CHI-SQUARE GOODNESS-OF-FIT TEST FOR ONE- AND MULTIDIMENSIONAL DISCRETE DISTRIBUTIONS

V. G. Voinov and M. S. Nikulin

UDC 519.2

The problem of the construction of chi-square type tests for discrete (one- and multidimensional) distributions of exponential type is considered in detail. In particular, as an example, the Stirling distribution of the second kind is investigated and a table of the best unbiased estimators of certain functions of the distribution parameters is given.

1. INTRODUCTION

The problems of the application of chi-square type tests have been considered in several investigations in mathematical statistics, especially following the well-known paper by Chernoff and Lehmann (1954), where for the first time one has pointed out the complications arising at the application of the Pearson chi-square test with the use of the minimal likelihood estimates. Later, in a survey article, Chibisov (1971) (see also Moore (1971)) has shown that the limiting distribution of the standard Pearson statistic \( \chi^2_n (\theta^*_n) \) depends on the method by which the estimate \( \theta^*_n \) of the unknown parameter \( \theta \) has been obtained.

In 1973, in [9]-[12] we have considered the question of the application of chi-square type tests for testing normality [9], for testing the hypothesis that a distribution function belongs to a family of continuous distribution functions \( \mathcal{F} (x; \theta) \) with shift and scale parameters [10], and also the general case [11] when as an estimate of the unknown parameter of a continuous distribution one makes use of an asymptotically efficient estimate (for example, the maximum likelihood estimate). In [11] we have given a general rule for the construction of the statistic of the test \( Y^2_n (\theta^*_n) \) which has as limit when \( n \to \infty \) (in the case of the validity of the hypothesis) the chi-square distribution with \( r - 1 \) degrees of freedom, where \( r \) is the number of classes in the grouping of the observations. In [12], [1], [15], [14] one has shown the possibility of the application of chi-square type tests, based on the statistic \( Y^2_n (\theta^*_n) \) for the construction of quantile tests [12], in the problems of homogeneity of two [1] and several [15] samples, and also in the bifactorial model of variance analysis [14]. In all the above mentioned investigations one has used ungrouped data for the estimation of the unknown parameters. In 1974, Dzhaparidze and Nikulin [6] have suggested the statistic \( W^2_n (\theta^*_n) \), whose limiting distribution is invariant with respect to the estimation methods, leading to \( \sqrt{n} \)-consistent estimates \( \theta^*_n \) of the unknown parameter \( \theta \). In [6], [24] and McCulloch (1985) one has pointed out the relations among the above mentioned statistics. In this connection one should also mention the paper by Rao and Robson (1974) in which one has considered the problem of the construction of the statistic \( Y^2_n (\theta^*_n) \) for the exponential distribution, whose density is \( f(x; \theta) = \frac{1}{\theta} \exp \left\{ -\frac{x}{\theta} \right\} \), \( \theta > 0 \), and also the paper by Le Cam, Mahan, and Singh [24], and that of Dudley [20], [21]. Finally, we mention here the last results of Drost (1988, 1989) in which one has considered, in particular, questions of the power and of the selection of the number of grouping intervals of the observations at the use of the statistics \( Y^2_n \) and \( W^2_n \).

As far as discrete distributions are concerned, here the first results are due to Kolchin, Sevast'yanov, and Chistyakov (1968), Park (1973), Bol'shev and Mirvaliev (1978) (see also [4]), who have considered the problems of testing the presence of a Poisson, geometric, and negative binomial distributions, respectively. We note that, in all the cases just mentioned, the sum of the observations is a sufficient statistic and, in addition, in these investigations one has started to use methods of unbiased estimation for the construction of unbiased estimates with minimal variance. Precisely because of this, these investigations, together with Gupta's paper [22] (1974), have given the possibility to make a natural generalization to an entire class of the so-called modified power series distributions, for which the sum of the observations is a complete sufficient statistic.

