Wrocław University of Economics

EKONOMICKÁ FAKULTA

VŠE

CONFERENCE PROCEEDINGS
FULL TEXT PAPERS

edited by
Zofia Rusnak and Beata Zmyślona

17th AMSE

APPLICATIONS OF MATHEMATICS AND STATISTICS IN ECONOMICS

International Scientific Conference | Poland • 27-31 August 2014

ISBN 978-83-7695-421-9

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
53-345 Wrocław, ul. Komandorska 118/120
www.ue.wroc.pl
Sprzedaż książek tel./fax 71 36-80-602
e-mail: econbook@ue.wroc.pl www.ksiegarnia.ue.wroc.pl

Contents

17th AMSE APPLICATIONS OF MATHEMATICS AND STATISTICS IN ECONOMICS
International Scientific Conference | Poland • 27-31 August 2014

CONSEQUENCES OF ASSUMPTION VIOLATIONS REGARDING CLASSICAL LOCATION TESTS

TOMÁŠ MARCINKO
University of Economics in Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability, nám. W. Churchilla 4, 130 67 Prague, Czech Republic
email: xmart14@vse.cz

Abstract
Nearly all classical statistical hypothesis tests are derived under a few fundamental assumptions, which may or may not be met in real world applications. The main aim of this article is to study consequences of a normality assumption violation concerning classical statistical methods, mainly its effect on type I and type II errors when dealing with one-sample or two-sample location tests. The focus will be on a very popular one-sample t-test, as well as on a Behrens-Fisher problem, i.e. on hypothesis testing concerning the difference between expected values of two random variables with unknown and possibly different variances. Based on a simulation study the consequences of different forms of non-normality will be examined for various sample sizes. Type I and type II errors of the classical tests will be then compared with those of appropriate nonparametric tests, specifically with the errors of the Wilcoxon signed-rank and rank-sum tests, as well as the tests based on bootstrap methodology. Based on the results of the conducted simulation study it can be inferred that the classical t-tests tend to be conservative or liberal depending on a form of non-normality. It will be also demonstrated that in case of a contaminated distribution with possible outliers the Wilcoxon tests should be always considered, and that for skewed data and a large sample size the bootstrap BC_a method may also be preferable.

Key words: *one-sample t-test, Behrens-Fisher problem, normality violation, Wilcoxon tests, bootstrap.*

DOI: 10.15611/amse.2014.17.20

1. Introduction

Location tests are arguably the most important statistical tests, which are used to determine, if the location parameter is equal to a given constant (one-sample problem), or if the location parameters of two populations are the same (two-sample problem). Most commonly, the location parameter of interest is the expected value, although in some cases the median or some other measures of location may be used.

We will focus on two very common statistical problems: the one-sample location test with the expected value being the location parameter and the variance of the population being unknown, and the Behrens-Fisher problem concerning difference between expected values of two random variables with unknown variances, which are not assumed to be equal.

Undoubtedly the most popular tests regarding these two problems are the one-sample Student's t-test and the approximate two-sample Welch's t-test. However, both these parametric tests were derived under a couple of assumptions, which may not be met in real world applications. Namely, we assume that the random sample come from populations that follow a normal distribution and the data are sampled independently (i.e. the observations in

the sample from any population are assumed to be independent and identically distributed following a normal distribution with the same expectation and the same variance).

The main aim of this article is a simulation study that will examine the consequences of non-normality, mainly its negative effect on type I and type II errors. The results obtained by parametric t-tests will be then compared with those of appropriate nonparametric tests, specifically with the Wilcoxon signed-rank and rank-sum tests, as well as the tests based on bootstrap methodology.

2. Parametric and nonparametric approach to location tests

When dealing with location tests, the most popular approach of many statisticians, researchers or data analysts is the parametric one, i.e. using the Student's t-test when dealing with a one-sample problem or the Welch's t-test when dealing with a two-sample problem.

