i}$$

and variance ([2, 3])

$$\mathbf{D}X^2_{N,T} = \frac{1}{T}\left((T-1)(6-4T)\left[\sum_{i=1}^{N} \frac{\pi_i^2}{p_i}\right]^2 + 4(T-1)(T-2)\sum_{i=1}^{N} \frac{\pi_i^3}{p_i^2} - \right. \tag{2}$$

$$\left. -4(T-1)\sum_{i=1}^{N} \frac{\pi_i^2}{p_i} \sum_{i=1}^{N} \frac{\pi_i}{p_i} + 6(T-1)\sum_{i=1}^{N} \frac{\pi_i^2}{p_i^2} - \left[\sum_{i=1}^{N} \frac{\pi_i}{p_i}\right]^2 + \sum_{i=1}^{N} \frac{\pi_i}{p_i^2} \right);$$

in this case the "noncentrality parameter" $\mathbf{E}X^2_{N,T} - (N-1)$ tends to infinity as a linear function of T. So, there are a vast space between the conditions of these two theorems. Moreover, in the case of convergence to the non-central chi-square distribution the latter depends on the noncentrality parameter and on N only, whereas in the case of asymptotic normality the asymptotic variance of $X^2_{N,T}$ depends essentially on all probabilities $\pi_1, \ldots, \pi_N$ (and usually in practice these probabilities are unknown).

2 Results

Using the algorithms of exact computation of Pearson statistics distributions (see [4, 5]) we investigate the accuracy of approximations of these distributions by noncentral chi-square and normal distributions.

The character of the dependence of the variance on $\pi_1, \ldots, \pi_N$ may be illustrated by the case $p_1 = \ldots = p_N = \frac{1}{N}$: here

$$\mathbf{E}X^2_{N,T} = N - 1 + (T-1)N\sum\nolimits_{i=1}^{N}\left(\pi_i - \tfrac{1}{N}\right)^2,$$

$$\mathbf{D}X^2_{N,T} = \frac{N^2}{T}\left((T-1)(6-4T)\left[\sum_{i=1}^{N}\pi_i^2\right]^2 + 4(T-1)(T-2)\sum_{i=1}^{N}\pi_i^3 + 2(T-1)\sum_{i=1}^{N}\pi_i^2\right).$$

For fixed values of N, T and of the noncentrality parameter $(T-1)N\sum_{i=1}^{N}\left(\pi_i - \frac{1}{N}\right)^2$ (i. e. fixed value of $\sum_{i=1}^{N}\pi_i^2$) the extremal values of $\sum_{i=1}^{N}\pi_i^3$ (and, consequently, $\mathbf{D}X^2_{N,T}$) are realized on the sets of probabilities of the form $(u_1, \ldots, u_1, u_2, \ldots, u_2, 0, \ldots, 0)$.

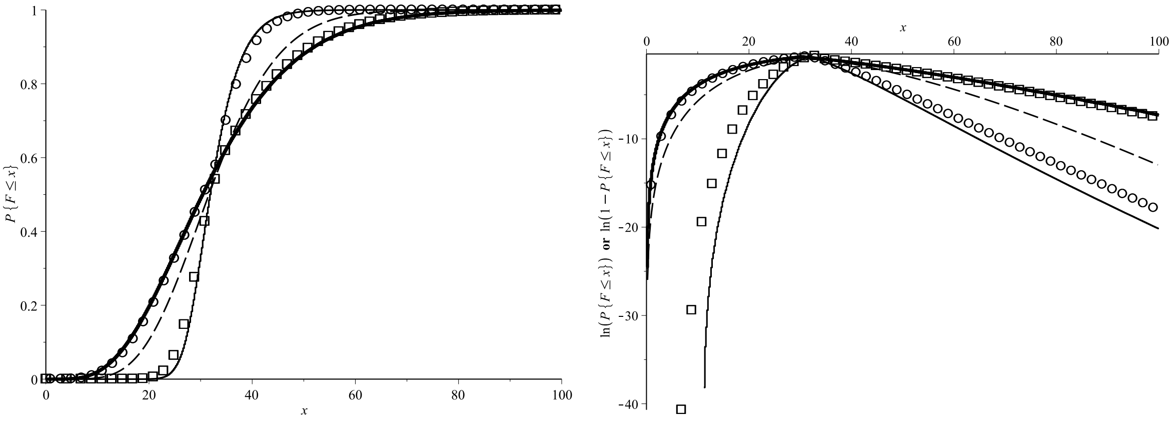

Figure 1: Distribution functions of $X^2_{10,100}$ with extremal values of $\mathbf{D}X^2_{10,100}$ and logarithms of their tails for $\lambda = 23.56$, and of noncentral chi-square with 9 degrees of freedom and noncentrality parameter $\lambda = 23.56$.

On the left part of Fig.1 for the case $N = 10$, $T = 100$, $\lambda = 23.56$ the graph of noncentral chi-square distribution with 9 degrees of freedom and noncentrality parameter λ (dotted line) and the graphs of exact distributions of $X^2_{10,100}$ for sets of probabilities realizing the minimal and maximal variances are presented; squares and circles correspond to minimal and maximal values of distribution function of $X^2_{10,100}$ observed for the random sample of sets of probabilities $\pi_1, \dots, \pi_N$ giving $\lambda = 23.56$. On the right part of Fig.1 for the same distribution functions $F(x)$ the graphs of $\ln \min\{F(x), 1 - F(x)\}$ are presented.

