Nonparametric Goodness-of-Fit Tests for Discrete, Grouped or Censored Data[1]

Boris Yu. Lemeshko[2], Ekaterina V. Chimitova[2] and Stepan S. Kolesnikov[2]

[2] Novosibirsk State Technical University
Department of Applied Mathematics
20 Karl Marx
630092 Novosibirsk, Russia
(e-mail: headrd@fpm.ami.nstu.ru)

Abstract. The problems of application of nonparametric Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling goodness-of-fit tests for discrete, grouped and censored data have been considered in this paper. The use of these tests for grouped and censored data as well as samples of discrete random variables is based on Smirnov transformation. The convergence of statistic distributions to the corresponding limiting distribution laws has been investigated under true null hypothesis by means of statistical simulation methods, as well as the test power against close competing hypotheses. For discrete and grouped data the criteria have been compared by power with Pearson chi-squire test. The criteria have been also compared by power with the modified nonparametric tests for censored samples.
Keywords: Goodness-of-fit tests; discrete, grouped, censored data; Smirnov transformation; Kolmogorov test, Cramer-von Mises-Smirnov test, Anderson-Darling test.

1 Introduction

In case of discrete or grouped data there are no evident problems with testing simple hypotheses about goodness-of-fit of an empirical distribution to theoretical law only if χ^2 goodness-of-fit tests are being used. Direct application of Kolmogorov, ω^2 Cramer-von Mises-Smirnov or Ω^2 Anderson-Darling tests is impossible, as the limiting statistic distributions for these criteria are obtained on the assumption of random variable continuity.

For testing simple goodness-of-fit hypotheses from right and/or left censored samples one can use the Renyi test [Renyi, 1953], Kolmogorov-Smirnov [Barr and Davidson, 1973], ω^2 Cramer-von Mises-Smirnov or Ω^2 Anderson-Darling [Pettitt and Stephens, 1976] modified tests. However in case of censored data, these criteria have a number of disadvantages embarrassing their application in practice.

In particular, Renyi statistic distribution converges to the limiting law very slowly, especially for high or, on the contrary, low censoring degree [Lemeshko and Chimitova, 2004]. The distributions of modified Kolmogorov-Smirnov, Cramer-von Mises-Smirnov and Anderson-Darling tests converge rather quickly

[1] This research was supported by the Russian Foundation for Basic Research, project no. 06-01-00059

to the corresponding limiting laws for small censoring degree [Lemeshko and Chimitova, 2004]. Application of the criteria for censored data hasn't been realized almost in any known for us software system of statistical analysis. And hence they are hardly available for a large number of specialists.

M. Nikulin has attracted our attention to the possibility of effective application of nonparametric goodness-of-fit tests for the analysis of grouped and censored data and samples of discrete random variables by means of Smirnov transformation and the "randomization", enabling to move from "staircase" and discontinuous distribution function to the continuous one [Greenwood and Nikulin, 2006]. The advantages of such approach are evident as we move to the problem of testing goodness-of-fit of the empirical distribution obtained after transformations to the continuous (uniform) distribution law.

Smirnov transformation is used rather often in statistical analysis. Let us test whether the random sample $X_1, X_2, ..., X_n$ corresponds to the law with distribution function $F(x)$. The transformation $U_i = F(x_i)$ converts the observed sample of random variables $X_1, X_2, ..., X_n$ into the sample of values uniformly distributed on the interval [0, 1]. Then the hypothesis about belonging of $U_1, U_2, ..., U_n$ to the uniform law can be tested, for example, using the Kolmogorov criterion with statistic

$$D_n = \sup_{0 \le u \le 1} \sqrt{n} \left| F_n(u) - u \right|, \tag{1}$$

where $F_n(u)$ is the empirical distribution function.

The "randomization" as a technique of conversion of grouped and censored data and discrete variable observations to the continuous variable observation is really applicable only in computer analysis.

The purpose of the paper is to investigate some practical aspects of application of classical goodness-of-fit tests for the analysis of grouped and censored data and discrete variable observations in case of using the Smirnov transformation with "randomization". In the paper it has been studied the convergence of statistic distributions to the corresponding limiting laws, as well as the power of the considered criteria for testing close competing hypotheses.

