Hacettepe Journal of Mathematics and Statistics
Volume 47 (4) (2018), 1061 – 1069

A bootstrap test for symmetry based on quantiles

Vali Zardasht*

Abstract

In this paper, we construct a nonparametric test for the symmetry assumption of an underling distribution based on the sample quantiles. Bootstrap re-sampling from a symmetric empirical distribution function is used to obtain the p-value of the test. The power of the new test statistic is compared with some well-known symmetry tests using a simulation study. The results show that the proposed test preserves its level and it has reasonable power properties on the family of distribution evaluated.

Keywords: Symmetry, Quantile, Test Power, Bootstrap, Simulation.

Mathematics Subject Classification (2010): 62G10, 62G09

Received : 07.09.2016 *Accepted* : 07.11.2016 *Doi* : 10.15672/HJMS.2016.386

1. Introduction

It is evident from the statistics literature that the assumption of symmetry plays an important role in statistical inferences. For example, if one is interested in estimating the measure of location, having a skewed distribution would give rise to consideration of more than one such measure. It is obvious that the validity of the inference procedures relies on holding the symmetry assumption. For instance, Munzel [9] warns against the use of the popular Wilcoxon-signed-rank test unless we accept a symmetry hypothesis for the distribution of the differences of the pairs. Many robust statistical methods (see [4]) depend on the assumption of symmetry. In case symmetry is not valid, one would need to determine a symmetrizing transformation before applying the statistical procedures. For more instances, we refer the reader to [5]. Furthermore, Ngatchou- Wandji [10] has pointed out that, in economy and finance for example, it is sometimes desirable to know whether a distribution is skewed or not. It may thus be interesting to test for skewness of distributions. Since then, the symmetry tests have received a great deal of attention in the research literature. See for example, [7], [1], [6], [15] and [12].

*Department of Statistics, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil, IRAN, Email: zardasht@uma.ac.ir

A random variable X with continuous distribution function F is said to be symmetrically distributed about point c if $X-c$ and $c-X$ are identically distributed or equivalently, for any real x

$$F(c+x) = 1 - F(c-x). \tag{1.1}$$

It can be shown that if X is symmetrically distributed about c, then c is equal to the mean (if exist), median and mode of the distribution (if it is unique). Assume that F is continuous, strictly increasing and let $\xi_p = F^{-1}(p), 0 < p < 1$ be corresponding pth quantile. It follows from equation(1.1) that

$$p = F(\xi_p) = F(\xi_p - c + c) = 1 - F(2c - \xi_p),$$

which implies that $2c-\xi_p = \xi_{1-p}$ or equivalently $2\xi_{0.5}-\xi_p = \xi_{1-p}$. Therefore, the absolute value of the difference $\xi_{1-p} + \xi_p - 2\xi_{0.5}$ can be used as a measure of the deviation of the distribution from the symmetry hypothesis. Note that for symmetric distributions this difference is equal to zero and its large value shows the asymmetry of the distribution. In this paper, we construct a test statistic based on the above measure for testing the symmetry hypothesis.

The rest of the paper is organized as follows. In section 2 we give the proposed test statistic. Section 3 is devoted to the bootstrap procedure and corresponding algorithm for obtaining the critical values of the proposed test. In Section 4, we compare the power of the test with those of some well-known tests via a simulation study.

2. Test Statistic

Let $X_1, \ldots, X_n$ be a random sample from an unknown continuous distribution function F. Let also

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

be the empirical distribution function and $\hat{\xi}_p = F_n^{-1}(p)$ be pth sample quantile. It is well-known that $\hat{\xi}_p$ is a consistent estimator of ξ_p. Regarding this, we propose our test statistic Q_n for testing the hypothesis

$$H_0 : F(c+x) = 1 - F(c-x)$$

against the alternative

$$H_1 : F(c-x) \neq 1 - F(c-x)$$

by the following:

$$Q_n = \max_{0<p<0.5} |\hat{\xi}_{1-p} + \hat{\xi}_p - 2\hat{\xi}_{0.5}|, \tag{2.1}$$

for which the large values of Q_n leads us to reject the null hypothesis in favor of the alternative H_1. The following theorem gives the asymptotic distribution of $\sqrt{n}Q_n$.

