Some recent characterization based goodness of fit tests

Bojana Milošević[*1]

[1] *University of Belgrade, Faculty of Mathematics*

In this paper some recent advances in goodness of fit testing are presented. Special attention is given to goodness of fit tests based on equidistribution and independence characterizations. New concepts are described through some modern exponentiality tests. Their natural generalizations are also proposed. All tests are compared in Bahadur sense.

Keywords: asymptotic efficiency, order statistics, independence,V-statistic

1 Introduction

Goodness of fit testing occupy a significant part of nonparametric statistic. Most of classical tests are based on distance between the assumed distribution function (d.f.) and its consistent estimate, empirical d.f. Symmetry tests are analogously constructed. A new approach, that is especially attractive in recent years, is making tests based on characterizations of different types. Those tests mostly use U-empirical d.f.'s, U-empirical integral transforms, U-empirical moments etc. The main advantage of these tests is that they are often free of some distribution parameters. Therefore they are suitable for testing composite hypothesis. In addition, there is an abundance of characterization theorems for some families of distributions, in particular for exponential and other life distributions, uniform, normal distribution, and characterizations of the family of symmetric distributions around zero. An extensive survey is given in classical monograph by Galambos and Kotz (see [4]) as well as in recent monograph by Ahsanullah (see [1]). Hence, many different modern goodness of fit tests (GOF) can be constructed (see [17]).

For purpose of comparison of tests, the Bahadur efficiency has become very popular. One of the reasons is that, unlike Pitman efficiency, it does not require the asymptotic normality of test statistics. For more details we refer to [14].

*Corresponding author: bojana@matf.bg.ac.rs

Bahadur efficiency can be expressed as the ratio of the Bahadur exact slope $c(\theta)$, a function describing the rate of exponential decrease for the attained level under the alternative, and $2K(\theta)$, the double Kullback-Leibler distance between the alternative and the set of null distributions. Under closed alternatives we use the *local Bahadur efficiency* given by

$$e = \lim_{\theta \to 0} \frac{c(\theta)}{2K(\theta)}.$$

According to Bahadur's theory slopes can be calculated in the following way. Suppose that, under an alternative indexed by a parameter θ, the sequence T_n converges in probability to some finite function $b(\theta)$. Suppose also that the large deviation limit

$$\lim_{n \to \infty} n^{-1} P_{H_0}\{T_n > t\} = -f(t)$$

exists for any t in an open interval I, on which f is continuous and $\{b(\theta), \theta > 0\} \subset I$. Then the Bahadur exact slope is

$$c(\theta) = 2f(b(\theta)).$$

Very often, the main obstacle is to find large deviation function. Crucial results concerning this problem are given in [15], [16] and [10].

In the following section we describe two types of characterization theorems and show how, using them, we can construct an integral and a Kolmogorov type statistic. In order to illustrate the general idea we provide two examples.

2 Characterizations and tests statistics

In this section we focus on so called "equidistribution based" and "independence based" characterizations and GOF tests based on them.

The first group contains the characterization of the following form.

Let $X_1, ..., X_{\max(m,p)}$ be i.i.d with d.f. F, $\omega_1 : R^m \mapsto R^1$ and $\omega_2 : R^p \mapsto R^1$ two sample functions such. Then the following equivalence hold

$$\omega_1(X_1, ..., X_m) \overset{d}{=} \omega_2(X_1, ..., X_p)$$

if and only if F belongs to some family $\mathcal{F}_0$.

Natural estimators of d.f.'s of ω_1 and ω_2 are V-empirical d.f.'s given by

$$G_{n1}(x) = \frac{1}{n^m} \sum_{1 \le i_1, ..., i_m \le n} I\{\omega_1(X_{i_1}, ..., X_{i_m}) < x\}$$

$$G_{n2}(x) = \frac{1}{n^p} \sum_{1 \le i_1, ..., i_p \le n} I\{\omega_2(X_{i_1}, ..., X_{i_p}) < x\}.$$

Alternatively, one can consider corresponding symmetrized U-empirical d.f.'s.