Let's assume we have a one-sample problem and we wish to determine, whether the population mean μ is equal to a specified value μ_0. The Student's t-test uses the statistic

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}, \tag{1}$$

where $\bar{x}$ is a sample mean, s is a sample standard deviation and n is a sample size. If the observations are independent, identically distributed and follow a normal distribution then it can be easily shown that under the null hypotheses $H_0: \mu = \mu_0$ the statistic t follows a Student's t distribution on $n-1$ degrees of freedom. Therefore, if we state the two-tailed alternative hypothesis $H_1: \mu \neq \mu_0$, we will reject the null hypothesis in favor of the alternative hypothesis when the absolute value of the statistic t is greater than a critical value from the Student's t distribution. Moreover, it can be shown that the Student's t-test is in fact a uniformly most powerful unbiased test, for details see Lehmann and Romano (2005).

Although the statistic t follows under the null hypothesis the Student's t distribution exactly only under the assumption of normality, this test is also often used for larger samples (e.g. $n > 30$). The reason for this is the fact that by the central limit theorem the mean of a sufficiently large number of iterates of independent and identically distributed random variables will be approximately normal, even if the underlying distribution is not. However, in this case the t-test may not be the most powerful.

For a Behrens-Fisher problem the two-sample Welch's t-test is probably the most often used parametric solution. Let μ_1 and μ_2 be the population means in first and second population, respectively, and let the null hypothesis be of the form $H_0: \mu_1 - \mu_2 = \Delta$. Then the test statistic for the Welch's test is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \tag{2}$$

where $\bar{x}_1$, $\bar{x}_2$ are sample means, s_1, s_2 are sample standard deviations and n_1, n_2 are sample sizes for respective populations. Although, unlike the one-sample Student's t test, this test statistic does not follow under the null hypothesis the Student's t distribution on a given degrees of freedom exactly, Welch (1947) proposed an approximation of the degrees of freedom associated with this statistic via the Welch-Satterthwaite equation

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}. \tag{3}$$

For details on derivation of the Welch-Satterthwaite equation refer to Satterthwaite (1946) and Welch (1947). Although other solutions to the Behrens-Fisher problem have been developed, the Welch's test still tends to be the most often used. In the **R** programming language, both Student's and Welch's tests are computed by t.test function.

Probably the best known nonparametric alternatives to the aforementioned parametric t-tests are the one-sample Wilcoxon signed-rank test and the two-sample Wilcoxon rank-sum test, which were both proposed by Wilcoxon (1945). However, these tests based on ranks assume the median rather than the expected value as the location parameter, i.e. they may not be a suitable alternative for the t-tests in case of an asymmetric underlying distribution. In the **R** programming language, both Wilcoxon tests are performed by wilcox.test function.

Another nonparametric approach to location tests is based on bootstrap methodology. Bootstrap as a computer-intensive method has an obvious advantage of being free of assumptions concerning underlying distribution, i.e. location tests based on bootstrap methodology can be used for any distribution and even for any measure of location besides the expected value. Bootstrap hypothesis testing is often derived from respective bootstrap confidence intervals, however the coverage probabilities of these intervals are only asymptotically accurate, i.e. for a small sample sizes bootstrap hypothesis testing can lead to a type I error that is higher than the given significance level.

In this article we will consider four bootstrap confidence intervals described by Efron and Tibshirani (1993). The first one, called a bootstrap-t confidence interval, is a modification of the Student's t interval, which in case of a one-sample location problem has the form

$$\left(\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}\,;\, \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right), \tag{4}$$

where $t_{1-\alpha/2}$ is 100(1–α/2)th percentile of Student's t distribution on $n - 1$ degrees of freedom and α is the given significance level. The bootstrap modification is achieved by the sample standard deviation s being replaced by a bootstrap estimate of a standard deviation and the percentiles $t_{1-\alpha/2}$ being replaced by the empirical percentiles of the bootstrap t statistic

$$t^*(b) = \frac{\bar{x}^*(b) - \bar{x}}{s^*(b)}\sqrt{n}, \quad b = 1, 2, ..., B, \tag{5}$$

where B is a number of bootstrap samples, $\bar{x}^*(b)$ is the mean in the bth bootstrap sample and $s^*(b)$ is the estimated standard error of $\bar{x}^*(b)$. For details see Efron and Tibshirani (1993, chapter 12.5).