2.1. Theorem. *Suppose that F is absolutely continuous and has a density f in neighborhood of ξ_p, and that f is positive and continuous at ξ_p, $p \in (0, 1/2)$. Then,*

$$\lim_{n\to\infty} P(\sqrt{n}Q_n > x) = 2[1 - \Phi_{\sigma^2}(x)].$$

Proof. First, note that the Theorem B in [11](p. 80) and the standard delta method imply that under H_0,

$$\sqrt{n}(\hat{\xi}_{1-p} + \hat{\xi}_p - 2\hat{\xi}_{0.5}) \xrightarrow{d} N(0, \sigma^2(p)),$$

where $\sigma^2(p) = \sigma_{11} - 4\sigma_{12} + 4\sigma_{22} - 4\sigma_{23} + 2\sigma_{13} + \sigma_{33}$,

$$\sigma_{11} = \frac{p(1-p)}{[f(\xi_p)]^2}, \quad \sigma_{12} = \frac{p}{2f(\xi_p)f(\xi_{0.5})}, \quad \sigma_{13} = \frac{p^2}{f(\xi_p)f(\xi_{1-p})},$$
$$\sigma_{22} = \frac{1}{4[f(\xi_{0.5})]^2}, \quad \sigma_{23} = \frac{p}{2f(\xi_{1-p})f(\xi_{0.5})}, \quad \sigma_{33} = \frac{p(1-p)}{[f(\xi_{1-p})]^2}.$$

On the other hand, since the limiting distribution of the sample quantile process

$$\{\sqrt{n}(\hat{\xi}_p - \xi_p), 0 < p < 1\}$$

is a Brownian process(cf. [11], p. 112), the limiting distribution of the process

$$\{\sqrt{n}(\hat{\xi}_{1-p} + \hat{\xi}_p - 2\hat{\xi}_{0.5}), 0 < p < 0.5\}$$

is also a Brownian process, say $B(p)$. Hence, the limiting distribution of $\sqrt{n}Q_n$ is just the maximum of the Brownian process $B(p)$ on $\epsilon \leq p \leq 0.5 - \epsilon$, $\epsilon > 0$ and its survival function is given by (cf. [13], p. 493)

$$P\{\max_{\epsilon \leq p \leq 0.5-\epsilon} B(p) > x\} = 2[1 - \Phi_{\sigma^2}(x)],$$

where $\Phi_{\sigma^2}(x)$ is the distribution function of a normal random variable with mean zero and variance $\sigma^2 = \sigma^2(0.5 - \epsilon)$, $\epsilon > 0$. This completes the proof. □

It follows from the theorem that the mean and variance of the limiting distribution of $\sqrt{n}Q_n$ can be given by $\frac{\sigma}{\sqrt{\pi}}$, $\frac{\sigma^2(\pi-2)}{2\pi}$, respectively, which depend on the density of the underling distribution F. In addition, in practice the size of the data set is not always large enough for applying the asymptotic results. Thus, to obtain the critical values and making decision on whether or not the null hypothesis is rejected, we apply the bootstrap resampling method which is described in the next section.

3. The bootstrap test

Obtaining critical values and drawing statistical inferences rely on sampling distribution of statistics which is often difficult. For this reason, statistical inferences are usually drawn using limiting distributions and asymptotic results. Furthermore, limiting distributions are not readily available and, when they are available, they may not be valid for the finite-sized sample at hand. Research results show that applying the bootstrap procedure is really helpful in such cases. To make a decision on rejecting or accepting H_0, it is enough to obtain the p-value of the test using the bootstrap method. In order to this, we need to take samples from the symmetric version of the empirical distribution function or equivalently from the closest symmetric distribution to F_n. Modarres [8] has shown that nonparametric maximum likelihood estimator of distribution function under the symmetry assumption is

$$F_n^s(x) = \frac{1}{2}[F_n(x) + 1 - F_n(2c - x)],$$

which is indeed the symmetrized version of the empirical distribution function F_n. Thus, we take bootstrap samples from the distribution F_n^s. An unknown point c, the center of symmetry, can be replaced with an appropriate estimator like the original sample median. In similar testing problem [2] uses the sample mean for estimating the center of symmetry which yields good performance as well. For a random sample of size n, the following algorithm gives steps to decide whether reject or accept H_0 on the basis of test statistic Q_n.

(1) Compute the observed Q_n for the original sample.
(2) Generate $B = 1000$ samples of size n from distribution function F_n^s.
(3) Compute Q_n for each of these bootstrap samples.

(4) Compute the percentage of bootstrap samples with Q_n values greater than the observed Q_n (the bootstrap p-value).
(5) If this percentage is greater than $\alpha = 0.05$, accept H_0; otherwise reject H_0.

In next section, the simulation results for assessing the performance of test Q_n is given.