2 Grouped and discrete data

Let us test simple hypothesis about goodness-of-fit of grouped sample to the theoretical distribution law $F(x)$. Grouped sample of the size n is given with the boundary points $x_0 < x_1 < ... < x_{k-1} < x_k$, where k is the number of intervals, x_0 and x_k are the left and right boundaries of the random variable domain respectively, and the number of observations n_i fallen into the i-th

interval, $\sum_{i=1}^{k} n_i = n$. Assume Y_{ij} $(i = 1,...k,\ j = 1,...n_i)$ are n independent realizations of the random variable uniformly distributed on [0,1]. Then the random variables obtained with "randomization" on the grouping intervals $(x_{i-1}, x_i]$,

$$U_{ij} = F(x_{i-1}) + Y_{ij}[F(x_i) - F(x_{i-1})],\ i = 1,...,k,\ j = 1,...,n_i, \quad (2)$$

are independent and uniformly distributed on [0,1].

The statement (2) allows [Greenwood and Nikulin, 2006] to move from grouped sample to "complete" sample of individual observations uniformly distributed on [0,1]. After that one can test the simple hypothesis about goodness-of-fit of the empirical distribution, built by the sample of values U_{ij}, $i = 1,...,k$, $j = 1,...,n_i$, to the uniform distribution using any nonparametric goodness-of-fit test.

A sample of observations $X_1, X_2,..., X_n$ of some discrete random variable can be similar to the grouped case transformed to the sample of uniformly distributed observations

$$U_i = F(X_i -) + Y_i[F(X_i) - F(X_i -)],\ i = 1,...,n, \quad (3)$$

where $F(x-) = \lim_{z \downarrow 0} F(x - z)$ and $Y_1, Y_2,..., Y_n$ are n independent realizations of the random variable uniformly distributed on [0,1]. In "randomization" the values Y_{ij} and Y_i in the statements (2) and (3) have to be simulated in accordance with the uniform distribution on [0,1].

In [Lemeshko and Postovalov, 2001] it was shown that nonparametric goodness-of-fit test statistic distributions in case of continuous distribution laws and complete samples converge to corresponding limiting laws very quickly. The limiting laws can be already used with $n \geq 20$ without risk of making a great mistake.

Nonparametric goodness-of-fit test statistic distributions have been investigated for discrete random variables and grouped samples of continuous values with the usage of considered approach. It has been shown that empirical distributions of nonparametric test statistics also converge with the sample size n growth to the corresponding limiting laws very fast. For example, in the figure 1 the limiting Kolmogorov law $K(S)$ and obtained after simulation of empirical distribution of Kolmogorov test statistic $G(K_n \mid H_0)$ are shown. The true hypothesis H_0 under test is about goodness-of-fit to the normal law. The empirical distribution is built by $N = 10000$ grouped samples of the size $n = 20$ with $k = 10$ grouping intervals in case of asymptotically optimal grouping method.

As the Kolmogorov test statistic we have used the statistic with Bolshev's correction

$$K_n = \frac{6nD_n + 1}{6\sqrt{n}}\ , \tag{4}$$

where $D_n = \max\{D_n^+, D_n^-\}$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(X_{(i)})\right\}$,

