NONPARAMETRIC GOODNESS-OF-FIT TESTS
WHEN CHECKING COMPOSITE HYPOTHESES

Lemeshko B.Yu., Postovalov S.N.
Novosibirsk State Technical University,
Novosibirsk, Russia, e-mail: headrd@fpm.ami.nstu.ru

Abstract. Using methods of computer simulation it has been shown that when checking
composite hypotheses the distributions of statistics of nonparametric goodness-of-fit tests depend
not only on the law \( F(x, \theta) \), corresponding to hypothesis \( H_0 \), but also on the method of estimating
the parameter \( \theta \). It has been shown that tests have a maximum power when using maximum
likelihood method.

With usage of goodness-of-fit tests the simple hypotheses of the form \( H_0: F(x) = F(x, \theta) \),
where \( F(x, \theta) \) - the cumulative distribution function, with which is checked goodness-of-fit of
observed sample, \( \theta \) - known value of the parameter (scalar or vectorial), and composite hypotheses
\( H_0: F(x) = F(x, \hat{\theta}) \), where \( \hat{\theta} \) - the estimate of the parameter calculated from the same sample (if
an estimate \( \hat{\theta} \) is calculated from other sample, a hypothesis is simple) can be checked. During
check from the sample the value \( S^* \) of used criterion statistics is calculated. Further, to make
conclusion about that, to accept or to reject a hypothesis \( H_0 \), it is necessary to know a conditional
distribution \( G(S|H_0) \) of statistics \( S \) when \( H_0 \) is valid. And if probability
\[
P\{S > S^*\} = \int_{S^*}^{\infty} g(s|H_0)ds
\]
is large enough, at least \( P\{S > S^*\} > \alpha \), where \( g(s|H_0) \) - conditional density, and \( \alpha \) - given
significance level (probability of an error of first kind - to reject a true hypothesis \( H_0 \)), there is no
reason to reject a hypothesis \( H_0 \). If during the analysis of sample the alternative \( H_1 \) is considered:
\( F(x) = F_1(x, \theta) \), it is connected with conditional distribution \( G(S|H_1) \) and probability of an error
of second kind \( \beta \) (to accept a hypothesis \( H_0 \), while the hypothesis \( H_1 \) is correct). The
representation of \( \alpha \) for a used goodness-of-fit test also uniquely determines \( \beta \):
\[
\alpha = \int_{S_a}^{\infty} g(s|H_0)ds, \quad \beta = \int_{0}^{S_{a}} g(s|H_1)ds.
\]
In this case the more is the power of criterion \( 1 - \beta \), the better it distinguishes appropriate
hypotheses.

Nonparametric goodness-of-fit tests (Kolmogorov’s–Smirnov’s test, Cramer-von-Mises tests) have
property of "freedom from distribution". But when distribution's parameters are estimated those
tests lose that property. First, dependence of statistics distributions on the law's form was noted in
[1]. Since then many publications were devoted to this problem. However solution was received
only in particular situations.

In papers [2, 3] with a help of computer analysis of statistics’ distributions \( G(S|H_0) \), where
\( S \) is corresponding statistics, \( H_0: F(x) = F(x, \hat{\theta}) \) is basic hypothesis, \( \hat{\theta} \) is parameter estimate,
were constructed simple and sufficiently exact approximations for \( G(S|H_0) \) with the different laws of random variable \( F(x, \theta) \) and maximum likelihood method of parameters estimation.

Further studies have shown that distributions of nonparametric tests' statistics when checking composite hypotheses essentially depend not only on the laws form \( F(x, \theta) \), on the estimated parameter's find, on the number of estimated parameters, on the specific parameter's values, but also on the method of estimating.

In the present moment the problems of application of nonparametric goodness-of-fit tests when checking composite hypotheses in connection with distinction in methods of estimation are not reflected in the scientific publications. Russian and international standards of statistical method’s applications don't contain recommendations to application of nonparametric tests when checking composite hypotheses. Therefore application of those tests in practice often leads to incorrect conclusions.

The distributions of statistics of nonparametric goodness-of-fit tests when checking composite hypotheses depend on character of this composite hypothesis. The law of distribution of a statistic \( G(S|H_0) \) is influenced by a lot of factors, determining "complexity" of a hypothesis: the form of the observed law of distribution \( F(x, \theta) \), appropriate to a true hypothesis \( H_0 \); the type of the estimated parameter and the amount of estimated parameters of the law \( F(x, \theta) \), and in some situations a specific value of the parameter, as in case of gamma - distribution; a used method of parameter’s estimation. For small sample sizes \( n \) the distribution \( G(S_n|H_0) \) depends also on from \( n \). Though truth essential dependenc from \( n \) is observed only for small sample sizes.

Using methods of statistical simulation, we researched how the size of observed sample tells on distributions of statistics of nonparametric goodness-of-fit tests for simple and composite hypotheses. As things turned out already for \( n \geq 15 \div 20 \) the distribution \( G(S_n|H_0) \) is close enough to limiting \( G(S|H_0) \), and the dependence from \( n \) can be neglected.

