sciendo

NS global economics

ECONOMICS
Volume 7, No.2 2019
ISSN 2303-5005

APPLICATION OF THE T-TEST IN HEALTH INSURANCE COST ANALYSIS: LARGE DATA SETS

Bojan Kresojević[1], Milica Gajić[1]

date of paper receipt: **09.12.2019.**

date of sending to review: **12.12.2019.**

date of review receipt: **19.12.2019.**

Review Article

doi: **10.2478/eoik-2019-0024**

UDK: **336.767:330.3(497.6RS)**

[1]University of Banja Luka, Faculty of Economics, **Bosnia and Herzegovina**

ABSTRACT

In this paper will be analyzed the application of the t-test against the nonparametric Mann - Whitney test in the analysis of health insurance benefit costs in the Republic of Srpska on large samples. This research aims to examine which method produces better results when testing statistical hypotheses. The adequacy of the statistical tests will be tested on primary health insurance cost data for 1,044,690 insureds in 2017. For two samples of size 4,000, the sampling distribution of the difference in two means has a skewness coefficient of 0.05 and a kurtosis coefficient of 3.09. Jarque - Bera test does not reject the hypothesis of normality of distribution with a p-value of 0.135. On the other hand, in the Mann - Whitney test, the real risk of the first species, when there is a difference in skewness between the samples, may be less than 0.001 compared to the nominal risk level of 0.05. Based on the results obtained, it is suggested to use the t-test instead of the Mann - Whitney test if the sample is large enough, which should be verified by the bootstrap method.

Keywords:
t-test, Mann -Whitney test, health insurance costs, large samples, distribution normality, skewness, kurtosis.

JEL: C12, I13, G22

INTRODUCTION

It is well known in the theory and practice of insurance that the cost of health insurance benefits has a distribution that, most often, deviates significantly from the normal. On the other hand, in a lot of situations, there is a need to analyze the association between the categorical variable and the cost of health insurance. If an independent variable has two categories, testing the relation is reduced to testing the statistical hypotheses with two independent samples. The most commonly used statistical tests with two samples are t-test and Mann - Whitney test. Many authors suggesting that the condition for applying the t-test is the normality of distribution. The Mann - Whitney test is usually recommended as an alternative to the t-test when the assumption of normality of distribution is not fulfilled.

In this paper will be analyzed the application of the t-test against the nonparametric Mann - Whitney test in the analysis of health insurance benefit costs in the Republic of Srpska on large samples. The analysis will be conducted on the health insurance cost data for all insured persons in 2017 that are covered by compulsory health insurance.

This study aims to test whether the t-test or the Mann - Whitney test is a better choice when we are analyzing health insurance costs in the Republic of Srpska on large samples. In other words, the key research question is which statistical test has a smaller deviation of the actual risk of the first kind relative to the nominal level of risk. This question also arises in many other areas where the data in the sample follow a distribution that is statistically significantly different from normal.

The paper is divided into four parts. Firstly, the literature review part will be presented opinions of key authors that suggest or disclaim usage of t-test on non-normal data. Secondly, the methodology part will be described in the data and methods. Then, the results will be presented in the third part. Finally, the results of the paper will be compared with the results of other authors in the discussion part.

1. LITERATURE REVIEW

Statistical methods can be divided into parametric and non-parametric methods. Parametric methods are those methods that imply application to data that have an approximately normal distribution (Pandey, 2015). When it comes to comparing one, two or more means, t-test or ANOVA (analysis of variance) is applied. Non-parametric methods are applied when the normality of distribution can't be ensured and then methods such as Mann – Whitney and Kruskal - Wallis tests, are used.

Fay &Proschan (2010) concluded that the Wilcoxon - Mann - Whitney rank-based test produces significantly better results than the t-test when it comes to data with pronounced skewness. However, a nonparametric rank-based test does not directly test the mean. According to the creators of this test (Mann & Whitney, 1947), this test examines the equality of distributions, which means that differences in other statistical moments (standard deviation, skewness, and kurtosis) can lead to rejection of the null hypothesis even when there is no difference in means. The Wilcoxon - Mann - Whitney test is often used to compare means and medians, predominantly in data with a distribution that deviates from normal. However, Fagerland&Sandvik (2009a) have shown by simulation that small deviations in skewness can lead to large deviations of type I error relative to the nominal level. This shows that the lack of robustness of this test to skewness is much more serious than previously thought.

According to the Central Limit Theorem, the distribution that deviates from normal becomes normal when the sample increases (Cundill& De Alexander, 2015). More specifically, the sample distribution of the mean becomes normal when the sample is enlarged enough. The rule of thumb in the literature is that samples larger than 30 are large samples and that their distribution is approximately normal (see Lovric, Komic&Stevic, 2017). The problem is that each distribution

does not deviate equally from the normal one, thus no universal limit can be set when it comes to approaching the normal distribution. Chang, Wu, Ho & Chen (2008) found by simulation that a sample size of 150 is required for with a gamma distribution (with beta and gamma coefficients of 1, and a variable alpha coefficient of 1 to 60) had an approximately normally distributed sample distribution. Central Limit Theoremhas found application in both life and non-life insurance to determine the aggregate amount of damage (Blanchet & Lam, 2013; Restrepo-Morales & MedinaHurtado, 2012).