$$D_n^- = \max_{1\le i\le n}\left\{F(X_{(i)}) - \frac{i-1}{n}\right\}.$$

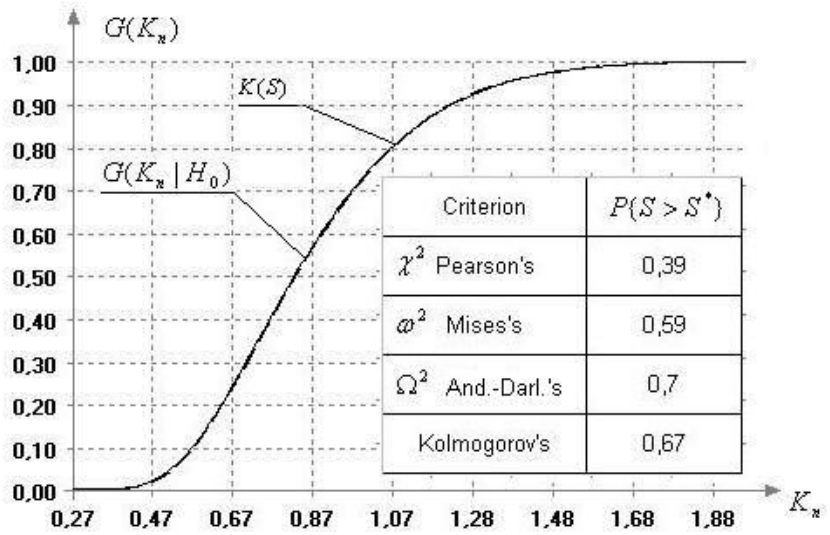

Fig. 1. The empirical distribution function of statistic (4) and the limiting Kolmogorov distribution law

The empirical distribution of the Kolmogorov statistic perfectly fits the Kolmogorov law $K(S)$ even for $n = 20$. This fact is also confirmed with the values of achieved significance level $P\{S > S^*\}$ while testing hypothesis about goodness-of-fit of the sample of statistic's (4) values to the Kolmogorov distribution $K(S)$ with χ^2 Pearson, ω^2 Cramer-von Mises-Smirnov, Ω^2 Anderson-Darling and Kolmogorov criteria. S^* is the value of corresponding goodness-of-fit test statistic.

The similar results about convergence of statistic distributions to the limiting laws for grouped data and discrete random variables have been obtained for Cramer-von Mises-Smirnov and Anderson-Darling criteria. It has been also

shown that the rate of convergence of $G(S_n \mid H_0)$ to the limiting laws of statistic S does not depend on the grouping method and the number of grouping intervals k.

3 Censored data

Let $X_1, X_2, ..., X_n$ be a sample of independent similarly distributed random variables. A set of values $X_{(1)} \le X_{(2)} \le ... \le X_{(r)}$ $(X_{(n-r)} \le X_{(n-r+1)} \le ... \le X_{(n)})$ is called a right/left censored sample, where $r < n$ is the number of complete observations, and the rest $n-r$ observations – are censored.

The modifications of nonparametric Kolmogorov-Smirnov, Cramer-von Mises-Smirnov and Anderson-Darling tests are introduced in [Barr and Davidson, 1973], [Pettitt and Stephens, 1976] for testing goodness-of-fit by censored samples. In particular the Kolmogorov statistic for censored data is defined by $K_n^c = \sup_M |F(x) - F_n(x)|$, where $M = \{x : F(x) \ge a\}$ for left censoring and $M = \{x : F(x) \le 1-a\}$ for right censoring, $a \in (0,1)$ is the censoring degree. The limiting distribution of the Kolmogorov statistic K_n^c for censored data is given as [Barr and Davidson, 1973]

$$P\{K_n^c < S\} = \sum_{i=-\infty}^{+\infty} (-1)^i \exp(-2i^2 S^2) P\left\{ \left| X - 2iS\sqrt{\frac{a}{1-a}} \right| < \frac{S}{\sqrt{a-a^2}} \right\} = K_c^a(S)$$

where X is the standard normal random variable. When $a = 0$ the limiting distribution of statistics K_n^c coincides with the Kolmogorov distribution $K(S)$.

As before it is possible to move from a censored sample to the sample of random variables $U_1, U_2, ..., U_n$, uniformly distributed on [0,1]. In case of right censoring we have $U_1 = F(X_{(1)})$, $U_2 = F(X_{(2)})$, …, $U_r = F(X_{(r)})$, and the values $U_{r+1}, U_{r+2}, ..., U_n$ are simulated uniformly on the interval $[F(x_c), 1]$, where x_c is the censoring point. In case of the first type censoring the point x_c is fixed and the number of complete observations r is random. In the second type censoring the last (first) observed value in sample is taken as x_c.

Classical Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests can be applied to analyze transformed sample.

