DOI 10.1007/s11018-016-0992-3
Measurement Techniques, Vol. 59, No. 5, August, 2016

BIAS OF NONPARAMETRIC GOODNESS-OF-FIT TESTS RELATIVE TO CERTAIN PAIRS OF COMPETING HYPOTHESES

B. Yu. Lemeshko, P. Yu. Blinov, and S. B. Lemeshko

UDC 519.24

The application of the nonparametric Anderson–Darling, Cramer–Mises–Smirnov, Kuiper, Watson, Kolmogorov, and Zhang goodness-of-fit tests in verification of simple and composite hypotheses is considered. Based on an investigation of the power, it is shown for the first time that there exist pairs of competing hypotheses which these tests are not able to distinguish in the case of small sample sizes n and type 1 error probabilities. It is shown that the reason for this lies in the bias of the tests in corresponding situations.
Keywords: *Anderson–Darling test, Cramer–Mises–Smirnov test, Kuiper test, Watson test, Kolmogorov test, Zhang test, power of test.*

Using tests to verify statistical hypotheses in the analysis of the results of experiments, the researcher will now and then blindly trust in the results of an inference without thinking that the test is itself a mathematical tool intended for the measurement (detection, estimation) in analyzed data of some deviation from previous suggestions, and that this tool also requires appropriate "adjustment and verification." The set of statistical "tools" the use of which is justified for measurement of corresponding quantities is quite broad, and the tools themselves belong to different "accuracy classes." The skillful application of any measuring tool presupposes knowledge of its capabilities and field of application.

It is with this in mind that we wish to discuss certain facts that attest to the limiting capabilities of nonparametric goodness-of-fit tests (Kolmogorov, Cramer–Mises–Smirnov, Anderson–Darling, Kuiper, Watson, and Zhang goodness-of-fit tests) in distinguishing between probability distribution laws. The existence of such facts in connection with nonparametric goodness-of-fit tests has not been previously reported in the literature. Knowledge of these facts will help achieve a more understandable interpretation of the results from the use of tests for the verification of statistical hypotheses.

Recall that when goodness-of-fit tests are used it is necessary to distinguish between verification of simple and composite hypotheses. A simple verifiable hypothesis has the form H_0: $F(x) = F(x, \theta)$, where $F(x, \theta)$ is a probability distribution function with which the goodness of fit of an observed sample is verified and θ a known value of a parameter (scalar or vector).

A composite verifiable hypothesis has the form H_0: $F(x) \in (F(x, \theta), \theta \in \Theta\}$, where Θ is the domain of definition of the parameter θ. A difference arises if in the course of verification of a composite hypothesis an estimator $\hat{\theta}$ of the distribution parameter is calculated on the same sample based on which the goodness of fit is verified, since in that case the distribution of the statistic corresponding to the validity of the hypothesis H_0 differs significantly from that which exists in the case of a simple verifiable hypothesis [1].

In the process of verifying a hypothesis H_0 of the membership of a sample to a law with distribution function $F(x, \theta)$, the Kolmogorov goodness-of-fit test [2] relies on the statistic

Novosibirsk State Technical University, Novosibirsk, Russia; e-mail: Lemeshko@ami.nstu.ru. Translated from Izmeritel'naya Tekhnika, No. 5, pp. 16–20, May, 2016. Original article submitted July 27, 2015.

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

where $F_n(x)$ is an empirical distribution function; $F(x, \hat{\theta})$, the distribution function of the law; $\hat{\theta}$, estimator of the vector of parameters found from the same sample; and n, size of sample.

In verifying a hypothesis with the use of the Kolmogorov goodness-of-fit test it is recommended that a statistic with Bol'shev correction in the form [2]

$$S_{\mathrm{K}} = \sqrt{n} D_n + \frac{1}{6\sqrt{n}},$$

where

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

$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, here and below, elements of a variational series constructed from an initial sample $X_1, X_2, \ldots, X_n$, should be used.

In verification of a simple hypothesis H_0, the statistic obeys a Kolmogorov distribution $K(s)$ [2].