2. RESEARCH METHODOLOGY

2.1. DATA

The paper uses primary data on health insurance costs in the Republic of Srpska in 2017. Data refer to the costs of the Republic of Srpska Health Insurance Fund for financing health services covered by compulsory health insurance. For the analysis, all health insurance costs (family medicine, outpatient treatment, hospital treatment, prescription treatment, etc.) that are available at the Health Insurance Fund information system are collected at the ID number level. Data were collected at the individual service level, accounting for 80% of total health insurance costs. About 30,000,000 individual services were processed and thus 80% of total health insurance costs were allocated to insured. The remaining 20% is due to the unavailability of data in proportion to the previous 80%.

2.2. COMPARISON OF STATISTICAL TESTS

Considering the formula for the t-test assuming that the sample variance is not the same from equation (1), we can conclude that the t-test statistic is a ratio of the difference between the means and the standard error of the difference between the mean. In other words, the test statistic represents the ratio between the mean and the standard deviation of the sample distribution for the difference between the means. It follows that the distribution of t-test statistics depends on the shape of the distribution of the sample distribution for the difference between the means.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (1)$$

Therefore, the t-test is adequate if and only if the sample distribution of the difference between the means has a distribution that approximates the t distribution, that is, the normal distribution on large samples. Therefore, the sample size for which the sample distribution will have a distribution close to the normal should be examined

To determine the sample size at which the distribution of health insurance costs becomes closer to normal, it is necessary to determine the sample distribution of health insurance costs for a given sample size. A convenient method for determining the distribution is to simulate with the help of randomly generated numbers of so-called Monte Carlo simulation if random values are generated based on theoretical distribution. Otherwise, when empirical distribution data is available (based

on a large sample) sampling based on random number generation is called bootstrap sampling, which by its nature, is a simulation. With the help of a random number generator, one number from 1 to n is taken (the number of available values to the sample that reflects the distribution), and the value of the cost is read for the given ordinal number. One sample unit was thus simulated. Other values of the sample are then simulated in the same way (for example, for a sample distribution of a sample of size 30, it is necessary to simulate a value for 30 units). For one simulated sample, the mean is calculated. This simulation procedure should be performed as many times as possible to increase the reliability of the simulation. For this work, the simulation will be performed for 10,000 samples.

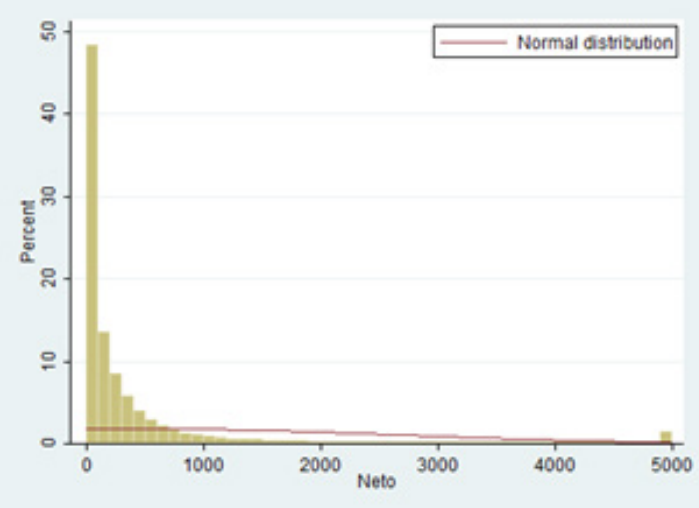
To show the effect of the Central Limit Theorem, or to examine the approximation of the sample distribution to the normal distribution, as the sample increases, more bootstrap simulations will be performed in this paper. The bootstrap simulation will be shown for the sample distribution of one mean for a sample of sizes 30, 500, and 4,000, as well as for the sample distribution of the difference between the means for a sample size of 4,000.

Based on the sampling distribution, the actual level of risk of the first species will be determined and then a comparison with a nominal level of 0.05 will be made. For the Mann - Whitney test, 10 repetitions of the test will be performed for two samples with the same mean and different second, third and fourth distribution points. Since the starting point is the null hypothesis that the means are equal, it will be tested whether the Mann - Whitney test allows for adequate testing of differences in means.

3. RESULTS

When it comes to health insurance costs, costs are insured asymmetrically - more than 25% of the insured have a cost of 0, while the maximum cost exceeds 300,000 KM.