The empirical distributions of statistic (4) and modified Kolmogorov statistic by the censored sample are represented in the figure 2. The corresponding limiting distributions are given in the figure for comparison. Statistic's values are

calculated by right censored samples from the exponential distribution of the sample size $n = 20$ and censoring degree 80% (the right part of random variable domain is inaccessible for observation, probability to fall in which is equal to $a = 0.8$).

The empirical distribution of Kolmogorov statistic K_n, calculated from the transformed samples, perfectly agrees with the limiting law $K(S)$ already for $n = 20$. At the same time the empirical distribution of the modified Kolmogorov statistic K_n^c, applied directly to censored samples of the same size, essentially differ from the limiting law $K_c^a(S)$.

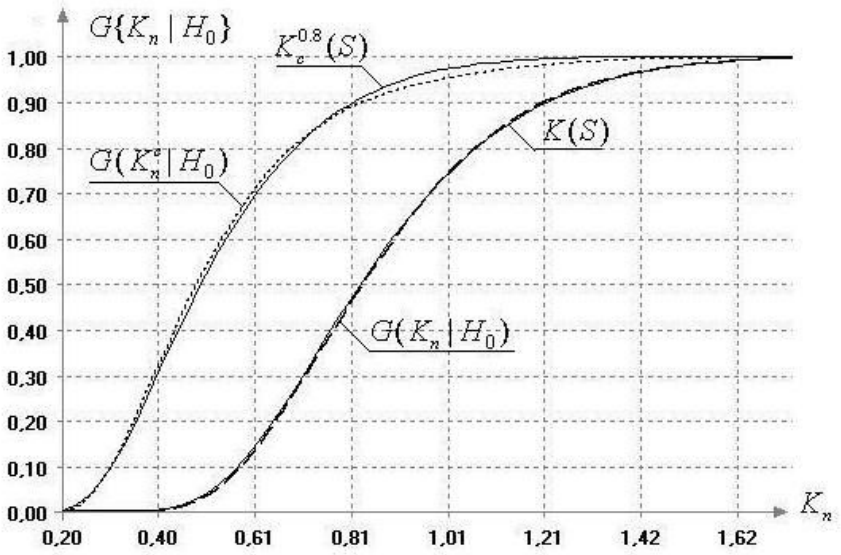

Fig. 2. The distributions of Kolmogorov test statistic in testing goodness-of-fit to the exponential law in case of $n = 20$ and censoring degree 80%

Distributions of statistic K_n (with Smirnov transformation and randomization) have been investigated with different types and degrees of censoring. It has been shown that the rate of convergence of empirical distributions $G(K_n \mid H_0)$ to $K(S)$ does not depend on the type and degree of censoring. Similar results have been obtained for ω^2 Cramer-von Mises-Smirnov and Ω^2 Anderson-Darling tests.

Empirical distributions $G(K_n^c \mid H_0)$ agree with the limiting law $K_c^a(S)$ rather well beginning with $n = 30$ only when censoring degree is less than 50% ($a < 0.5$). If the censoring degree increases up to 95%, sufficient closeness of

$G(K_n^c \mid H_0)$ to $K_c^a(S)$ takes place if $n \geq 500$ [Lemeshko and Chimitova, 2004].

4 Some remarks on the test power

There is no doubt that conclusions obtained in [Lemeshko et al., 2007], concerning the comparative analysis of the test power for close competing hypotheses, are also place for grouped samples.

For censored data it is worth comparing the power of classical criteria applied to the transformed data with the power of modified for censored samples tests [Barr and Davidson, 1973], [Pettitt and Stephens, 1976].

For example, the power of modified Kolmogorov test essentially depends on the censoring degree. By means of statistical modeling methods we have shown that the higher censoring degree the more modified Kolmogorov test exceeds by power the Kolmogorov test with Smirnov transformation and randomization. For small censoring degrees (approximately up to 30%) these criteria are close by power.