The biggest disadvantage of the bootstrap-t interval is that this method of confidence interval estimation is neither transformation-respective nor range-preserving. The percentile interval proposed by Efron (1979) on the other hand has both these properties. After generating B independent bootstrap data sets, the percentile interval is given by the 100(α/2)th and 100(1–α/2)th empirical percentile acquired from the $\bar{x}^*(b)$ values. This interval tends to

be less erratic then the bootstrap-t interval in actual practice, but in some cases may have less satisfactory coverage properties. Another modification of the percentile method called the BC_a (bias-corrected and accelerated) interval was proposed by Efron (1987). The last bootstrap method to be considered in this article is the ABC method, which generates only approximate bootstrap confidence intervals, but significantly reduces the amount of computation needed for the BC_a intervals. For details on the BC_a and ABC methods see Efron and Tibshirani (1993, chapter 14). In the **R** programming language, the bootstrap confidence intervals can be computed by using the boott, bootstrap, bcanon and abcnon functions of the bootstrap package. Another option is the boot.ci function of the boot package.

3. Simulation study

The following simulation study will focus on the Monte Carlo estimation of type I and type II errors of the location tests shortly discussed in the previous section under various violations of normality. For this purpose we will simulate data from the following distributions:

- normal distribution: $X \sim N(\mu = 100; \sigma^2 = 100)$
- modified Student's distribution: $X \sim 100 + 5{,}7735\ t(\nu = 3)$
- uniform distribution: $X \sim U(a = 82{,}6795; b = 117{,}3205)$
- gamma distribution: $X \sim \Gamma(\kappa = 100; \theta = 1)$
- log-normal distribution: $X \sim LN(\mu = 4{,}6002; \sigma^2 = 0{,}00995)$
- skew normal distribution: $X \sim SN(\xi = 88{,}417; \omega = 15{,}303; \alpha = 3)$
- shifted exponential distribution: $X \sim 90 + Ex(\lambda = 0{,}1)$
- contaminated normal distribution: $X \sim (1 - \varepsilon)\ N(100; 100) + \varepsilon\ N(100; 10\ 000)$.

Without the loss of generality, all of these distributions (except for a contaminated normal distribution, which has larger variance) were calibrated so that they have the population mean 100 and the variance also 100. The first three distributions (normal, modified Student's and uniform) are symmetric around the population mean and the other four distributions (gamma, log-normal, skew normal and shifted exponential) are asymmetric with a gamma distribution being the least skewed ($\gamma_1 = 0{,}2$) and a shifted exponential distribution being the most skewed ($\gamma_1 = 2$). The skewness of the other two distributions is only moderate ($\gamma_1 = 0{,}301$ for a log-normal distribution and $\gamma_1 = 0{,}667$ for a skew normal distribution). Lastly, the contaminated normal distribution is a mixture distribution, where the majority of the population comes from a specified normal distribution, whereas a small proportion of the population ($\varepsilon = 0{,}05$) comes from a normal distribution with the same mean but much larger variance, i.e. outliers can be drawn from such a population.

The simulation study consisted of 10 000 simulated data sets so that the Monte Carlo error was sufficiently low (see Figure 1 comparing the exact power function of the Student's t-test under a normal distribution, which was derived using a non-central t distribution, and the Monte Carlo estimate of this power function). For the bootstrap methods 5 000 bootstrap samples were generated.

The simulation study was computed in the programming language **R** version 2.15.3.