The quantity

$$V_n = D_n^+ + D_n^-,$$

where D_n^+ and D_n^- are defined above, is used as a measure of the distance between the empirical and theoretical law in the Kuiper goodness-of-fit test.

The statistic [4]

$$V = V_n \left(\sqrt{n} + 0.155 + 0.24 / \sqrt{n} \right)$$

or the statistic [5]

$$V_n^{\mathrm{mod}} = \sqrt{n} V_n + \left(3\sqrt{n} \right)^{-1}$$

may be used as the statistic of the test.

The limiting distribution of these statistics where the simple hypothesis H_0 is valid was found in [3] and presented in [1].

The ω^2 statistic of the Cramer–Mises–Smirnov goodness-of-fit test has the form

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

In the case of a simple verifiable hypothesis when the parameters of the theoretical law $F(x, \theta)$ are known, this statistic (where the H_0 hypothesis is valid) obeys in limit a law with distribution function $a1(s)$ [2].

The statistic of the Watson goodness-of-fit test [6, 7] is used in the following form which is convenient for calculations:

$$U_n^2 = \sum_{i=1}^{n} \left(F(x_{(i)}, \hat{\theta}) - \frac{i - 1/2}{n} \right)^2 - n \left(\frac{1}{n} \sum_{i=1}^{n} F(x_{(i)}, \hat{\theta}) - \frac{1}{2} \right)^2 + \frac{1}{12n}.$$

The limiting distribution of this statistic where the simple verifiable hypothesis is valid is found in [6, 7].

The statistic

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

is used in the Anderson–Darling goodness-of-fit Ω^2 test [8, 9].

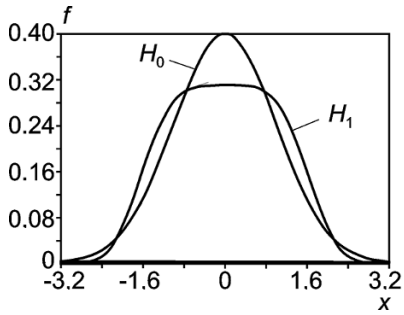

Fig. 1. Densities corresponding to normal law (H_0) and generalized normal law (H_1).

If a simple verifiable hypothesis H_0 is valid when the parameters of the theoretical law $F(x, \theta)$ are known, this statistic in limit obeys a law with distribution function $a2(s)$ [2].

Nonparametric goodness-of-fit tests whose statistics have the following form are proposed in [10, 11]:

$$Z_A = -\sum_{i=1}^{n}\left[\frac{\ln\left\{F(x_{(i)},\hat{\theta})\right\}}{n-i+1/2}+\frac{\ln\left\{1-F(x_{(i)},\hat{\theta})\right\}}{i-1/2}\right];$$

$$Z_C = \sum_{i=1}^{n}\left[\ln\left\{\frac{\left[F(x_{(i)},\hat{\theta})\right]^{-1}-1}{(n-1/2)/(i-3/4)-1}\right\}\right]^2;$$

$$Z_K = \max_{1\le i\le n}\left(\left(i-\frac{1}{2}\right)\ln\left\{\frac{i-1/2}{nF(x_{(i)},\hat{\theta})}\right\}+\left(n-i+\frac{1}{2}\right)\ln\left[\frac{n-i+1/2}{n\left\{1-F(x_{(i)},\hat{\theta})\right\}}\right]\right).$$

The use of goodness-of-fit tests with these statistics complicates the strong dependence of the distributions of the statistics on the sample size n.

In verification of composite hypotheses when the parameters of an observed probability distribution law are estimated on the same sample, all of the nonparametric goodness-of-fit tests that are being considered here lose the property of being "distribution-free" [12]. Moreover, the limiting distributions of the statistics of nonparametric goodness-of-fit tests depend on a number of factors that define the "complexity" of the hypothesis.

The distribution law of the statistic $G(S|H_0)$ is influenced by the following factors [1]:

- form of the observed distribution law $F(x, \theta)$ corresponding to the true hypothesis H_0;
- type of estimated parameter and number of observed parameters;
- concrete value of parameter in certain situations, for example, in the case of gamma and beta distributions; and
- method used to estimate the parameters.