Therefore GOF tests can be constructed based on the difference between functions G_{n1} and G_{n2}. Mostly used in last few years (see e.g [18],[11],[8], etc.) are those of integral type

$$I_n = \int_{-\infty}^{\infty} (G_{n1}(x) - G_{n2}(x)) dF_n(x),$$

as well as Kolmogorov type statistic

$$K_n = \sup_x |G_{n1}(x) - G_{n2}(x)|.$$

Usually large values of statistics are significant. Under some additional conditions both statistics are often free of some distribution parameter. For example, in case of testing exponentiality sufficient condition is that ω_1 and ω_2 are homogenuous functions of sample elements.

The group of independence based characterization contains the characterizations of the following form.

Let $X_1, ..., X_m$ be i.i.d with d.f. F, $\omega_1 : R^m \mapsto R$ and $\omega_2 : R^p \mapsto R$ two sample function such that $p \leq m$. Then the following equivalence hold

$$\omega_1(X_1, ..., X_m) \textit{ and } \omega_1(X_1, ..., X_p) \textit{ are independent}$$

if and only if F belongs to some family $\mathcal{F}_0$.

Thus we may reformulate our null hypothesis into

$$H_0 : \quad H(x_1, x_2) = G_1(x_1)G(x_2), \quad \text{for all } x_1, x_2 \in R$$

where G_1, G_2 i H are marginal and joint d.f.'s of ω_1 and ω_2, respectively. Natural choice for test statistics is

$$\begin{aligned} I_n &= \int_{x_1,x_2} (G_n(x_1, x_2) - H_n(x_1, x_2)) dF_n(x_1) dF_n(x_2) \\ K_n &= \sup_{x_1,x_2} |G_n(x_1, x_2) - H_n(x_1, x_2)|, \end{aligned} \tag{1}$$

where

$$\begin{aligned} G_n(x_1, x_2) &= G_{n1}(x_1)G_{n2}(x_2) \\ H_n(x_1, x_2) &= \frac{1}{n^m} \sum_{1 \leq i_1, \ldots, i_m \leq n} I\{\omega_1(X_{i_1}, ..., X_{i_m}) < x_1, \omega_2(X_{i_1}, ..., X_{i_p}) < x_2\} \end{aligned}$$

are suitable V-empirical d.f.'s. As previously, large values of proposed statistics are significant. Those type of GOF tests have been firstly considered by Milošević and Obradović in [10].

Notice that all integral type statistics are U-statistics, V-statistics or hybrid U-statistic. It turns up that in most cases their kernels are bounded and non-degenerate. Hence, the limiting distributions of these statistics, appropriately normalized, are normal. In case of Kolmogorov type statistics we have supremum over some family of U-statistics, usually non-degenerate in the sense of [10, 16]. Therefore their limiting distributions, appropriately normalized, coincide with that of a supremum of of absolute value of some centered Gaussian process (field). The critical values can be found using Monte Carlo simulations.

3 Examples and discussion

Recently Milošević and Obradović proved the following characterization theorem (see [6]). The theorem generalizes results from [12] based on original ideas from [2] and [19].

Let $X_1,\dots,X_m$ be a random sample from the distribution whose density $f(x)$ has the Maclaurin expansion for $x>0$, and let X_0 be a random variable independent of the sample that follows the same distribution. Let k be a fixed number such that $1<k\le m$. X is exponentially distributed if and only if one of the following three statement holds:

$$X_{(k;m)} \overset{d}{=} X_{(k-1;m-1)} + \frac{1}{m}X_m \tag{2}$$

$$X_{(k;m)} \overset{d}{=} X_{(k-1;m)} + \frac{1}{m-k+1}X_0 \tag{3}$$

$$X_{(k;m)} \overset{d}{=} \frac{1}{m}X_1 + \frac{1}{m-1}X_2 + \cdots + \frac{1}{m-k+1}X_k \tag{4}$$

In the spirit of the previous section many GOF tests based on this characterization theorem can be constructed. Denote with $I_k^{(j)}$ integral type, and with $K_k^{(j)}$ $(j=1,2,3)$ Kolmogorov type tests based on j-th part of the theorem for $m=k$.