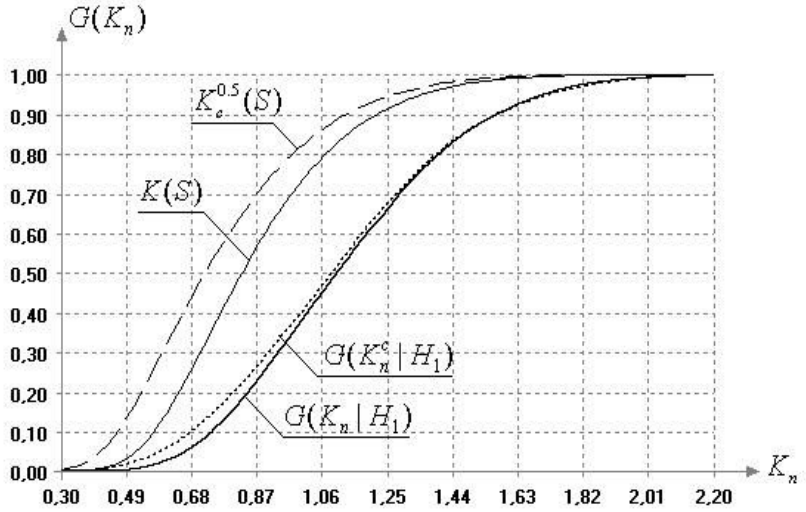

Fig. 3. The distributions of Kolmogorov statistic for the true hypothesis H_0 and H_1

The illustration (fig. 3) shows two cases of modified Kolmogorov test statistic distributions applied to censored sample and two cases of test statistic distributions calculated by the transformed sample $U_1, U_2, ..., U_n$. In the first case the hypothesis H_0, the Weibull distribution with the form parameter 3, is true; and in the second case the competing hypothesis H_1, the Weibull

distribution with the form parameter 3.5, is true. The sample size $n = 300$, second type right censoring, the censoring degree $a = 0.5$.

5 Conclusion

The results of investigation enable to conclude a good possibility to use the approach considered (Smirnov transformation with randomization) for correct application of classical nonparametric goodness-of-fit tests for grouped and censored data and samples of discrete random variables.

In case of simple hypothesis testing, nonparametric statistic distributions converge to statistic limiting distributions very quickly. For the sample size $n \geq 20$ one can use the limiting laws without risk of making a great mistake.

The influence of grouping methods on the power of nonparametric goodness-of-fit tests should be investigated in more detail.

The application of Smirnov transformation with randomization is quite efficient for realization in software systems of statistical analysis. It expands the possibilities of the classical nonparametric goodness-of-fit tests` application to grouped data and discrete random variables.

References

[Barr and Davidson, 1973] Barr D.M., Davidson T. A Kolmogorov-Smirnov test for censored samples. Technometrics, 1973. V. 15. N. 4.

[Greenwood and Nikulin, 1996] Greenwood P.E., Nikulin M.S. A Guide to Chi-Squared Testing. – John Wiley & Sons, Inc. 1996. – 280 p.

[Lemeshko and Chimitova, 2004] Lemeshko B.Yu., Chimitova E.V. Investigation of the estimates properties and goodness-of-fit test statistics from censored samples with computer modeling technique // Proceedings of the Seventh International Conference "Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods", September 6-10, 2004, Minsk. Vol. 1. – P. 143-146

[Pettitt and Stephens, 1976] Pettitt A.N., Stephens M.A. Modified Cramer von Mises statistics for censored data // Biometrika, 1976. V. 63. N. 2.

[Renyi, 1953] Renyi A. On the theory of order statistics // Acta Mathem. Acad. Sci. Hung. 1953. Vol. 4. – P. 191-232.

[Lemeshko and Postovalov, 2001] Lemeshko B.Yu., Postovalov S.N. On the dependence of nonparametric test statistic distributions and the test power on parameter estimation method // Zavodskaya Laboratoriya. Diagnostika materialov. 2001. Vol. 67. - № 7. - P. 62-71. (in Russian)

[Lemeshko et al., 2007] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. Power goodness-of-fit tests at close alternatives // Izmeritelnaya Technika. 2007. № 2. – P. 22-27. (in Russian)