4. Simulation study

In this section, we assess the performance of the test statistic Q_n in terms of the closeness of its size to nominal $\alpha = 0.05$ level. We also compare the power of the test with that of some well-known competitive tests. In a simulation study, while comparing different symmetry tests, [15] have found that the bootstrap test in [6] performs well and has Type I error rate close to the nominal 0.05 level. The test statistic in [6] is given by

$$T = \frac{\bar{X} - M}{J}, \tag{4.1}$$

where $\bar{X}$ and M are the sample mean and median, respectively and

$$J = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{n} \sum_{i=1}^{n} |X_i - M|,$$

is a robust estimate of standard deviation. They have shown that under H_0, $\sqrt{n}T$ is asymptotically normal with mean zero and approximate variance 0.5708. Throughout this section, we denote the asymptotic version of the test by T^{NA} and its bootstrap one by T^B.

Staudte [12] has considered a test based on the statistic

$$\hat{\gamma}_r = \frac{X_{([nr])} + X_{(n-[nr]+1)} - 2X_{([\frac{n+1}{2}])}}{X_{(n-[nr]+1)} - X_{([nr])}},$$

and compared its power (for $r = 0.1$) with tests based on the Studentized $S_r = X_{([nr])} + X_{(n-[nr]+1)} - 2X_{([\frac{n+1}{2}])}$ (used also in [10]) and sample skewness measure $Sk = \frac{\hat{\mu}_3}{S^3}$, where $\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3$ and S is the sample standard deviation.

In this section, we denote the above three test by T_{ST}, T_{NW}, and T_{SK}, respectively. In Clear way, $T_{ST} = \sqrt{n}K(\hat{\gamma}_r)$, where

$$K(\hat{\gamma}_r) = \frac{1}{\sqrt{a_2}}[\sinh^{-1}\{l(\hat{\gamma}_r)/D\} - \sinh^{-1}\{l(a_1)/D\}], \quad l(x) = a_1 + 2a_2 x,$$

$$a_0 = [r(1-r)(g_r^2 + g_{1-r}^2) + g_{0.5}^2 + 2r^2 g_r g_{1-r} - 2r g_{0.5}(g_r + g_{1-r})]/R_r^2,$$
$$a_1 = -2[r(1-r)(g_{1-r}^2 - g_r^2) - r g_{0.5}(g_{1-r} - g_r)]/R_r^2,$$
$$a_2 = [r(g_r^2 + g_{1-r}^2) - r^2(g_r + g_{1-r})^2]/R_r^2,$$
$$D^2 = 4a_0 a_2 - a_1^2, \quad R_r = X_{(n-[nr]+1)} - X_{([nr])}, \quad g_r = [f(F^{-1}(r))]^{-1}.$$

For more details about the statistic T_{ST} see [12]. The asymptotic distribution of T_{ST} under the null hypothesis is the standard normal distribution. In addition, $T_{NW} = \frac{\sqrt{n}\hat{S}_r}{R_r\sqrt{a_0}}$ which is also asymptotically normal. Furthermore, $T_{SK} = \sqrt{n}Sk/\hat{\tau}$, where

$$\hat{\tau}^2 = (\hat{\mu}_6 - 6\hat{\mu}_2\hat{\mu}_4 + 9\hat{\mu}_2^3)/\hat{\mu}_2^3, \quad \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k,$$

which have a limiting standard normal distribution provided that 6 moments exist (cf. [3]). In our simulation, we use the asymptotic version of the above three tests.

We also consider the bootstrap test based on statistic $\hat{\gamma}_r$, for $r = 0.1$, and denote it by T_γ^B. Hence, we compare the power and size of six tests T^{NA}, T^B, T_{ST}, T_{NW}, T_{SK} and T_γ^B with that of our proposed test Q_n.

To compare the size of the tests, we choose five symmetric distributions standard normal, t-student, standard logistic, Beta distribution with parameters $(a, b) = (2, 2)$ and uniform distribution with parameter $(a, b) = (0, 1)$. Figure 1 depicts plot of the density function of these distributions. To assess the power performance of the above tests, we

