Limit distributions of the Pearson statistics for nonhomogeneous polynomial scheme

M.P. Savelov

Lomonosov Moscow State University,
Moscow Institute of Physics and Technology

Introduction

We consider nonhomogeneous polynomial scheme and find conditions on the probabilities of outcomes under which it is possible to describe limit distributions of the Pearson statistics.

Definitions

For a fixed $N \geq 2$ consider a series of trials with N outcomes and n independent trials in the n-th series.
H_0: the probability of the j-th outcome in the t-th trial of the n-th series is $p_j^t(n), j = 1, \ldots, N$.

Let $\nu_i(n)$ be the frequency of the j-th outcome. The Pearson statistics

$$X(n) := \sum_{j=1}^{N} \frac{(\nu_j(n) - np_j)^2}{np_j}$$

is widely used to test the hypothesis $H(\mathbf{p})$: «probabilities of outcomes constitutes a vector $\mathbf{p} = (p_1, \ldots, p_N)$».

Definitions

So, if $\pi_{N-1}(\alpha)$ is the $(1-\alpha)$-quantile of the χ^2_{N-1} distribution, then the rule «reject the hypothesis $H(\mathbf{p})$ if $X(n) > \pi_{N-1}(\alpha)$ and accept $H(\mathbf{p})$ otherwise» will give erroneous answer with probability approximately equal to α if the hypothesis $H(\mathbf{p})$ is true and n is sufficiently large.

Can the «large» value of the statistics $X(n)$ be explained by the fact that (for every j) p_j^t are periodic functions (with period L) of the variable t with $p_j = \frac{p_j^1 + \ldots + p_j^L}{L}$?

Assumptions

Suppose that the following conditions are satisfied:

1. $\sqrt{n}\max_{1\le i\le N}\left(\frac{p_i^1+\ldots+p_i^n}{n}-p_i\right)\to 0$ for $n\to\infty$,

2. There exists element-wise limit of martices $\lim_{n\to\infty}\left\|\frac{1}{n}\sum_{t=1}^n p_i^t p_j^t\right\|_{i,j=1}^N$.

Example 1. *Suppose for every* $1\le j\le N$ $p_j^t(n)$ *doesn't depend on* n *and is a periodic function of* t*. Then conditions 1–2 are satisfied.*

By definition, $A = \left\| \delta_{ij} - \lim_{n\to\infty} \frac{\sum_{t=1}^{n} p_i^t p_j^t}{n\sqrt{p_i p_j}} \right\|_{i,j=1}^{N}$. Denote by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$$

the eigenvalues of the matrix A (real by virtue of the symmetry A). We prove that $\lambda_1 \leq 1$, $\lambda_N = 0$.

Theorem 1. *If conditions 1 and 2 are satisfied then*

$$X(n) \xrightarrow{d} \sum_{i=1}^{N-1} \lambda_i \xi_i^2, \ n \to \infty,$$

where ξ_i are independent standard normal random variables.

Note that very similar statements about the behavior of $X(n)$ can also be obtained in the situation when instead of conditions 1–2 the stationarity of the sequence of observations is assumed (see Chanda, 1980).

Similar effects (there is the inequality between variances instead of inequality in the sense of stochastic order) in another situation are noted in Samuels, 1965.

Finally, in the paper of Selivanov, 2008, in a problem close to the one under consideration the behavior of Pearson statistics is established in the case when frequencies are centered by their expectations.

Suppose now that for some positive numbers $p_1, \ldots, p_N$ ($\sum_{j=1}^{N} p_j = 1$) the conditions 1–2 are satisfied, but the numbers $p_1, \ldots, p_N$ are unknown. In this case it is not possible to compare the values $\nu_j(n)$ and np_j by means of $X(n)$. Nevertheless, in this case it is possible to consider two disjoint parts of sample of the sizes m and n, and consider the natural statistics (used, for example, in tests for homogeneity of two samples, see Kramer, 1975) of the following form :

$$Y(n,m) = nm \cdot \sum_{j=1}^{N} \frac{\left(\frac{\nu_j(m+n)-\nu_j(n)}{m} - \frac{\nu_j(n)}{n}\right)^2}{\nu_j(n+m)}.$$

Theorem 2. *Suppose that the conditions 1–2 are satisfied. Let* ξ_i *be independent standard normal random variables. If* $m(n) \geq 1$ *is such that* $m(n) \to \infty$ *for* $n \to \infty$ *and* $\frac{n}{m(n)}$ *is bounded then* $Y(m(n), n) \xrightarrow{d} \sum_{i=1}^{N-1} \lambda_i \xi_i^2$ *for* $n \to \infty$.

Theorem 2 is a generalization of the classical result on the limit distribution of the statistic $Y(m, n)$ used in testing the homogeneity hypothesis (see Kramer, 1975).

Corollary 1. *Suppose for every* $1 \le j \le N$ $p_j^t(n)$ *doesn't depend on* n *and is a periodic function of* t *then*

$$Y(n,m) = nm \cdot \sum_{j=1}^{N} \frac{\left(\frac{\nu_j(m+n)-\nu_j(n)}{m} - \frac{\nu_j(n)}{n}\right)^2}{\nu_j(n+m)} \xrightarrow{d} \sum_{i=1}^{N-1} \lambda_i \xi_i^2,$$

$$X(n) := \sum_{j=1}^{N} \frac{(\nu_i(n) - np_j)^2}{np_j} \xrightarrow{d} \sum_{i=1}^{N-1} \lambda_i \xi_i^2, \ n \to \infty,$$

where $\xi_i, 1 \le i \le N$, *are independent and* $\xi_i \sim N(0,1)$.

Proposition 1. *Suppose that there exists* $1 \leq j_0 \leq N$ *such that* $\sqrt{n}\left(\frac{p_{j_0}^1+\ldots+p_{j_0}^n}{n} - p_{j_0}\right) \to \infty$. *Then* $\mathbf{E}X(n) \to +\infty$.

Thank you for attention!