In case of giving a specific alternative (competing hypothesis \( H_1 \), to which distribution \( F_1(x, \theta) \) corresponds) cumulative distribution function of a statistic \( G(S|H_1) \) also depends on all enumerated factors. Moreover as distinct from \( G(S|H_0) \), the distribution of a statistic \( G(S|H_1) \) with a true hypothesis \( H_1 \) very strongly depends on the sample size \( n \) : due to this fact the ability of criteria to distinguish close hypotheses increase with growth of \( n \) the power of criteria increase.

In present work applying the methods of computer simulation we researched distributions \( G(S|H_0) \) and \( G(S|H_1) \) of statistics of Kolmogorov’s test and \( \omega_2^2 \) of Mises in case of usage of various methods of estimation for various laws \( F(x, \theta) \), most frequently used in applications.

In criteria such as Kolmogorov’s test (Kolmogorov-Smirnov’s test) most frequently is used the following statistic [4]

\[
S_k = \frac{6nD_n + 1}{6\sqrt{n}},
\]

where \( D_n = \max(D_n^+, D_n^-), D_n^+ = \max_{i \leq n} \left\{ \frac{i}{n} - F(x_i, \theta) \right\}, D_n^- = \max_{i \leq n} \left\{ F(x_i, \theta) - \frac{i-1}{n} \right\}, n \) - sample size, \( x_1, x_2, \ldots, x_n \) - sample values, ordered by increase, \( F(x) \) - function of the law of distribution,
goodness-of-fit with which is checked. The distribution of variable $S_K$ for simple hypothesis in a limit obeys the Kolmogorov’s law $K(S)$ [4].

Statistic $\omega^2$ of Mises’s test (Cramer-von-Mises-Smirnov test)

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

with a simple hypothesis obeys the distribution $aI(S)$ [4].

Distributions of statistics in this case were compared when using of maximum likelihood estimates (MLE) and MD-estimations. When evaluation of MD-estimations appropriate distance between empirical and theoretical distributions is minimized. When using a statistic $S_K$ as an estimate of the parameter vector $\theta$ the values minimizing this statistic are selected: $\hat{\theta} = \arg \min_0 S_K$ (MD-estimation $S_K$). Similarly, when using a statistic $S_\omega$ it (MD-estimate $S_\omega$) is minimized on $\theta$: $\hat{\theta} = \arg \min_0 S_\omega$.

In fig. 1 the distributions $G(S_K|H_0)$ of Kolmogorov’s statistics $S_K$ are represented when checking a composite hypothesis with usage of maximum likelihood method for estimation of two parameters of the law, corresponding to a hypothesis $H_0$ (1 - normal with density $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$), 2 - logistics $f(x) = \frac{\pi}{\sqrt{3}} \exp\left\{ -\frac{\pi(x-\mu)}{\sqrt{3}} \right\}/\left[1 + \exp\left\{ -\frac{\pi(x-\mu)}{\sqrt{3}} \right\}\right]^2$, 3 - Laplace $f(x) = \frac{\theta_0}{2} e^{-|x-\theta_0|}$, 4 - least value $f(x) = \frac{1}{\theta_0} \exp\left\{ \frac{x-\theta_1}{\theta_0} - \exp\left\{ \frac{x-\theta_1}{\theta_0} \right\} \right\}$, 5 - Cauchy with density $f(x) = \frac{\theta_0}{\pi[\theta_0^2 + (x-\theta)^2]}$. In fig. 2 the distributions $G(S_K|H_0)$ of the same statistic $S_K$ are represented when checking the same hypotheses, but with usage of MD-estimates $S_K$.

In fig. 3 the distributions of a statistic $S_\omega$ for similar hypotheses $H_0$ are presented while using MLE, and in fig. 4 - while using MD-estimates minimizing a statistic $S_\omega$ on parameters.

When using of MD-estimates minimizing a statistic of criterion, the empirical distributions $G(S_\omega|H_0)$, corresponding to various hypotheses $H_0$, have a minimum scatter, that allows to speak about a certain “freedom from distribution” of criteria. If we base only on this fact, it would seem, that only such methods of estimation should be applied when checking composite hypotheses.
However research of a power of considered criteria for various methods of estimation has shown, that the nonparametric criteria with close alternatives have maximum power in case of a parameter estimation by maximum likelihood method. In fig. 5 distribution $G(S|H_0)$ and distribution $G(S|H_1)$ of the statistics $S_ω$ of Cramer-von-Mises-Smirnov’s test are represented depending on $n$ for composite hypothesis ($H_0$ - normal distribution, $H_1$ - logistic) with $n = 20, 100, 500, 1000$, when using MLE, and in fig. 6 - when using of MD-estimates $S_ω$.

In the work the distributions $G(S|H_0)$ and $G(S|H_1)$ for various hypotheses $H_0$ and various close $H_1$ were researched depending on sample size. On the grounds of the obtained results development of “Rules for checking goodness-of-fit of experienced distribution with theoretical one” when checking composite hypotheses with nonparametric tests becomes possible.

References