Questions related to verification of composite hypotheses with respect to different distribution laws and the construction of models of limiting distributions of the statistics of the goodnesss-of-fit tests enumerated above have been discussed in a number of studies [13–18]. Recommendations on the use of goodness-of-fit tests in verification of simple and composite hypotheses may be found in the applications handbook [1].

Using a goodness-of-fit test, the experimenter may propose that a particular test being used constitutes a tool that, in principle, may be applied to distinguish a law $F(x, \theta)$ corresponding to a verifiable hypothesis H_0 from (similar) competing distribution laws. Recall that two types of errors are associated with the verification of statistical hypotheses. Errors of the first kind consist in a deviation of a valid verifiable hypothesis H_0 and its probability is denoted α. An error of the second kind

TABLE 1. Power of Zhang Goodness-of-Fit Test z_A Relative to Hypothesis H_1

n	α				
	0.15	0.1	0.05	0.025	0.01
10	0.127	0.078	0.033	0.014	0.005
20	0.148	0.090	0.036	0.014	0.004
30	0.199	0.128	0.056	0.023	0.006
40	0.263	0.180	0.087	0.039	0.012
50	0.333	0.239	0.127	0.063	0.022
100	0.650	0.548	0.389	0.259	0.139
150	0.844	0.775	0.641	0.503	0.335
200	0.939	0.901	0.815	0.706	0.545
300	0.992	0.985	0.962	0.923	0.841

TABLE 2. Power of Zhang Goodness-of-Fit Test z_C Relative to Hypothesis H_1

n	α				
	0.15	0.1	0.05	0.025	0.01
10	0.163	0.101	0.041	0.017	0.004
20	0.211	0.130	0.049	0.014	0.002
30	0.277	0.179	0.071	0.020	0.002
40	0.348	0.238	0.104	0.033	0.003
50	0.421	0.300	0.142	0.049	0.005
100	0.715	0.599	0.390	0.201	0.045
150	0.879	0.806	0.635	0.420	0.150
200	0.955	0.917	0.808	0.634	0.322
300	0.995	0.988	0.961	0.895	0.688

consists in a nondeviation of H_0 in the case where some competing hypothesis H_1 is valid and its probability is denoted β. The experimenter's assumption as to the desired properties of a goodness-of-fit test includes the requirement that it be unbiased, i.e., that for any given probability of an error of the first kind α and any competing hypothesis H_1 (any competing law), the power of the test $1 - \beta$ relative to H_1 must obey the inequality $\alpha \le 1 - \beta$.

Examples that demonstrate the bias of nonparametric goodness-of-fit tests relative to certain competing hypotheses are not found in the literature. Apparently, most of tests used under practical conditions are asymptotically unbiased, and "flaws" appear in the case of limited sample sizes and relatively similar competing hypotheses.

In [19, 20], bias (in the case of limited sample sizes n and low α) of a whole series of special tests (Shapiro–Wilk, Epps–Palley, Hegasi–Green, Spigelhapter, Roystone) was found in a study of the properties of a set of tests oriented towards the verification of normality. A generalized normal law with density

$$f(x) = \frac{\theta_2}{2\theta_1 \Gamma(1/\theta_2)} \exp\left\{-\left(\frac{|x-\theta_0|}{\theta_1}\right)^{\theta_2}\right\}$$

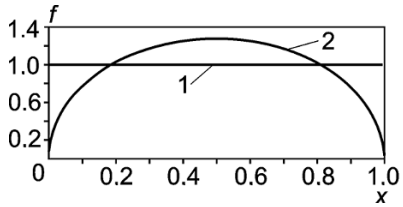

Fig. 2. Densities corresponding to *1*) uniform law (H_0) and *2*) beta distribution of the first kind (H_1).

and shape parameter $\theta_2 = 4$ is considered as the competing hypothesis H_1. The corresponding density of the normal and of the generalized normal law are represented in Fig. 1. The competing hypothesis H_1, which corresponds to a generalized normal law with shape parameter $\theta_2 = 4$, constitutes a "litmus paper" on which previously concealed drawbacks of the set of tests becomes apparent.