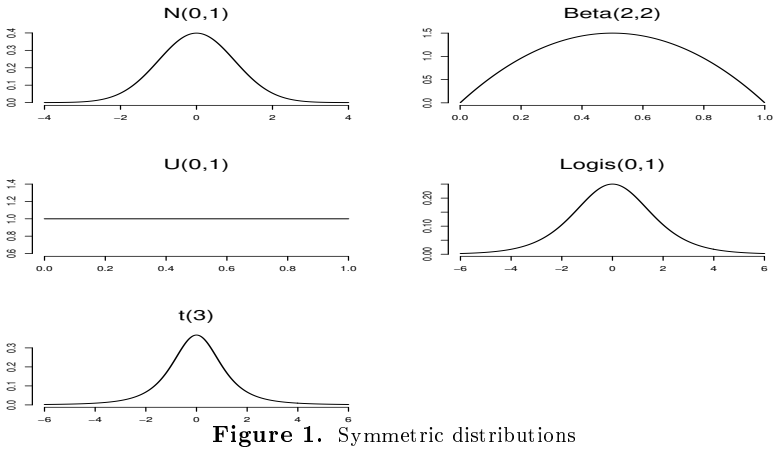

Figure 1. Symmetric distributions

choose eight asymmetric distributions with different values of skewness measures. The first is a mixture of two normal distribution. We choose four distributions coming from the generalized lambda distribution family. We denote this family by $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The corresponding inverse of distribution function is given by

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}.$$

We also use Gamma distributions with parameters $(\alpha, \beta) = (5, 7)$ and $(\alpha, \beta) = (3, 5)$ and F distribution with parameter $(n_1, n_2) = (8, 9)$ in our evaluation. Figure 2 gives the plot of the density function of these distributions along with their skewness measure $\gamma = \frac{E(X-\mu)^3}{\sigma^3}$.

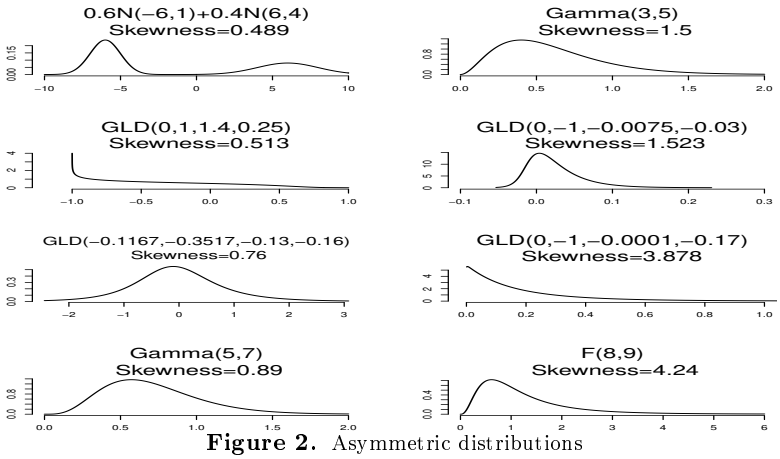

Figure 2. Asymmetric distributions

The comparison is based on different sample sizes 30, 50, 100, 300 and for each distribution and sample size, 1000 samples are simulated. It should be mentioned that to draw the maximum value in statistic Q_n, 50 equispaced points in $(0, 0.5)$ are used. The size of tests and power comparison results are reported on Tables 1 to 3.

Table 1 shows that, in terms of size of test, statistics Q_n performs well and have Type I error rate close to the nominal 0.05 level. Furthermore, Table 2 shows that, for normal mixture and $GLD(-0.1167, -0.3517, -0.13, -0.16)$, the T^{NA} test is more powerful than the other tests. But for other distributions, the Q_n test shows a better performance. Tables 2 and 3 show that Q_n is more powerful than the other tests for most of the simulated distributions.

Table 1. Size of the Tests

Dist.	n	T^B	T^{NA}	Q_n	T^B_γ	T_{ST}	T_{SK}	T_{NW}
$N(0,1)$	30	0.046	0.035	0.018	0.017	0.055	0.026	0.042
	50	0.051	0.042	0.020	0.027	0.065	0.020	0.056
	100	0.057	0.062	0.020	0.040	0.068	0.045	0.064
	300	0.063	0.056	0.025	0.041	0.064	0.049	0.063
$t(3)$	30	0.029	0.074	0.027	0.007	0.077	0.028	0.049
	50	0.044	0.093	0.023	0.017	0.085	0.018	0.064
	100	0.032	0.112	0.027	0.025	0.080	0.020	0.068
	300	0.051	0.133	0.058	0.032	0.090	0.023	0.083
$Beta(2,2)$	30	0.048	0.047	0.024	0.029	0.067	0.020	0.063
	50	0.061	0.053	0.025	0.036	0.066	0.041	0.059
	100	0.058	0.061	0.034	0.037	0.056	0.036	0.054
	300	0.065	0.067	0.042	0.050	0.067	0.051	0.067
$U(0,1)$	30	0.103	0.094	0.039	0.056	0.090	0.041	0.087
	50	0.077	0.096	0.035	0.048	0.077	0.046	0.075
	100	0.066	0.089	0.042	0.058	0.069	0.042	0.067
	300	0.071	0.109	0.045	0.058	0.073	0.040	0.073
$Logis(0,1)$	30	0.044	0.041	0.020	0.018	0.062	0.019	0.046
	50	0.054	0.058	0.024	0.019	0.077	0.026	0.066
	100	0.050	0.051	0.024	0.028	0.071	0.031	0.065
	300	0.056	0.056	0.039	0.039	0.092	0.039	0.085

