The 7th Korea-Russia International Symposium on Science and Technology (KORUS 2003). University of Ulsan. June 28-July 6, 2003. – Vol.3. – P.126-132.

RULES OF APPLICATION OF GOODNESS-OF-FIT TESTS IN SIMPLE AND COMPOSITE HYPOTHESIS TESTING

B.Yu. Lemeshko, S.N. Postovalov, E.V. Chimitova
Novosibirsk State Technical University,
20, Karl Marx Prospect, Novosibirsk 630092, Russian Federation
Tel: +7 (383-2) 463457, Fax: +7 (383-2) 460209
E-mail: headrd@fpm.ami.nstu.ru

Abstract. **Problems of application of nonparametric goodness-of-fit tests and χ^2 type tests have been considered.**

The following points concerning the tests of χ^2 type have been considered: (a) correctness problems in usage of χ^2_{k-r-1}-distributions as the limiting distribution laws depending on estimation method used; (b) grouping methods providing maximal test power for close alternative hypotheses; (c) choice of the optimal interval number by the maximal test power. Optimal grouping tables have been constructed.

Nonparametric tests of Kolmogorov type, ω^2 and Ω^2 Mises type lose their property of "independence from distribution" in composite hypothesis testing. The limiting distribution laws depend on the distribution corresponding to the hypothesis under test, number and type of estimated parameters, their values and the estimation method. The models of limiting distributions for nonparametric test statistics have been constructed for a number of composite hypotheses.

Obtained results are included in the GOSSTANDART recommendations of Russia "Applied statistics. Rules of check of experimental and theoretical distribution of the consent". Part 1 – Goodness–of–fit tests of a type chi-square (P 50.1.033-2001), part 2 – Nonparametric goodness–of–fit tests (P 50.1.037-2002)

Keywords: Tests of chi-square type, tests of Kolmogorov type, tests of Mises type, composite hypothesis testing.

The usage of the tests to verify empirical data goodness-of-fit to theoretical distribution law is coursed by a number of conditions ensuring an adequate problem solution. Unfortunately, these conditions are seldom discussed in the works that used as a manual. Hence, in spite of the apparent simplicity, the practice of using the goodness-of-fit tests abounds with the examples of incorrect or inefficient use of the tests, especially in case of composite hypothesis testing.

With goodness-of-fit tests it is possible to verify simple hypotheses in the form H_0: $F(x) = F_0(x,\theta)$, where $F_0(x,\theta)$ is the probability distribution function, to which the sample of independent identically distributed observations $X_1, X_2, \ldots, X_n$ is tested for fit, θ is known parameter value (scalar or vector), and composite hypotheses in the form H_0: $F(x) \in \{F_0(x,\theta), \theta \in \Theta\}$, where Θ is the parameter space. In composite hypothesis testing the parameter estimate $\hat{\theta}$ is calculated from the same sample.

The procedure of hypothesis checking by means of χ^2 tests provides for splitting the random variable domain into k intervals by boundary points

$$x_0 < x_1 < \ldots < x_{k-1} < x_k .$$

The Pearson’s statistic X_n^2 is calculated according to the statement:

$$X_n^2 = n\sum_{i=1}^{k}\frac{(n_i / n - P_i(\theta))^2}{P_i(\theta)}, \tag{1}$$

where n_i – the number of observations fallen into the i-th interval, $P_i(\theta) = \int_{x_{i-1}}^{x_i} f_0(x,\theta)dx$ – the probability of an observation being in the i-th interval, $n = \sum_{i=1}^{k} n_i$, $\sum_{i=1}^{k} P_i(\theta) = 1$. If a simple hypothesis H_0 is true the limiting statistic distribution $G(X_n^2|H_0)$ is the χ^2-distribution with the freedom degree number equal to $k-1$. If m parameters of distribution law are estimated from the sample by minimizing the statistic X_n^2 than the statistic obeys the χ^2- distribution with the freedom degree number $k-m-1$. If an alternative hypothesis H_1 is true the limiting statistic distribution $G(X_n^2|H_1)$ is the noncentral χ^2- distribution with the same freedom degree number and the noncentrality parameter

$$s(\theta) = \sum_{i=1}^{k}\frac{c_i^2(\theta)}{P_i(\theta)}, \tag{2}$$

where $c_i(\theta) = \sqrt{n}\int_{x_{i-1}(\theta)}^{x_i(\theta)} \left(f_1(x,\theta) - f_0(x,\theta)\right)dx$ and $f_1(x,\theta)$ corresponds to an alternative hypothesis.