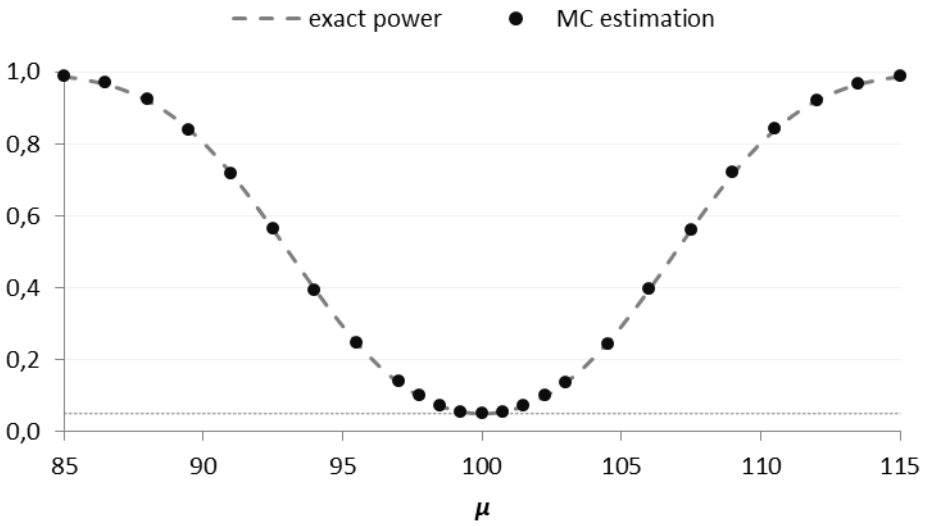

Figure 1 Exact vs estimated power function of Student's t-test (normal distribution, $n = 10$)

3.1. One-sample location test

Based on a Monte Carlo simulation study, it is simply possible to estimate the power function of particular tests under the given violations of normality. Apart from type I and type II errors, the average length of relevant confidence intervals was computed. If several tests have type I error equal to or less than a chosen significance level (for the purpose of this study we will use a significance level 0,05), the test having the shortest length of the relevant confidence interval can be preferred, as such tests tend to be more powerful in most situations.

Table 1 provides estimates of type I errors for various one-sample two-tailed location tests, underlying distributions and sample sizes. It is quite obvious that for small sample sizes different violations of normality can have a different effect on type I errors of the Student's t-test. When dealing with symmetric distributions, it seems that distributions with higher kurtosis (such as a Student's t distribution) tend to make the Student's t-test more conservative (having type I error lower than the significance level), whereas the distributions with smaller kurtosis (such as a uniform distribution) tend to make the Student's t-test more liberal (having type I error higher than the significance level). On the other hand, skewed distributions always tend to make the Student's t-test more liberal, nevertheless, unless the skewness is substantial (e.g. in case of an exponential distribution), the effect of skewness is relatively small. And, as expected, with larger sample sizes the type I errors get closer to the nominal significance level.

The only exception, when the Student's t-test was too conservative even for large sample sizes, is the case of a contaminated normal distribution. This is no surprise, as outliers can have a very big influence on the sample mean. Consequently, when we are dealing with symmetric distributions with substantially long tails or with possible outliers, the Wilcoxon test should always be preferred (see Table 2 comparing the average length of relevant confidence intervals and Figure 2 comparing power functions of the Student's and Wilcoxon tests for a contaminated normal distribution).

On the other hand, the only exception, when the Student's t-test was too liberal even for large sample sizes, is the case of a shifted exponential distribution, i.e. a distribution with substantial skewness. In such cases, bootstrap-t can give better results.

Another conclusion that can be inferred from the simulation study is that the bootstrap percentile method, BC_a method and ABC method should not be recommended for small and even moderate sample sizes, as the coverage probability of the relevant confidence intervals is significantly lower than the required confidence level. However, for larger sample sizes ($n > 50$) the type I error is reasonably close to the given significance level and these tests tend to have indeed more power than the Student's t-test, although the difference in power is slim (Table 3 provides the average length of Student's and bootstrap tests for $n = 100$). Although all of these bootstrap methods tend to give very similar results for the considered location problem, from theoretical point of view the BC_a method should be preferred when available.

