Models of Statistic Distributions of Nonparametric Goodness-of-fit Tests in Composite Hypotheses Testing in Case of Double Exponential Law

Boris Yu. Lemeshko\textsuperscript{1} and Stanislav B. Lemeshko\textsuperscript{1}

\textsuperscript{1}Department of Applied Mathematics, Novosibirsk State Technical University, K.Marx pr., 26, Novosibirsk, Russia.
E-mail: Lemeshko@fpm.ami.nstu.ru

Abstract: In this paper are presented more precise results (tables of percentage points and statistic distribution models) for the nonparametric goodness-of-fit tests in testing composite hypotheses using the maximum likelihood estimate (MLE) for double exponential distribution law. Statistic distributions of the nonparametric goodness-of-fit tests are investigated by the methods of statistical simulation.

Keywords: goodness-of-fit test, composite hypotheses testing Kolmogorov test, Cramer-Mises-Smirnov test, Anderson-Darling test, double exponential distribution.

1 Introduction

In composite hypotheses testing of the form \( H_0 : F(x) \in \{ F(x, \theta), \theta \in \Theta \} \), when the estimate \( \hat{\Theta} \) of the scalar or vector distribution parameter \( F(x, \theta) \) is calculated by the same sample, the nonparametric goodness-of-fit Kolmogorov, Cramer-Mises-Smirnov, Anderson-Darling tests lose the free distribution property.

The value
\[
D_n = \sup_{x \in \Theta} |F_n(x) - F(x, \theta)|,
\]
where \( F_n(x) \) is the empirical distribution function, \( n \) is the sample size, is used in Kolmogorov test as a distance between the empirical and theoretical laws. In testing hypotheses, a statistic with Bolshev correction (Bolshev, 1987) of the form (Bolshev and Smirnov, 1983)
\[
S_K = \frac{6nD_n + 1}{6\sqrt{n}},
\]
where \( D_n = \max(D_n^+, D_n^-) \),
\[
D_n^+ = \max_{i=1}^{n} \left\{ \frac{i}{n} - F(x_i, \theta) \right\}, \quad D_n^- = \max_{i=1}^{n} \left\{ F(x_i, \theta) - \frac{i-1}{n} \right\},
\]
\( n \) is the sample size, \( x_1, x_2, \ldots, x_n \) are sample values in increasing order is usually used. The distribution of statistic (1) in testing simple hypotheses obeys the Kolmogorov distribution law \( K(S) \) (Bolshev and Smirnov, 1983).

In \( \omega^2 \) Cramer-Mises-Smirnov test, one uses a statistic of the form
\[
S_\omega = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left( \frac{F(x_i, \theta) - \frac{2i-1}{2n}}{2n} \right)^2.
\]
and in test of $\Omega^2$ Anderson-Darling type (Anderson and Darling, 1952, 1954), the statistic of the form

\[
S_\Omega = -n - 2 \sum_{i=1}^{n} \left\{ \frac{2i-1}{2n} \ln F(x_i, \theta) + \left( 1 - \frac{2i-1}{2n} \right) \ln(1 - F(x_i, \theta)) \right\}. \tag{3}
\]

In testing a simple hypothesis, statistic (2) obeys the distribution $a_1(S)$ and statistic (3) obeys the distribution $a_2(S)$ (see Bolshev and Smirnov, 1983).

In composite hypotheses testing, the conditional distribution law of the statistic $G(S|H_0)$ is affected by a number of factors: 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 specific value of the parameter (e.g., in the case of gamma-distribution and beta-distribution families); the method of parameter estimation. The distinctions in the limiting distributions of the same statistics in testing simple and composite hypotheses are so significant that we cannot neglect them.

The paper (Kac et al., 1955) was a pioneer in investigating statistic distributions of the nonparametric goodness-of-fit tests with composite hypotheses. Then, for the solution to this problem, various approaches where used (Darling, 1955, 1957), (Durbin, 1973, 1975), (Martinov, 1978), (Pearson and Hartley, 1972), (Stephens, 1970, 1974), (Chandra et al., 1981), (Tyurin, 1984), (Dzhaparidze and Nikulin, 1982), (Nikulin, 1992).

In our research (Lemeshko and Postovalov, 1998, 2001a, 2001b, 2002), (Lemeshko and Maklakov, 2004), (Lemeshko, 2004), statistic distributions of the nonparametric goodness-of-fit tests are investigated by the methods of statistical simulating, and for constructed empirical distributions approximate models of law are found. The results obtained were used to develop recommendations for standardization (R 50.1.037-2002, 2002).