4.1. Application on real life data. In this subsection, we apply the above tests for the data set used also in [15] which reports revenues available per-pupil in the 89 Educational Agencies or school districts in New Mexico and available in the R package, `lawstat`. The results in Table 4 show that all the tests accepted the symmetry hypothesis. These results are consistent with the right skewness of the data seen in the histogram in Figure 3.

5. Conclusion

In this paper, we proposed a nonparametric test for testing symmetry based on the maximum of the sum of two symmetric quantiles minus twice the median in which the bootstrap resampling method is used to compute the corresponding p-value. An extensive simulation exercise was undertaken to compare the performance of this test statistic and six other well-known test statistics. The results revealed that the proposed test has Type I error rate close to the nominal level. These results also indicated that, in most cases, the proposed test statistic is more powerful than the other competitive tests. Finally,

Table 2. Power of the Tests

Dist.	n	T^B	T^{NA}	Q_n	T^B_γ	T_{ST}	T_{SK}	T_{NW}
$0.6N(-6,1)+0.4N(6,4)$	30	0.883	0.902	0.813	0.837	0.851	0.317	0.851
	50	0.954	0.972	0.915	0.954	0.939	0.484	0.940
	100	0.981	0.991	0.967	0.979	0.973	0.747	0.973
	300	1.000	1.000	1.000	1.000	1.000	0.995	1.000
$GLD(0,1,1.4,0.25)$	30	0.188	0.214	0.359	0.170	0.330	0.220	0.296
	50	0.279	0.341	0.629	0.287	0.437	0.494	0.418
	100	0.469	0.585	0.922	0.548	0.685	0.881	0.676
	300	0.862	0.946	1.000	0.947	0.981	1.000	0.980
$GLD(-0.1167,-0.3517,$ $-0.13,-0.16)$	30	0.050	0.084	0.049	0.021	0.111	0.039	0.078
	50	0.053	0.096	0.069	0.046	0.115	0.047	0.093
	100	0.093	0.146	0.092	0.054	0.151	0.050	0.133
	300	0.188	0.275	0.184	0.161	0.259	0.108	0.253
$G(5,7)$	30	0.133	0.162	0.208	0.087	0.245	0.135	0.199
	50	0.219	0.261	0.458	0.177	0.325	0.292	0.276
	100	0.422	0.486	0.848	0.445	0.534	0.683	0.507
	300	0.916	0.935	1.000	0.917	0.934	0.982	0.933

Table 3. Power of the Tests

Dist.	n	T^B	T^{NA}	Q_n	T^B_γ	T_{ST}	T_{SK}	T_{NW}
$G(3,5)$	30	0.233	0.301	0.397	0.142	0.356	0.199	0.288
	50	0.369	0.461	0.715	0.325	0.524	0.424	0.467
	100	0.662	0.734	0.991	0.697	0.756	0.794	0.729
	300	0.986	0.993	1.000	0.994	0.995	0.979	0.994
$GLD(0,-1,0.0075,-0.03)$	30	0.229	0.317	0.379	0.153	0.389	0.182	0.299
	50	0.429	0.539	0.650	0.350	0.594	0.368	0.522
	100	0.748	0.837	0.928	0.765	0.830	0.627	0.810
	300	0.997	0.999	0.998	0.994	0.995	0.926	0.994
$GLD(0,-1,-0.0001,-0.17)$	30	0.695	0.873	0.963	0.474	0.830	0.276	0.716
	50	0.886	0.967	0.999	0.775	0.947	0.408	0.874
	100	0.999	0.999	0.999	0.996	0.996	0.513	0.990
	300	1.000	1.000	0.996	1.000	1.000	0.751	1.000
$F(8,9)$	30	0.491	0.695	0.793	0.294	0.667	0.220	0.535
	50	0.777	0.891	0.982	0.637	0.844	0.343	0.764
	100	0.979	0.989	1.000	0.974	0.976	0.496	0.950
	300	1.000	1.000	0.994	1.000	1.000	0.676	1.000

Table 4. P-values of the Tests

T^B	T^{NA}	Q_n	T^B_γ	T_{ST}	T_{SK}	T_{NW}
0.0079	0.0000004	0.01	0.017	0.0017	0.03544	0.00665

using a numerical example, the use of the test statistic for testing symmetry of data was illustrated.