In [27] and [28] one presents in detail a technique for the construction of chi-square type tests, based on the statistic $Y_n^2$, for modified power series discrete distributions. In the next section we present briefly the fundamental results of these communications. We mention only that in the same investigations one has considered the problem of the construction of a chi-square test for testing the hypothesis that the density $f(x; \theta)$ of the distribution of the observation $X_i$ of a sample $X_1, ..., X_n$ belongs to the family of continuous distributions of exponential type

$$f(x; \theta) = \begin{cases} \frac{\alpha(x)}{\beta(\theta)} \exp \{x \psi(\theta)\}, & \alpha \in \mathcal{A}, \quad \theta \in \Theta, \\ 0, & \text{otherwise}, \end{cases}$$

where for the construction of the statistic $Y_n^2$ one makes use of the best unbiased estimator $\hat{f}(x)$ of the value $f(x; \theta)$, constructed from the sufficient statistic $S_n = (X_1 + ... + X_n)/n$, similarly to the manner in which this has been suggested in [10] for the construction of a chi-square test for testing normality, and used also in [7] for the application of the Kolmogorov and $\omega^2$ tests to continuous distributions with the shift and scale parameters. A more detailed survey of the results, presented here, on the theory of tests of chi-square type can be found in [5], [17], [18], [24], [28], [29], [35].

2. A CHI-SQUARE TEST FOR MODIFIED POWER SERIES DISCRETE DISTRIBUTIONS (ONE-DIMENSIONAL CASE)

Suppose that one tests the hypothesis $H_0$, according to which the independent, identically distributed random variables $X_1, ..., X_n$ are subjected to the modified power series distribution, i.e.,

$$P_\theta(x) = P_\theta\{X_i = x \mid H_0\} = \begin{cases} \frac{\alpha(x) \psi^2(\theta)}{f(\theta)}, & x \in \mathcal{X}, \quad \theta \in \Theta, \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathcal{X} = \{0, 1, 2, \ldots\}$, $\Theta = \{\theta : 0 < \theta < \rho\}$, $\rho$ is the radius of convergence of the series $f(\theta) = \sum x, \psi^2(\theta)$, $\alpha(x)$ is a positive function on $\mathcal{X}$, $\psi(\theta)$ is a positive, finite, differentiable function on $\Theta$. The probability distribution (1) has been investigated for the first time by Gupta [22]. The family (1) contains the binomial, the Poisson, the negative binomial distributions, the logarithmic series distribution, the generalized Poisson distributions, and also several other known probability distributions, including left, right, and two-sided truncated distributions.

For example, selecting $\alpha(x) = 1/\mathcal{I}!$, $\psi(\theta) = \theta$ and $f(\theta) = e^\theta$, we obtain the Poisson distribution with parameter $\theta$, $\theta > 0$. Setting $\psi(\theta) = (1-\theta)^k$, $\theta \in \Theta \subset (0; 1)$, $k > 0$, $\psi(\theta) = \theta$ and $\omega(x) = \Gamma(k+x)/\Gamma(k) x!$, we obtain the negative binomial distribution. Setting $\psi(\theta) = \theta(1-\theta)^k$, $f(\theta) = (1-\theta)^k$, $\theta \in \Theta \subset (0; 1)$, $\omega = \{1, 2, \ldots\}$, we obtain the generalized negative binomial distribution, etc. Other examples can be found, for example, in [3].

Since $\psi^2(\theta) = \exp \{x \ln \psi(\theta)\}$, we can see that formula (1) gives a family of discrete probability distributions of exponential type of rank (order) 1. From the factorization theorem it follows that the family (1) has the sufficient statistic $S_n = X_1 + ... + X_n$, where for any $s \in \mathcal{X}$ we have

$$P_\theta\{S_n = s\} = \begin{cases} \frac{\delta(s, n) \psi^2(\theta)}{f^n(\theta)}, & s \in \mathcal{X}, \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\delta(s, n) = \sum_{x_1+...+x_n = s} \prod_{i=1}^n a(x_i), \quad x_i \in \mathcal{X}, \quad \text{i.e.} \quad \delta(s, n) - \text{...}$$