Originally for composite hypothesis testing and estimating of the parameters from the sample the use of χ^2_{k-m-1}-distributions as the limiting law was assumed to be correct only if the estimates were calculated by minimizing the statistic X_n^2. Later it was shown that X_n^2 has the same χ^2_{k-m-1}-distribution even in case if maximum likelihood estimates (MLE) for grouped observations are used.

By statistical modeling technique we have investigated distributions of this statistic for composite hypothesis testing and using MLE from grouped observations (for finite samples) and it has been confirmed there is a close fit of obtained empirical statistic distributions to the χ^2_{k-m-1}-distributions. Furthermore, we have revealed that the χ^2_{k-m-1}-distributions have every reason to be used as the limiting law for the statistic X_n^2 also if the shift and scale parameters of distribution laws under test are estimated as the linear combinations of sample quantiles). Statistic modeling results confirmed that the statistic X_n^2 also has the χ^2_{k-m-1}-distributions if these estimates are used.

Data grouping is evident to result in information losses and these losses depend on grouping technique. In practice observations are usually splitted into intervals of equal length or equiprobable intervals at the best. In these cases information losses as well as the ability of criterions to distinguish close hypotheses are different.

The Fisher information is a measure of inside closeness between random variable distributions, and this inside nature is associated to the power of distinguishing between close parameter values. As a statistic doesn’t have more information than a source sample than the distinguishing power by a statistic is not higher than by all sample. Hence, it is necessary to choose such statistic, for which information losses are minimal.

In other words, the less information losses because of observation grouping, the higher power of corresponding goodness-of-fit tests for close alternative hypotheses. Information losses can be decreased by selecting boundary points so, that $\mathbf{J}_{\Gamma}(\theta)$ tends to the information matrix for nongrouped data $\mathbf{J}(\theta)$, i.e. by solving asymptotically optimal grouping problem. In case of scalar parameter the problem reduces to the maximization of Fisher information quantity on the parameter for grouped sample

$$\max_{x_0<x_1<\ldots<x_{k-1}<x_k} \sum_{i=1}^{k} \left(\frac{\partial \ln P_i(\theta)}{\partial \theta} \right)^2 P_i(\theta) .$$

And in case of vector parameter different functionals of Fisher information matrix can be chosen. For example, the determinant of information matrix can be maximized, as it has been done in this case, i.e. to solve the problem

$$\max_{x_0<x_1<\ldots<x_{k-1}<x_k} \det \mathbf{J}_{\Gamma}(\theta) .$$

Generally the Fisher information matrix depends not only on boundary points x_i, but on the parameters of a tested distribution. However by solving the asymptotically optimal grouping problem interval boundary points in the invariant form relative to distribution parameters have been obtained and the corresponding asymptotically optimal grouping tables have been made for rather broad number of distributions.

Pearson’s chi-squire statistic distributions depend on the method of splitting random variable domain into intervals [1]. Grouping method influences the distributions $G(X_n^2 | H_1)$ and hence it affects the Pearson’s test power: the criterion has the maximal power for close alternatives and asymptotically optimal grouping.

When using chi-squire goodness-of-fit tests, there is an ambiguity in test (statistic) construction due to the choice of number of intervals and boundary points, i.e. how the random variable domain is splitted into intervals. It is obvious that the number of intervals and the grouping method should be chosen by the maximal test power. Though, this subject escapes from any regulating documents or literary sources.