The nonparametric goodness-of-fit tests (Kolmogorov, Kuiper, Cramer–Mises–Smirnov, Watson, and Anderson–Darling tests) have been shown to be nonbiased relative to a given pair of hypotheses [21]. However, the Zhang goodness-of-fit tests with statistics Z_C and Z_A in verification of normality relative to the same competing hypothesis exhibit substantial bias [21]. Estimators of the power of the Zhang goodness-of-fit test with statistic Z_A with respect to a competing hypothesis H_1 are presented in Table 1, and with the statistic Z_C, in Table 2.

Estimators of power less than α are indicated by gray shading in Tables 1 and 2. Figuratively speaking, from the standpoint of a goodness-of-fit test, this means that the law corresponding to H_1 is "more normal than normal." Thus, a goodness-of-fit test with statistic Z_A for $n = 10$ or 20 cannot distinguish a law corresponding to the hypothesis H_1 from a normal law.

Note that in the general case, the Anderson–Darling, Watson, and Cramer–Mises–Smirnov goodness-of-fit tests (and even the Zhang goodness-of-fit test with statistics Z_A and Z_C) are not so greatly inferior in power (and not always) to special normality criteria.

It should be noted that the situation that has been described here as regards the bias of Zhang goodness-of-fit tests with statistics Z_A and Z_C in verification of normality is not the only example with such a drawback relating to the use of nonparametric goodness-of-fit tests. Another example has to do with the verification tests of (simple and composite) hypotheses on the membership of analyzed samples to a uniform law and the use for these purposes, in particular, of nonparametric goodness-of-fit tests.

In this case, a uniform law on the interval [0, 1] corresponds to a simple verifiable hypothesis H_0, while membership of an observable random variable, to a beta distribution of the first kind with density function

$$f(x) = \frac{1}{\theta_2 B(\theta_0, \theta_1)} \left(\frac{x - \theta_3}{\theta_2} \right)^{\theta_0 - 1} \left(1 - \frac{x - \theta_3}{\theta_2} \right)^{\theta_1 - 1},$$

where $B(\theta_2, \theta_1) = \Gamma(\theta_0)\,\Gamma(\theta_1)/\Gamma(\theta_0 + \theta_1)$ is a beta function with values of the parameters $\theta_0 = 1.5$, $\theta_1 = 1.5$, $\theta_2 = 1$, and $\theta_3 = 0$, corresponds to a competing hypothesis H_1.

The probability distribution functions corresponding to these hypotheses intersect, while the density distributions presented in Fig. 2 illustrate the essential difference between the competing laws.

From an investigation of the distributions of the statistics and an analysis of the power of goodness-of-fit tests relative to H_1 performed in [22], it may be established that the series of special tests oriented towards the verification of a hypothesis of uniformity (Moran, Sherman, Greenwood, Yang, Hegasi–Green, and others) are not capable of distinguishing this hypothesis from H_0 in the case of small sample sizes n and low significance levels α.

It turns out that most of the nonparametric goodness-of-fit tests considered in [22] (except for the Kuiper and Watson tests) also suffer from this drawback to a significant extent, the cause of which is seen in the bias of the corresponding tests.

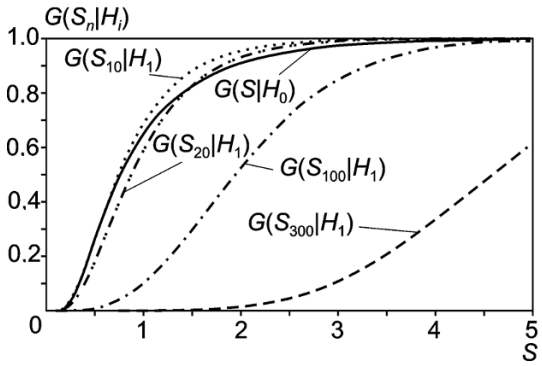

Fig. 3. Distribution $G(S|H_0)$ and $G(S_n|H_1)$ of the statistic of the Anderson–Darling goodness-of-fit test.