## 2 Statistic distributions of the tests in testing composite hypotheses concerning double exponential law

In testing composite hypotheses for distribution law with density

\[
f(x, \theta) = \frac{\theta_0}{2\theta_0 \Gamma(1/\theta_0)} \exp\left\{-\left(\frac{|x - \theta_0|}{\theta_0}\right)^{\theta_0}\right\}, \tag{4}
\]

distributions $G(S|H_0)$ of non-parametric goodness-of-fit tests statistics depend on specific value of shape parameter $\theta_0$.

The family (4) defines a set of symmetric laws, special cases of which are normal distribution ($\theta_0 = 2$) and Laplace distribution ($\theta_0 = 1$). Sometimes this distribution is called double-sided exponential, although usually $\theta_0 = 1$ is implied.

The feature in behavior of non-parametric goodness-of-fit tests statistics $G(S|H_0)$ $S_\theta$ when testing composite hypotheses for family (4) is that with
shape parameter growing up to $\theta_0 \approx 1.64$, distributions $G(S|H_0)$ are shifting to the right, and with the following growth, the shift starts in the opposite direction (see Fig.1).

Upper percentage points and the models of limiting statistic distributions of Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests have been constructed for the values of shape parameter $\theta_0 = 0.5, 0.75, 1, 1.6, 2, 3, 4, 5, 7$ when MLEs were used. The results for the shape parameter $\theta_0 = 1.6$ are presented as an example in table 1. These results define more accurately supplemented results that were presented in (Lemeshko et al., 2004). If the value of form parameter $\theta_0$ is not congruent with tabular, interpolation could be used to obtain approximate percentage points.

Distributions $G(S|H_0)$ of the Kolmogorov, Cramer-Mises-Smirnov and the Anderson-Darling statistics are best approximated by the family of the III type beta-distributions with the density function

$$B_3(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4) = \frac{\theta_0^{\theta_0}}{\theta_1 B(\theta_0, \theta_1)} \left( \frac{x-\theta_4}{\theta_3} \right)^{\theta_0-1} \left( \frac{1-x-\theta_4}{\theta_3} \right)^{\theta_1-1} \frac{1}{1+(\theta_2-1)\frac{x-\theta_4}{\theta_3}} \gamma(\theta_0, \theta_1),$$

by gamma-distributions family $\gamma(\theta_0, \theta_1, \theta_2) = \frac{1}{\theta_0^{\theta_0}} \frac{1}{\theta_1} \Gamma(\theta_0) x^{-\theta_0-1} e^{-x/\theta_0}.$

by the family of the $Sb$-Johnson distributions

$$Sb(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4) = \frac{\theta_0 \theta_2}{(x-\theta_1)(\theta_2 + \theta_3 - x)} \exp \left\{ -\frac{1}{2} \ln \frac{x-\theta_3}{\theta_2 + \theta_3 - x} \right\}.$$
or by the family of the SI-Johnson distributions

$$SI(\theta_0, \theta_1, \theta_2, \theta_3) = \frac{\theta_1}{x-\theta_3} \sqrt{2\pi} \exp \left\{ -\frac{1}{2} \left[ \theta_0 + \theta_1 \ln \frac{x-\theta_3}{\theta_2} \right]^2 \right\}.$$ 

The tables of percentage points and statistic distributions models were constructed by modeled statistic samples with the size $N = 10^6$ ($N$ is the number of runs in simulation). This number ensures the deviation of the empirical p.d.f. $G_n(S \mid H_0)$ from the theoretical (true) to be less than $10^{-3}$. In this case, the samples of pseudorandom variables, belonging to $F(x, \theta)$, were generated with the size $n = 10^3$. For such value of $n$ statistic p.d.f. $G(S \mid H_0)$ almost coincides with the limit p.d.f. $G(S \mid H_0)$.

**Table 1.** Upper percentage points and models of limiting statistic distributions of the nonparametric goodness-of-fit test when MLE are used (for $\theta_0 = 1.6$)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Percentage points</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0.9$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 0.95$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 0.99$</td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>1.216</td>
<td>B$(4.2366, 5.7254, 2.8969, 2.4200, 0.330)$</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>1.185</td>
<td>B$(4.3698, 5.2853, 3.3545, 2.3863, 0.318)$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.851</td>
<td>B$(5.4129, 7.6381, 2.1289, 1.3936, 0.290)$</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>1.141</td>
<td>B$(4.9730, 5.5743, 4.6422, 2.3576, 0.29)$</td>
</tr>
<tr>
<td>$\theta_4$, $\theta_5$</td>
<td>0.828</td>
<td>B$(6.2506, 7.4916, 2.5914, 1.4130, 0.275)$</td>
</tr>
<tr>
<td>$\theta_6$, $\theta_7$</td>
<td>0.770</td>
<td>B$(5.3623, 7.3149, 2.1379, 1.0542, 0.0393)$</td>
</tr>
<tr>
<td>$\theta_8$, $\theta_9$, $\theta_{10}$</td>
<td>0.704</td>
<td>B$(7.4853, 7.2752, 3.2095, 1.557, 0.260)$</td>
</tr>
</tbody>
</table>