For the same parameters $N = 10$, $T = 100$ the differences between distribution functions of $X^2_{10,100}$ with maximal (minimal) variance and of non-central chi-square distribution with 9 degrees of freedom for three values of λ (1.14, 7.40, 23.56) are shown in the upper part of Fig.2. In the lower part of Fig.2 the corresponding differences between logarithms of tails are shown.

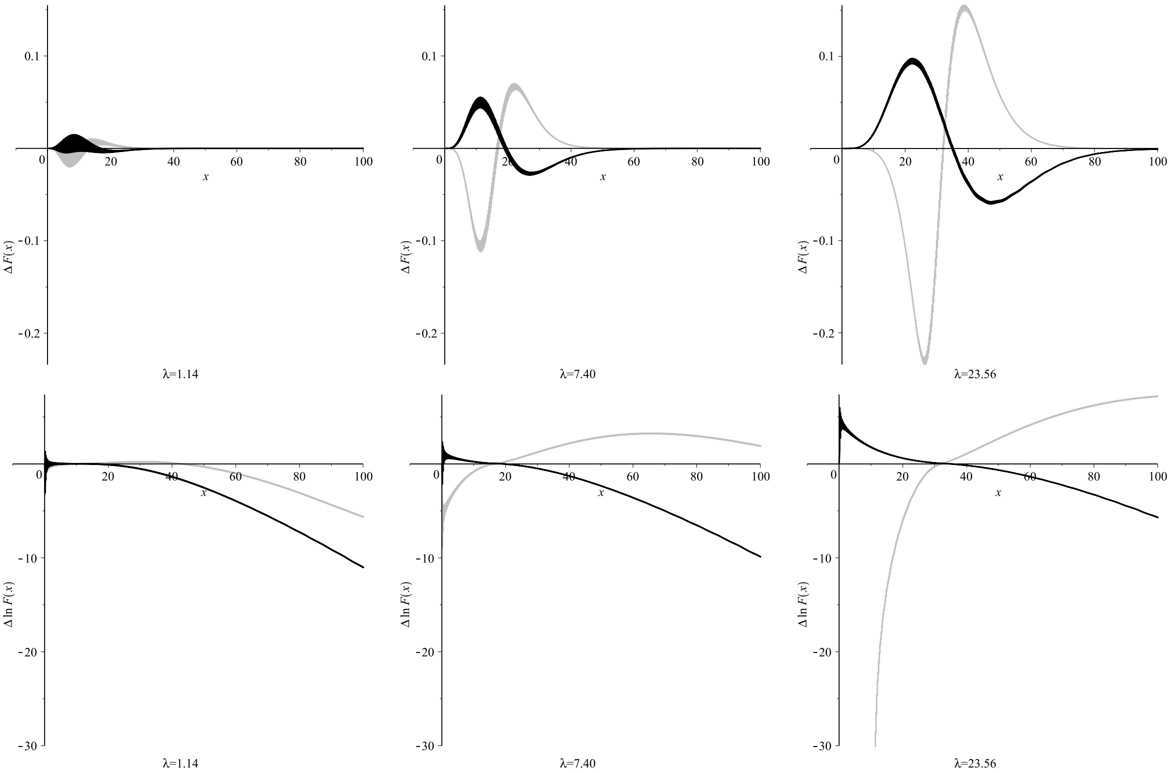

Figure 2: Differences between distribution functions and logarithms of tails for $N = 10, T = 100$.

If outcome probabilities $p_1, \dots, p_N$ are not equal, then the differences between distributions of the Pearson statistics (1) computed for samples with outcome probabilities $\pi_1, \dots, \pi_N$ with fixed value of the noncentrality parameter (i. e. the mean) appears to be larger and the sets of such N-dimensional vectors $\pi_1, \dots, \pi_N$ are asymmetrical. So, the investigation of forms and sizes of accurate confidence sets of probabilities based on the values of Pearson statistics as well as the power of tests appears to be a nontrivial problems.

References

[1] Broffitt J.D., Randles R.H. (1977). A power approximation for the chi-square goodness-of-fit test: Simple hypothesis case. *J. Amer. Stat. Assoc.* Vol. **72**(359), pp. 604-607.

[2] Yarnold J.K. (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *J. Amer. Stat. Assoc.* Vol. **65**(330), pp. 864-886.

[3] Yarnold J.K. (1972). Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. *Ann. Math. Stat.* Vol. **43**(5), pp. 1566-1580.

[4] Zubkov A.M., Filina M.V. (2008). Exact computation of Pearson statistics distribution and some experimental results. *Austrian J. Stat.* Vol. **37**(1), pp. 129-135.

[5] Zubkov A.M., Filina M.V. (2011). Tail properties of Pearson statistics distributions. *Austrian J. Stat.* Vol. **40**(1,2), pp. 47-54.