TABLE 3. Power of Anderson–Darling Goodness-of-Fit Test Relative to Hypothesis H_1

n	α				
	0.15	0.1	0.05	0.025	0.01
10	0.095	0.053	0.019	0.007	0.002
20	0.140	0.078	0.028	0.010	0.003
30	0.196	0.114	0.042	0.014	0.004
40	0.258	0.156	0.060	0.021	0.005
50	0.325	0.206	0.084	0.031	0.007
100	0.652	0.505	0.283	0.134	0.041
150	0.861	0.760	0.544	0.332	0.138
200	0.954	0.904	0.762	0.565	0.311
300	0.998	0.990	0.959	0.882	0.702

Estimators of power relative to the hypothesis H_1 are presented in Table 3 for the Anderson–Darling goodness-of-fit test used for verification of uniformity as an illustration of such a situation for a concrete example. Cells of the table with estimators of power less than the corresponding value of α are distinguished by gray shading. The distribution $G(S|H_0)$ of the statistic of this test where the verifiable hypothesis H_0 is valid and the distributions $G(S_n|H_1)$ of its statistic where H_1 is valid (for sample sizes $n = 10, 20, 100, 300$) are shown in Fig. 3. The distributions of the statistic $G(S_n|H_1)$ with $n = 10$ or 20 intersect $G(S|H_0)$, which explains why the power $1 - \beta$ proves to be less than the specified value of α. In Fig. 3, the distribution $G(S|H_0)$ is shown only for $n = 10$. With $n \geq 20$, the distributions $G(S_n|H_0)$ cannot be distinguished visually from $G(S_{10}|H_0)$ and practically coincide with the limiting distribution of the statistic of the Anderson–Darling goodness-of-fit test in verification of simple hypotheses $a2(s)$.

The examples that have been considered here demonstrate the bias of nonparametric goodness-of-fit tests in the case of small sample sizes n and low significance levels α relative to certain pairs of competing hypotheses. This is apparently the first mention of the existence of this type of drawback of nonparametric goodness-of-fit tests. Hence, it follows that the correct use of some test for the creation of a "reliable" statistical inference often may prove to be inadequate. The use of a set of tests that rely on different measures of deviation of an empirical distribution from the theoretical improves the quality of statistical inferences.

Through knowledge of the actual properties of tests and of the drawbacks of these tests that have been noted in the present article and observance of the recommendations of [1, 21, 22] that determine whether it is correct to use certain tests, specialists involved in the solution of problems of statistical analysis in the processing of the results of measurements in a particular application domain will be able to approach the selection of tests in a more rational manner without being concerned with the use of a particular test.

In conclusion, we would like to turn the reader's attention to the fact that, in our view, the use of methods of statistical analysis in problems of metrology and standardization lags considerably behind the level of development of modern applied mathematical statistics. This has to do both with the range of methods employed and their correct use, in particular, descriptions of the use of statistical tests in regulatory documents. For example, it would be useful to make corrections in the standard [23], where there are serious errors in the description of tests recommended for use in verification of normality. Of no lesser importance is the fact that the reasons for such states derive from the fact that out-of-date information and methods are published in textbooks in the series, *Standardization and Metrology* [24].

The present study was carried out with the support from the Ministry of Education of Russia within the framework of the design part of State Assignment No. 2.541.2014/K.