The tests $I_3^{(1)}$ and $K_3^{(1)}$ were proposed in [18], $I_2^{(2)}$ and $K_2^{(2)}$ in [7], $I_3^{(2)}$ and $K_3^{(2)}$ in [9], while the test $I_k^{(3)}$, $K_2^{(3)}$ and $K_3^{(3)}$ were proposed in [5]. Notice that the $I_2^{(1)}$ and $K_2^{(1)}$ coincide with $I_2^{(3)}$ and $K_2^{(3)}$, respectively. We propose $I_k^{(1)}$ and $I_k^{(2)}$ for arbitrary k and show that they are asymptotically equivalent to

U-statistics with nondegenerate symmetric kernel. We derive their asymptotic d.f.'s as well as large deviation functions.

We compare tests against some closed alternatives, namely Weibull, Makeham, Gamma, mixture of exponential distributions with negative weights (EMNW(3)) and linear failure rate (LFR) distribution. Their densities can be found e.g. in [9, 10]. The results are summarized in Tables 1 and 2.

Table 1: Bahadur efficiencies of integral type tests

	$j=1$			$j=2$			$j=3$		$I^{\mathcal{E}}$
Alt.	$e_2^{(1)}$	$e_3^{(1)}$	$\max_k e_k^{(1)}, k$	$e_2^{(2)}$	$e_3^{(2)}$	$\max_k e_k^{(2)}, k$	$e_3^{(3)}$	$\max_k e_k^{(3)}, k$	$e^{\mathcal{E}}$
Weibull	.621	.649	(.649, 3)	.750	.746	(.750, 2)	.664	(.710, 8)	.419
Makeham	.488	.654	(.783, 6)	.625	.772	(.872, 6)	.573	(.876, 14)	.714
Gamma	.723	.638	(.723, 2)	.796	.701	(.796, 2)	.708	(.723, 2)	.701
EMNW(3)	.694	.835	(.835, 3)	.844	.916	(.916, 3)	.799	(.885, 6)	.542
LFR	.104	.206	(.613, 20)	.208	.308	(.712, 23)	.159	(.804, 88)	.535

Table 2: Bahadur efficiencies of Kolmogorov type tests

	$j=1$		$j=2$		$j=3$	$K^{\mathcal{E}}$
Alt.	$e_2^{(1)}$	$e_3^{(1)}$	$e_2^{(2)}$	$e_3^{(2)}$	$e_3^{(3)}$	$e^{\mathcal{E}}$
Weibull	.092	.079	.277	.258	.152	.200
Makeham	.125	.123	.342	.370	.216	.375
Gamma	.093	.066	.267	.212	.138	.131
EMNW(3)	.149	.122	.396	.364	.230	.334
LFR	.052	.067	.155	.213	.106	.235

Now we pass to the example of GOF tests via independence based characterization. Fisz in [3] proved following theorem.

Let X and Y be i.i.d. random variables with continuous distribution function F. Then $\min\{X,Y\}$ *and* $|X-Y|$ *are independent if and only if* $F(x) = 1 - e^{-\lambda x}$, *for some positive constant* λ.

The Kolmogorov type test $K^{\mathcal{E}}$ based on this characterization have been proposed in [10]. Beside this test we propose the corresponding integral type test $I^{\mathcal{E}}$ of the form (1). We prove that the limiting distribution is normal and found the large deviation function. Bahadur efficiencies of tests are shown in Tables 1 and 2.

From the tables we can conclude that the "order" of proposed tests differ with alternative, the order of corresponding U-statistic and type of characterization and characterization itself. As far as Milošević-Obradović characterization based tests are considered, in case $m = k$ it can be noticed that for fixed k tests based on the second case are most efficient. However, it does not necessary hold for some other choices of m and k and alternative distributions.