for Kolmogorov’s test

<table>
<thead>
<tr>
<th></th>
<th>$p = 0.9$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0.95$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 0.99$</td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.339</td>
<td>Sb$(3.6139, 1.0337, 3.4000, 0.013)$</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.325</td>
<td>Sb$(2.7348, 0.9148, 1.8000, 0.016)$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.121</td>
<td>B$(4.5239, 3.7332, 15.6889, 0.6596, 0.009)$</td>
</tr>
<tr>
<td>$\theta_3$, $\theta_4$</td>
<td>0.314</td>
<td>Sb$(3.2111, 0.8115, 1.3500, 0.016)$</td>
</tr>
<tr>
<td>$\theta_5$, $\theta_6$</td>
<td>0.109</td>
<td>B$(4.2190, 3.9949, 12.6139, 0.5642, 0.0087)$</td>
</tr>
<tr>
<td>$\theta_7$, $\theta_8$</td>
<td>0.087</td>
<td>B$(4.5491, 4.8658, 9.0448, 0.4000, 0.008)$</td>
</tr>
<tr>
<td>$\theta_9$, $\theta_{10}$, $\theta_{11}$</td>
<td>0.069</td>
<td>B$(6.8750, 4.6392, 18.0200, 0.3937, 0.006)$</td>
</tr>
</tbody>
</table>

for Cramer-Mises-Smirnov’s test

<table>
<thead>
<tr>
<th></th>
<th>$p = 0.9$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 0.95$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p = 0.99$</td>
<td></td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>1.819</td>
<td>B$(3.7982, 2.4042, 26.2612, 10.000, 0.095)$</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>1.735</td>
<td>B$(3.6908, 2.1990, 32.1310, 10.000, 0.10)$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.864</td>
<td>B$(4.0782, 5.1594, 17.0570, 7.9000, 0.09)$</td>
</tr>
<tr>
<td>$\theta_3$, $\theta_4$</td>
<td>1.669</td>
<td>B$(4.6625, 1.4267, 33.5120, 4.5000, 0.09)$</td>
</tr>
<tr>
<td>$\theta_5$, $\theta_6$</td>
<td>0.716</td>
<td>B$(4.5576, 4.2326, 10.9573, 3.23142, 0.08)$</td>
</tr>
<tr>
<td>$\theta_7$, $\theta_8$</td>
<td>0.589</td>
<td>B$(4.5825, 5.3012, 7.9243, 2.5555, 0.0775)$</td>
</tr>
<tr>
<td>$\theta_9$, $\theta_{10}$, $\theta_{11}$</td>
<td>0.492</td>
<td>B$(5.08840, 5.2459, 10.6760, 2.4738, 0.068)$</td>
</tr>
</tbody>
</table>
3 Conclusions

In this work are presented more precise models of the statistic distributions of the nonparametric goodness-of-fit tests for testing composite hypotheses with the distributions family (4).

It should be stressed, that obtained percentage points and models guarantee proper implementation of the nonparametric goodness-of-fit tests in statistic analysis problems if MLE is used. These results can’t be used with other estimations because statistic distributions of these tests are essential depend on estimation method (Lemeshko et al., 2001). In the case of the I, II, III type beta-distribution families’ statistic distributions depend on a specific value of two form parameter of these distributions. Statistic distributions models and tables of percentage points for various combinations of values of two form parameters (more than 1500 models) were constructed in the thesis of Lemeshko S.B. and partly were published in the paper (Lemeshko et al., 2007).

Note that, in composite hypotheses testing, power of the nonparametric goodness-of-fit tests, generally, essentially higher (if MLE is used), than in simple hypotheses testing.

The results of comparative analysis of goodness-of-fit tests power (nonparametric and $\chi^2$ type) subject to some sufficiently close pair of alternative are presented in (Lemeshko et al., 2007), and are in more detail stated in (Lemeshko et al., 2008a, 2008b).

The authors hope that release of the article will be conductive to decrease mistake amount, committed in statistic analysis problems if nonparametric goodness-of-fit tests are used (Lemeshko, 2004).

This research was supported by Federal Education Agency of Russian Federation Ministry of Education in the Analytical departmental purposeful program framework ”Development of Higher School Potential”, (project № 2.1.2/3970).

References