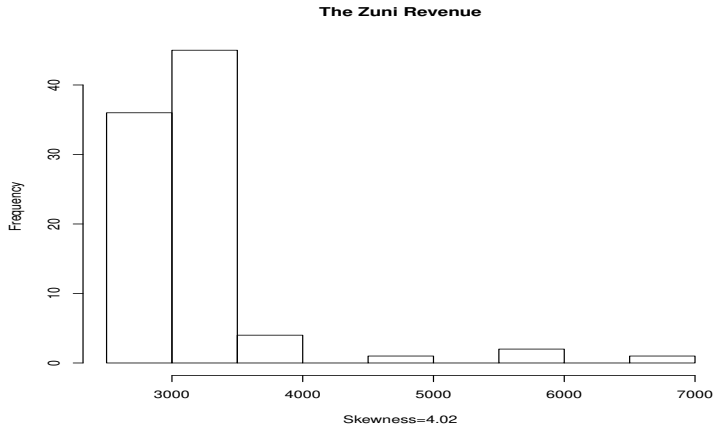

Figure 3. Histogram of the revenue data

Acknowledgement. The author would like to express his gratitude to the referees for their valuable comments, suggestions and careful reading of the manuscript which have significantly improved the original version of this paper.

References

[1] Cabilio, P. and Masaro, J. *A simple test of symmetry about an unknown median,* Canadian Journal of Statistics **24**, 349-361, 1996.

[2] Drikvandi, R., Modarres, R. and Jalilian, A. H. *A bootstrap test for symmetry based on ranked set samples,* Computational Statistics and Data Analysis **55**, 1807-1814, 2011.

[3] Gupta, M. K. *An asymptotically nonparametric test of symmetry,* The Annals of Mathematical Statistics **38**(3), 849-866, 1967.

[4] Huber, P. J. *Robust statistics: A review,* The Annals of Mathematical Statistics **43**, 1041–1067, 1972.

[5] Meintanis, S.G. and Ngatchou-Wandji, J. *Recent tests for symmetry with multivariate and structured data,* in: Jiang, J., Roussas, G. G. and Samaniego, F. J. (Eds.), Non-Parametric Statistical Methods and Related Topics: A Festschrift in honor of Professor P.K. Bhattacharya on the occasion of his 80th birthday (World Scientific, New Jersey, 2012,) 35-73.

[6] Miao, W., Gel, Y. R., and Gastwirth, J. L. *A New Test of Symmetry about an Unknown Median. Random Walk, Sequential Analysis and Related Topics - A Festschrift in Honor of Yuan-Shih Chow. Eds.: Agnes Hsiung, Cun-Hui Zhang, and Zhiliang Ying,* (World Scientific Publisher, 2006).

[7] Mira, A. *Distribution-free test for symmetry based on Bonferroni's measure,* Journal of Applied Statistics **26**, 959-972, 1999.

[8] Modarres, R. *Efficient nonparametric estimation of a distribution function,* Computational Statistics and Data Analysis **39**, 75-95, 2002.

[9] Munzel, U. *Nonparametric methods for paired samples,* Statistica Neerlandica **53**, 277-286, 1999.

[10] Ngatchou- Wandji, J. *On testing for the nullity of some skewness coefficients,* International Statatistics Review **74**, 47-65, 2006.

[11] Serfling, R. J. *Approximation Theorems of Mathematical Statistics* (John Wiley and Sons Inc, 1980).

[12] Staudte, R. G. *Inference for quantile measures of skewness,* TEST **23**, 751-768, 2014.

[13] Taylor, H. M. and Karlin, S. *An Introduction to Stochastic Modeling* (Academic Press, 1998).

[14] Van der vaart, A. W. *Asymptotic statistics* (Cambridge University Press, 1998).

[15] Zheng, T. and Gastwirth, J.L. *On Bootstrap Tests of Symmetry About an Unknown Median,* Journal of Data Science **8**(3), 413-427, 2010.