The power of chi-squire tests essentially depends on the interval number k. The test power is known to decrease from a certain value with increasing of interval number k. As a matter of fact, it is possible to find the optimal value of grouping interval number depending on a couple of alternative hypotheses, grouping method and the sample size n. Knowing the limiting distributions $G(S \mid H_0)$ and $G(S \mid H_1)$ of the statistic S, it is possible to estimate the test power for any significance level α given, considering it as the function of interval numbers $x(\alpha, r)$ with the sample size n given. In [2] the power of Pearson’s test as the function of $x(\alpha, r)$ and n was investigated analytically and by means of statistical modeling. And the results of analytical calculations turned out to be entirely justified with the power estimates, obtained by modeling.

The power value for the tests of χ^2 type can be calculated according to the statement [3]:

$$1-\beta = \mathrm{P}(\mathrm{s} \mid r, \alpha) = e^{-\mathrm{s}/2} \sum_{j=0}^{\infty} \frac{\mathrm{s}^j}{j!\,2^{2j-1+r/2}\,\Gamma(j+r/2)} \times \int_{\sqrt{x(\alpha,r)}}^{\infty} y^{2j-1+r} e^{-y^2/2} dy , \qquad (3)$$

where s is the noncentrality parameter defined by (2), $x(\alpha, r)$ is $(1-\alpha)$-percentage point of χ_r^2-distribution with r degrees of freedom (α is the given probability of error type I (alpha error), β is the probability of error type II (beta error)). All the power functions, represented below, have been constructed with the significance level $\alpha = 0.1$.

In the figure 1 the power functions of Pearson's χ^2 test are represented depending on the interval number k in case of equiprobable and asymptotically optimal grouping for the sample size n equal to 500 and 5000 in testing simple hypothesis of fit to the exponential distribution law (H_0: $f_0(x) = \theta \exp\{-\theta x\}$ for $\theta = 1$; against H_1: $f_1(x) = \theta \exp\{-\theta x\}$ for $\theta = 1.05$). In both cases the test power decreases with the growth of k, but for asymptotically optimal grouping it is higher than in equiprobable one.

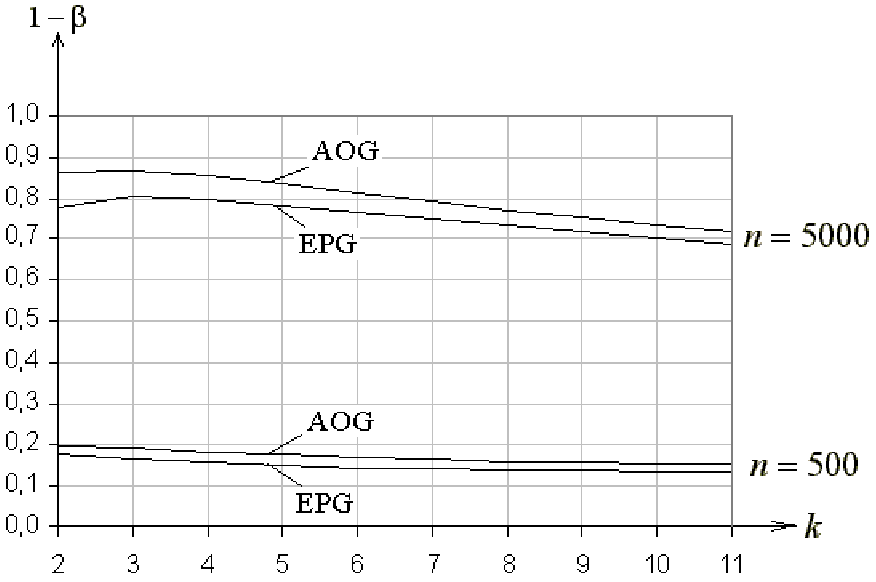

Figure 1

The ability of any statistical criterions to distinguish hypotheses, i.e. their power, increases with the sample size growth. When n is small it is very difficult to distinguish a pair of close hypotheses as distributions $G(S|H_0)$ and $G(S|H_1)$ turn out to be very close.