The results for tests based on Fisz's characterizations are reasonably good in comparison to considered tests based on Milošević-Obradović characterization. It confirms that this rather new approach has a potential.

General conclusion is that the integral type tests are much more efficient then the corresponding Kolmogorov ones. On the other hand the Kolmogorov type tests are consistent against all alternatives, while the integral type tests can be made consistent against almost all alternatives of practical purpose considering their two-tailed versions.

Acknowledgements: The research was supported by MNTRS, Serbia, Grant No. 174012.

References

[1] M. Ahsanullah, *Characterizations of Univariate Continuous Distributions.* Atlantis Studies in Probability and Statistics ser. 7., Atlantis Press, 2017.

[2] B. C. Arnold, J. A. Villasenor, Exponential characterizations motivated by the structure of order statistics in samples of size two. *Statist. Probab. Lett.* 83(2):596–601, 2013.

[3] M. Fisz, Characterization of some probability distributions. *Skand. Aktuarietidskr* 41(1-2):65–67, 1958.

[4] J. Galambos, S. Kotz, *Characterizations of Probability Distributions*, Springer-Verlag, Berlin-Heidelberg-New York, 1978.

[5] M. Jovanović, B. Milošević, Ya. Yu. Nikitin, M. Obradović and K. Yu. Volkova, Tests of exponentiality based on Arnold-Villasenor characterization, and their efficiencies. *Comput. Statist. Data Anal.* 90:100–113, 2015.

[6] B. Milošević, M. Obradović, Some characterizations of the exponential distribution based on order statistics. *Appl. Anal. Discrete Math.* 10:394–407, 2016.

[7] B. Milošević, M. Obradović, Some characterization based exponentiality tests and their Bahadur Efficiencies. *Publ. Inst. Math* 100(114):107–117, 2016.

[8] B. Milošević, M. Obradović, Characterization based symmetry tests and their asymptotic efficiencies. *Statist. Probab. Lett.* 119:155–162, 2016.

[9] B. Milošević, Asymptotic Efficiency of New Exponentiality Tests Based on a Characterization, *Metrika.* 79(2): 221–236, 2016.

[10] B. Milošević, M. Obradović, Two-dimensional Kolmogorov-type Goodness-of-fit Tests Based on Characterizations and their Asymptotic Efficiencies. *J. Nonparametr. Stat.* 28(2):413–427, 2016.

[11] M. Obradović, On Asymptotic Efficiency of Goodness of Fit Tests for Pareto Distribution Based on Characterizations. *Filomat* 9(10):2311–2324, 2015.

[12] M. Obradović, Three Characterizations of Exponential Distribution Involving Median of Sample of Size Three, *J. Stat. Theory Appl.* 14(3):257–264, 2015.

[13] M. Obradović, M. Jovanović, B. Milošević, Goodness-of-fit tests for Pareto distribution based on a characterization and their asymptotics. *Statistics* 49(5):1026–1041, 2015.

[14] Ya. Yu. Nikitin, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press, New York, 1995.

[15] Ya. Yu. Nikitin, E. V. Ponikarov, Rough large deviation asymptotics of Chernoff type for von Mises functionals and U-statistics. *Proceedings of Saint-Petersburg Mathematical Society* 7 (1999) 124–167; English translation in AMS Translations, ser. 2, 203:107–146, 2001.

[16] Ya. Yu. Nikitin, Large Deviations of U-empirical Kolmogorov-Smirnov Test, and Their Efficiency. *J. Nonparametr. Stat.* 22(5):649–668, 2010.

[17] Ya. Yu. Nikitin, Tests based on characterizations, and their efficiencies: a survey. *Acta Comment. Univ. Tartu. Math.* 21(1):4-24, 2017.

[18] K. Volkova, Goodness-Of-Fit Tests for Exponentiality Based on Yanev-Chakraborty Characterization and Their Efficiencies. *19th EYSM* 2015.

[19] G. P. Yanev, S. Chakraborty. Characterizations of exponential distribution based on sample of size three. *Pliska Stud. Math. Bulgar.* 22(1): 237–244, 2013.