REFERENCES

1. B. Yu. Lemeshko, *Nonparametric Goodness-of-Fit Tests: Handbook on Applications*, NITs INFRA-M, Moscow (2014), DOI: 10.12737/11873.
2. L. N. Bol'shev and N. V. Smirnov, *Tables of Mathematical Statistics*, Nauka, Moscow (1983).
3. N. H. Kuiper, "Tests concerning random points on a circle," *Proc. Konikl. Nederl. Akad. Van Wettenschappen*, Ser. A, **63**, 38–47 (1960).
4. M. A. Stephens, "EDF statistics for goodness of fit and some comparisons," *J. Amer. Stat. Assoc.*, **69**, No. 347, 730–737 (1974).
5. B. Yu. Lemeshko and A. A. Gorbunova, "On the application and power of the Kuiper, Watson, and Zhang nonparametric goodness-of-fit tests," *Izmer. Tekhn.*, No. 5, 3–9 (2013).
6. G. S. Watson, "Goodness-of-fit tests on a circle," *Biometrika*, **48**, No.1–2, 109–114 (1961).
7. G. S. Watson, "Goodness-of-fit tests on a circle," *Biometrika*, **49**, No. 1–2, 57–63 (1962).
8. T. W. Anderson and D. A. Darling, "Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes," *Ann. Math. Stat.*, **23**, 193–212 (1952).
9. T. W. Anderson and D. A. Darling, "A test of goodness of fit," *J. Amer. Stat. Assoc.*, **29**, 765–769 (1954).
10. J. Zhang, *Powerful Goodness-of-fit and Multi-sample Tests: PhD Thesis*, York University, Toronto (2001).
11. J. Zhang, "Powerful goodness-of-fit tests based on the likelihood ratio," *J. Roy. Stat. Soc.: Ser.* B, **64**, No. 2, 281–294 (2002).
12. M. Kac, J. Kiefer, and J. Wolfowitz, "On tests of normality and other tests of goodness of fit based on distance methods," *Ann. Math. Stat.*, **26**, 189–211 (1955).
13. B. Yu. Lemeshko, S. B. Lemeshko, and S. N. Postovalov, "Statistic distribution models for some nonparametric goodness-of-fit tests in testing composite hypotheses," *Comm. Stat. Theory and Methods*, **39**, No. 3, 460–471 (2010).
14. B. Yu. Lemeshko and S. B. Lemeshko, "Models of statistic distributions of nonparametric goodness-of-fit tests in composite hypotheses testing for double exponential law cases," *Comm. Stat. Theory and Methods*, **40**, No. 16, 2879–2892 (2011).
15. B. Yu. Lemeshko and S. B. Lemeshko, "Models of distributions of statistics of nonparametric goodness-of-fit tests in verification of composite hypotheses with the use of maximum likelihood estimators. Part I," *Izmer. Tekhn.*, No. 6, 3–11 (2009).
16. B. Yu. Lemeshko and S. B. Lemeshko, "Models of distributions of statistics of nonparametric goodness-of-fit tests in verification of composite hypotheses with the use of maximum likelihood estimators. Part II," *Izmer. Tekhn.*, No. 8, 17–26 (2009).

17. B. Yu. Lemeshko, A. A. Gorbunova, S. B. Lemeshko, and A. R. Rogozhnikov, "Solving problems of using some nonparametric goodness-of-fit tests," *Optoelectr., Instrum. Data Proces.*, **50**, 21–35 (2014).
18. B. Yu. Lemeshko and A. A. Gorbunova, "Use of Kuiper and Watson nonparametric goodness-of-fit tests in verification of composite hypotheses," *Izmer. Tekhn.*, No. 9, 14–21 (2013).
19. B. Yu. Lemeshko and S. B. Lemeshko, "A comparative analysis of tests for verification of deviation of a distribution from a normal law," *Metrologiya*, No. 2, 3–24 (2005).
20. B. Yu. Lemeshko and A. P. Rogozhnikov, "Investigation of features and power of certain normality tests," *Metrologiya*, No. 4, 3–24 (2009).
21. B. Yu. Lemeshko, *Tests for Verification of Deviation of a Distribution from a Normal Law: Handbook on Applications*, INFRA-M, Moscow (2015), DOI: 10.12737/6086.
22. B. Yu. Lemeshko, P. Yu. Blinov, *Tests for Verification of Deviation of a Distribution from a Normal Law: Handbook on Applications*, INFRA-M, Moscow (2015), DOI: 10.12737/11304.
23. GOST R 8.736–2011, *Direct Repeated Measurements. Methods of Processing the Results of Measurements. Basic Assumptions*.
24. S. S. Antsyferov, M. S. Afanas'ev, and K. E. Rusanov, *Processing the Results of Measurements: Teach. Aid*, Izd. IKAR, Moscow (2014).