The most commonly used nonparametric goodness-of-fit tests include Kolmogorov tests and also ω^2 and Ω^2 Mises tests. The value

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x, \theta)|,$$

where $F_n(x)$ is the empirical distribution function, $F(x, \theta)$ is the theoretical distribution function, and n is the sample size, is used as a distance between the empirical and theoretical laws in Kolmogorov test. For testing hypotheses, one usually uses statistic of the form [3]

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}},$$

where

$$D_n = \max(D_n^+, D_n^-)\ ,\ D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\},\ D_n^- = \max_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\},$$

$x_1, x_2, \ldots, x_n$ are sample values in increasing order, and $F(x)$ is the distribution function, fit to which is tested. The distribution of statistic S_K in testing the simple hypothesis in the limit obeys Kolmogorov law $K(S)$ [3].

In tests of the type of ω^2, the distance between the hypothetical and the true distributions is considered in the quadratic metric

$$\int_{-\infty}^{\infty} \{E[F_n(x)] - F(x)\}^2 \psi(F(x)) dF(x),$$

where $E[\cdot]$ is the mathematical expectation operator.

In choosing $\psi(t) \equiv 1$ in Mises ω^2 tests, one uses a statistic (Cramer – Mises – Smirnov statistic) of the form

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i,\theta) - \frac{2i-1}{2n}\right\}^2 .$$

In testing a simple hypothesis it obeys the distribution $a1(S)$ [3].

In choosing $\psi(t) \equiv 1/t(1-t)$ in Mises Ω^2 tests, the statistic (Anderson – Darling statistic) has the form

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i,\theta))\right\}.$$

In the limit, this statistic obeys the distribution $a2(S)$ [3].

In the case of simple hypotheses, the limiting statistic distributions of the nonparametric Kolmogorov, ω^2 and Ω^2 Mises tests are known for a long time. These tests are said to be “distribution-free” tests. However, the nonparametric test power in testing composite hypotheses for the same sample sizes is always much higher than that in testing simple ones. And whereas in testing simple hypotheses the nonparametric Kolmogorov, ω^2 and Ω^2 Mises tests have a lower power compared with the χ^2-type tests provided that the latter use the asymptotically optimal grouping, in testing composite hypotheses the nonparametric tests appear to be more powerful. To make use of their advantages, we must merely know the distribution $G(S|H_0)$ for the tested composite hypotheses.

While testing composite hypotheses, when the same sample is used to estimate the parameters of the observed law $F(x,\theta)$, the nonparametric goodness-of-fit tests lose the property of “distribution–freeness”.

Paper [4] was the pioneer in investigating the limiting statistic distributions of the nonparametric goodness-of-fit tests in testing the composite hypotheses. The literature presents several approaches to investigating the nonparametric goodness-of-fit tests in the case of testing the composite hypotheses [5-11].

It has been found that in composite hypothesis tests, the conditional distribution law of the statistic $G(S|H_0)$ is affected by a number of factors determining the hypothesis complexity:

the form of the observed law $F(x,\theta)$ corresponding to the true hypothesis H_0; the type of the parameter estimated and the number of parameters to be estimated; sometimes, it is a concrete value of the parameter (e.g., in the case of gamma-distribution); the method of parameter estimation.

Constructing of the limiting distribution by analytical methods is an extremely complicated problem. It is most suitable to use the method of computer analysis of statistical regularities. The method showed good results in simulating the test statistic distributions. Implementation of such procedure of computer analysis of statistic distributions contains neither difficulties of principal nor practical difficulties at present. In [12-14] we constructed models approximating the limiting statistic distributions for some composite hypotheses.

Conclusion

Asymptotically optimal grouping maximizes the power of Pearson's χ^2 test and likelihood ratio test with respect to close alternative hypotheses for both simple and composite hypotheses. Moreover, the asymptotically optimal grouping tables contain the probability values of an observation being in an interval, what facilitates the computation process.