Table 1. Estimated type I errors of various one-sample two-tailed location tests

size	distribution	Student's t-test	Wilcoxon test	bootstrap-t	percentile method	BC_a method	ABC method
$n = 10$	normal	0,0496	0,0487	0,0566*	0,0987*	0,1002*	0,0985*
	t(3)	0,0411*	0,0487	0,0737*	0,1076*	0,1379*	0,1377*
	uniform	0,0546*	0,0487	0,0320*	0,0890*	0,0666*	0,0656*
	gamma	0,0510	-	0,0567*	0,0999*	0,1007*	0,1000*
	log-normal	0,0520	-	0,0538	0,1005*	0,1017*	0,0997*
	skew normal	0,0566*	-	0,0559*	0,1039*	0,1015*	0,0990*
	exponential	0,1024*	-	0,0626*	0,1413*	0,1228*	0,1202*
	contaminated	0,0357*	0,0496	0,0683*	0,1061*	0,1466*	0,1442*
$n = 50$	normal	0,0494	0,0473	0,0557*	0,0570*	0,0566*	0,0566*
	t(3)	0,0450*	0,0473	0,0723*	0,0656*	0,0878*	0,0877*
	uniform	0,0486	0,0473	0,0455*	0,0562*	0,0496	0,0487
	gamma	0,0495	-	0,0540	0,0563*	0,0579*	0,0581*
	log-normal	0,0489	-	0,0549*	0,0565*	0,0584*	0,0585*
	skew normal	0,0500	-	0,0518	0,0567*	0,0567*	0,0576*
	exponential	0,0624*	-	0,0547*	0,0679*	0,0655*	0,0655*
	contaminated	0,0298*	0,0476	0,1109*	0,0790*	0,1561*	0,1577*
$n = 100$	normal	0,0480	0,0481	0,0511	0,0520	0,0517	0,0518
	t(3)	0,0462	0,0481	0,0716*	0,0603*	0,0788*	0,0782*
	uniform	0,0483	0,0481	0,0525	0,0508	0,0488	0,0481
	gamma	0,0484	-	0,0540	0,0533	0,0520	0,0516
	log-normal	0,0491	-	0,0549*	0,0527	0,0525	0,0522
	skew normal	0,0516	-	0,0546*	0,0550*	0,0539	0,0542
	exponential	0,0601*	-	0,0567*	0,0621*	0,0604*	0,0597*
	contaminated	0,0380*	0,0495	0,1111*	0,0765*	0,1371*	0,1373*

* estimated type I error differs significantly from the significance level 0,05 based on the exact binomial test using the procedure proposed by Clopper and Pearson (1934)

Table 2. Average length of relevant confidence intervals (normal vs long-tailed distributions)

Size	distribution	Student's t-test	Wilcoxon test	bootstrap-t	percentile method	BC_a method	ABC method
$n = 10$	normal	13,886	14,244	15,473	type I error too high		
	t(3)	12,477	12,746	16,488	type I error too high		
	contaminated	26,455	28,209	48,273	type I error too high		
$n = 50$	normal	5,656	5,799	5,891	5,454	5,465	5,466
	t(3)	5,340	4,199	5,867	5,142	5,326	5,319
	contaminated	12,542	6,401	15,687	12,001	13,427	13,383
$n = 100$	normal	3,961	4,049	4,137	3,889	3,893	3,894
	t(3)	3,811	2,901	4,125	3,737	3,836	3,834
	contaminated	9,234	4,444	10,550	9,063	9,633	9,574