The choice of too large interval number results test power loss. The optimal interval number k depends on the sample size n and the specified pair of alternative hypotheses H_0 and H_1. Usually the optimal k turns out to be much smaller than the numbers recommended by different regulating documents and given by a great quantity of empirical formulas. The maximal test power for the given sample size n is frequently reached with the minimally possible or rather small interval number k.

The results of research into χ^2 test power depending on grouping method and interval number as well as the tables of asymptotically optimal grouping underlie the standardization recommendations [15], developed by us.

The constructed approximations of the limiting statistic distributions of the nonparametric goodness-of-fit tests extend the region of correct application of these tests and may be recommended for construction of statistical regularities when it is impossible to solve the problem analytically.

On the basis of obtained models standardization recommendations [16] are developed. The percentile point tables and constructed models of nonparametric test statistic distributions of Kolmogorov type, ω^2 Mises and Ω^2 Anderson-Darling type are represented in recommendations [16] for testing various composite hypotheses of goodness-of-fit to exponential, seminormal, Rayleigh, Maxwell, Laplace, normal, log-normal, Cauchy, logistic, maximum-value, minimum-value, Weibull, gamma-, Sb-Johnson, Sl-Johnson, Su-Johnson distributions.

References

[1] Lemeshko B.Yu., Postovalov S.N. // Zavodskaya laboratoria. Diagnostika materialov. 1998. – Vol. 64. – № 5. – P. 56-63.

[2] Lemeshko B.Yu., Chimitova E.V. // Zavodskaya laboratoria. Diagnostika materialov. 2003. – Vol. 69. № 1. – P. 61-67.

[3] Bolshev L.N., Smirnov N.V. Tablitsy matematicheskoi statistiki. – M.: Nauka, 1983. – 416 p.

[4] Kac M., Kiefer J., Wolfowitz J. // Ann. Math. Stat. 1955. – Vol. 26. P. 189.

[5] Durbin J. // Lect. Notes Math. 1976. – Vol. 566. P. 33.

[6] Martynov G.V. Omega-Square Tests. Nauka, Moscow, 1978 (in Russian).

[7] Pearson E.S., Hartley H.O. Biometrica tables for Statistics, University Press, Cambridge, 1972, vol. 2.

[8] Stephens M.A. // Journ. Amer. Statist. Assoc. 1974. – Vol. 69. P. 730.

[9] Chandra M., Singpurwalla N.D., Stephens M.A. // Journ. Amer. Statist. Assoc. 1981. – Vol. 76. P. 375.

[10] Tyurin Yu. N. // Izv. AN SSSR. Ser. Mat. 1984. – Vol. 48, no. 6. P. 1314.

[11] Tyurin Yu. N., Savvushkina N.E. Izv. AN SSSR. Ser. Tekhn. Kibernetika. 1984. – No. 3, p. 109.

[12] Lemeshko B.Yu., Postovalov B.Yu. // Zavodskaya laboratoria. Diagnostika materialov. 1998. – Vol. 64, no. 3, p. 61.

[13] Lemeshko B.Yu., Postovalov S.N. // Optoelectronics, Instrumentation and Data Processing. 2001. - № 2. - P. 76-88.

[14] Lemeshko B.Yu., Postovalov S.N. // Zavodskaya laboratoria. Diagnostika materialov. 2001. Vol. 67. - № 7. – P. 62-71.

[15] P 50.1.033-2001. Standardization recommendations. Applied statistics. Rules for check of experimental and theoretical distribution on the consent. Part I. Goodness-of-fit tests of a type chi-squire. – M.: Izdatelstvo standartov. 2002. – 87 p.

[16] P 50.1.037-2002. Standardization recommendations. Applied statistics. Rules for check of experimental and theoretical distribution on the consent. Part II. Nonparametric goodness-of-fit test. – M.: Izdatelstvo standartov. 2002. – 67 p.