Table 3. Average length of relevant confidence intervals (non-normal distributions, $n = 100$)

size	distribution	Student's t-test	Wilcoxon test	bootstrap-t	percentile method	BCa method	ABC method
$n = 100$	uniform	3,9654	4,1901	4,1257	3,8919	3,8957	3,8976
	gamma	3,9603	-	4,1299	3,8892	3,8928	3,8939
	log-normal	3,9599	-	4,1282	3,8885	3,8941	3,8943
	skew normal	3,9580	-	4,1596	3,8870	3,8975	3,8980

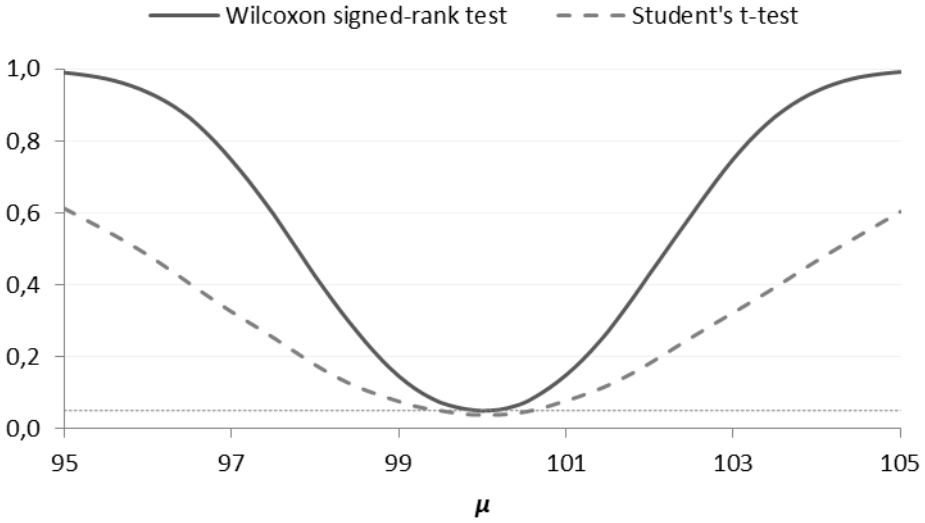

Figure 2 Power functions of the t-test and Wilcoxon test (contaminated distribution, $n = 100$)

3.2. Two-sample location test

In order to examine the mere effects of non-normality on the two-sample location tests, the sample sizes and all the distribution parameters were kept the same for both samples. The conducted simulation study showed that the effects of non-normality on the Welch's t-test are quite similar to the ones of the Student's t-test in the one-sample case (see Table 4 for estimated type I errors). However, some small differences were found.

First of all, there seems to be a smaller effect of skewness or kurtosis on type I errors of the two-tailed Welch's t-test, when a sample size is small. For larger samples ($n \geq 50$), the only time the Welch's t-test was proved to be conservative was the case of a contaminated normal distribution.

Table 4. Estimated type I errors of various two-sample two-tailed location tests

size	distribution	Welch's t-test	Wilcoxon test	bootstrap-t	percentile method	BCa method	ABC method
n = 10	normal	0,0488	0,0435*	0,1038*	0,0969*	0,0979*	0,0977*
	t(3)	0,0398*	0,0435*	0,1458*	0,1012*	0,1229*	0,1216*
	uniform	0,0503	0,0435*	0,0905*	0,0941*	0,0905*	0,0893*
	gamma	0,0488	-	0,1035*	0,0969*	0,0981*	0,0977*
	log-normal	0,0481	-	0,1033*	0,0961*	0,0995*	0,0981*
	skew normal	0,0451*	-	0,1082*	0,0956*	0,1018*	0,1005*
	exponential	0,0345*	-	0,1501*	0,1057*	0,1331*	0,1318*
	contaminated	0,0296*	0,0425*	0,2453*	0,1068*	0,1602*	0,1629*
n = 50	normal	0,0513	0,0502	0,0634*	0,0594*	0,0603*	0,0609*
	t(3)	0,0482	0,0502	0,0918*	0,0669*	0,0817*	0,0813*
	uniform	0,0509	0,0502	0,0601*	0,0580*	0,0555*	0,0554*
	gamma	0,0520	-	0,0663*	0,0603*	0,0609*	0,0609*
	log-normal	0,0521	-	0,0649*	0,0615*	0,0620*	0,0614*
	skew normal	0,0522	-	0,0684*	0,0621*	0,0636*	0,0625*
	exponential	0,0515	-	0,0817*	0,0666*	0,0771*	0,0770*
	contaminated	0,0401*	0,0527	0,1646*	0,0834*	0,1411*	0,1405*
n = 100	normal	0,0510	0,0510	0,0605*	0,0565*	0,0552*	0,0560*
	t(3)	0,0498	0,0510	0,0803*	0,0595*	0,0749*	0,0739*
	uniform	0,0506	0,0510	0,0590*	0,0527	0,0524	0,0516
	gamma	0,0503	-	0,0591*	0,0573*	0,0567*	0,0559*
	log-normal	0,0511	-	0,0615*	0,0558*	0,0558*	0,0569*
	skew normal	0,0506	-	0,0645*	0,0576*	0,0573*	0,0579*
	exponential	0,0508	-	0,0681*	0,0587*	0,0632*	0,0641*
	contaminated	0,0442*	0,0510	0,1161*	0,0679*	0,1080*	0,1082*

* estimated type I error differs significantly from the significance level 0,05 based on the exact binomial test using the procedure proposed by Clopper and Pearson (1934)

On the other hand, most of the results that were point out for the nonparametric one-sample tests remain valid also for the Behrens-Fisher problem, i.e. the Wilcoxon rank-sum test should always be preferred for contaminated or mixture distributions (or in presence of outliers, which cannot be omitted) and the bootstrap methods cannot be recommended for smaller sample sizes. Furthermore, the asymptotic nature of the bootstrap methods seems to work only for very large sample sizes ($n > 100$), as these tests were a bit liberal even at $n = 100$. Hence, the preference of the bootstrap methods in case of mere asymmetry of the two underlying distributions seems to be less justified.

4. Problem of symmetry of the Student's confidence intervals

The conducted simulation study showed that there is only a small effect of skewness on the type I error of the two-tailed Student's t-test. However, this is not true in case of one-tailed t-tests. In fact, the simulation study showed, inter alia, that for a skew normal distribution and a quite large sample size ($n = 100$) the estimated type I error of the one-tailed Student's test is 0,0579 in case of a left-tailed alternative hypothesis $H_1: \mu < \mu_0$ and 0,0440 in case of a right-tailed alternative hypothesis $H_1: \mu > \mu_0$. This means that, even though the overall coverage probability of the corresponding two-tailed confidence interval is good, in case of positively skewed data the coverage probability of the left-tailed confidence interval will be significantly larger than the coverage probability of the right-tailed one. On the other hand, the coverage probabilities of the one-tailed confidence intervals based on the bootstrap BC_a and ABC methods are very similar to the coverage probability of the respective two-tailed interval.

For better understanding of this problem, we will use an example data set, which provides annual income data of 143 highly-educated male employees, who had a percentile score on the AFQT intelligence test more than 0,9. The data come from the National Longitudinal Study of Youth and were also published by Ramsey and Shafer (2013).

As expected for income data, the underlying distribution is substantially skewed. In fact, the sample skewness is 2,3 and the sample kurtosis is 9,8. From the large positive skewness it can be inferred that both the Student's t-test and the test based on the bootstrap BC_a method will have type I error a little bit higher than the nominal significance level 0,05. If we compare Student's 95% confidence interval for the population mean (89 962; 117 444) and the corresponding BC_a confidence interval (91 790; 119 490), we see that the length of both confidence intervals is quite similar, but the Student's interval is shifted to the left compared to the BC_a interval. As the sample size is sufficiently large, so that the type I error of the bootstrap interval will be reasonably close to the nominal significance level, the BC_a confidence interval, as well as the corresponding location test, should be preferred.

5. Practical aspects of liberalness or conservativeness of a testing procedure

In the previous section we demonstrated on a simple example that the BC_a method should be preferred, when dealing with skewed data, provided a sample size is sufficiently large. Another practical aspect of the results obtained by the simulation study was the information that the classical one-sample t-test tend to be liberal for skewed data, whereas it is conservative for longer-tailed data, especially when a sample size is small. On the other hand, Welch's two-sample t-test showed no tendency to being liberal, although it can be conservative for both long-tailed and skewed distributions in case of a small sample. This information can be quite useful in real world applications (incl. economic ones) that deal with

non-normal data – while the rejection of a null hypothesis by a liberal test may be spurious, the conservativeness of a testing procedure should not invalidate the rejection of a null hypothesis. Consequently, when the t-test is deemed to be liberal, it is recommended to check the rejection of a null hypothesis by alternative methods. Similarly, acceptance of a null hypothesis should not be based solely on a parametric test that is proven to be conservative.

6. Conclusion

The classical location tests – the one-sample Student's t-test and the two-sample Welch's t-test – are derived under the assumption that the observations from the random sample are independent and identically distributed following a normal distribution. The aim of this article was to examine the effect of non-normality on type I and type II errors of these tests.

The conducted simulation study showed that both the Student's and Welch's two-tailed t-tests are a bit sensitive to even moderate deviations from normality, esp. in case of a small sample size, and can become liberal or conservative depending on a form of non-normality. When the sample size gets larger the type I error gets closer to the nominal significance level, however, for skewed data and the one-sample t-test there still might be a small problem with undesirable type I errors in case of one-tailed alternative hypotheses caused by inherent symmetry of the classical methods. This problem can be circumvented by the BC_a method provided a sample size is large. Furthermore, the classical tests cannot be recommended for contaminated data and the use of Wilcoxon tests should be considered instead.

Acknowledgements

The support of the Internal Grant Agency of the University of Economics in Prague (project IGA 128/2014 "Consequences of assumption violations of classical statistical methods and the possible use of alternative statistical techniques in economic applications") is gladly acknowledged.

References

1. CLOPPER, C.J., PEARSON, E.S. 1934. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. In Biometrika, vol. 26, iss. 4, pp. 404-413.
2. EFRON, B. 1979. Bootstrap Methods: another look at the jackknife. In Annals of Statistics, vol. 7, iss. 1, pp. 1-26.
3. EFRON, B. 1987. Better Bootstrap Confidence Intervals. In Journal of the American Statistical Association, vol. 82, iss. 397, pp. 171-185.
4. EFRON, B., TIBSHIRANI, R.J. 1993. An Introduction to the Bootstrap. Boca Raton: Chapman & Hall/CRC. 1993. ISBN 0-412-04231-2
5. LEHMANN, E.L., ROMANO, J.P. 2005. Testing Statistical Hypotheses. New York: Springer. 2005. ISBN 0-387-98864-5
6. RAMSEY, F.L., SCHAFER, D.W. 2013. The Statistical Sleuth: A Course in Methods of Data Analysis. Boston: Brooks/Cole, Cengage Learning. 2013. ISBN 1-133-49067-0
7. SATTERTHWAITE, F.E. 1946. An Approximate Distribution of Estimates of Variance Components. In Biometrics Bulletin, vol. 2, iss. 6, pp. 110-114.
8. WELCH, B.L. 1947. The generalization of "Student's" problem when several different population variances are involved. In Biometrika, vol. 34, iss. 1-2, pp. 28-35.
9. WILCOXON, F. 1945. Individual Comparisons by Ranking Methods. In Biometrics Bulletin, vol. 1, iss. 6, pp. 80-83.