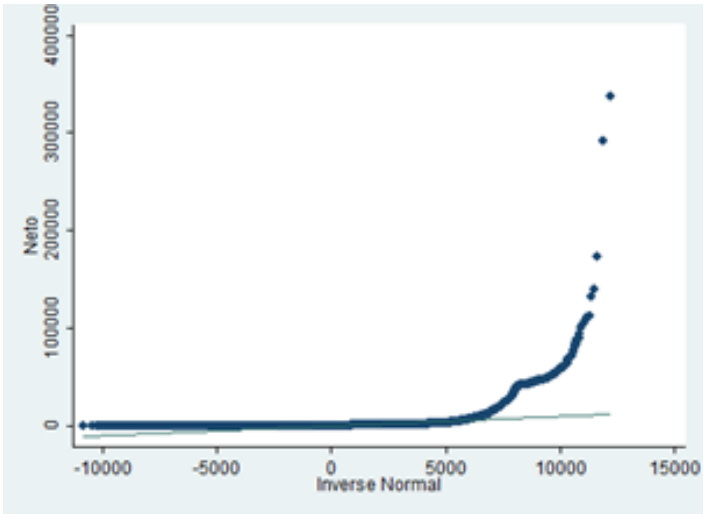
Figure 1. Distribution of health insurance costs



Source: Authors

The mean for health insurance costs is 523.96KM, the standard deviation is 2.144. The skewness coefficient is 20.52, and the kurtosis coefficient is 1.249. The foregoing indicates that the distribution is remarkably different from the normal, which can also be inferred from Figure 1 values greater than 5,000 replaced with a value of 5,000 to give a better histogram of the distribution of health insurance costs over the interval 0-5,000.

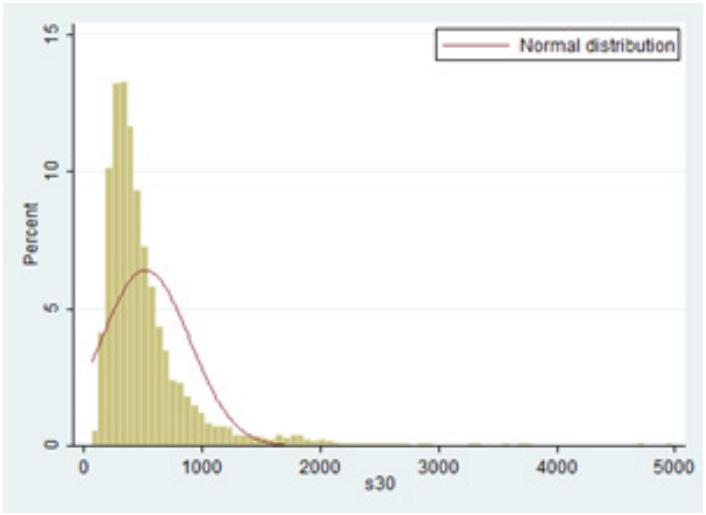
Figure 2. Q-Q diagram of health insurance cost distribution vs normal distribution



Source: Authors

From Figure 2 it is clear that the distribution deviates from the normal one, as indicated by the Kolmogorov - Smirnov test with a p-value of less than 0.001.
The sample distribution with samples of size 30 is described by mean of 521.30 which approximates the standard deviation of the original distribution, the standard error of 377.39, the skewness coefficient is 2.72, and the skewness coefficient is 14.65.

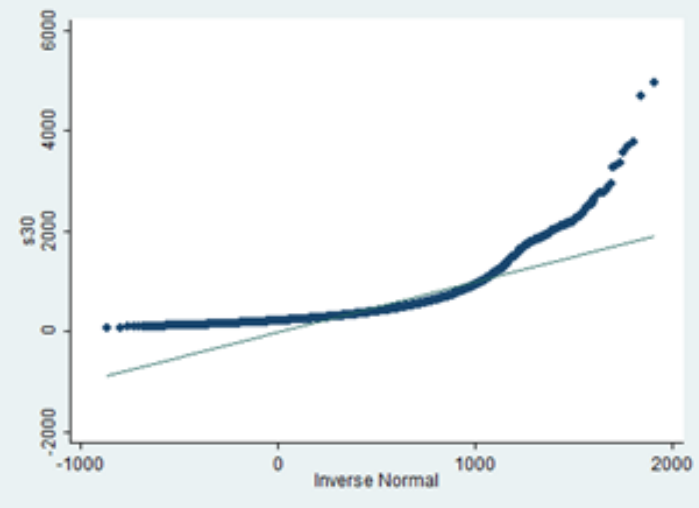
Figure 3. Sample distribution of health insurance costs for sample size 30 (n = 30)



Source: Authors

Based on Figure 3, as well as the statistics, it can be concluded that a sample of size 30 is not sufficient to achieve the action of the law of large numbers, that is, to make the sample distribution approximately normal. Such an outcome could be expected as a consequence of the distribution of health insurance costs in Figure 1, which shows exceptional skewness, which means that the sample needs to be increased significantly more to bring the distribution closer to normal. Jarque - Bera and Kolmogorov - Smirnov test normality tests hypothesize the normality of this distribution at risk of 0.0001.
Figure 3 one can see the approximation of this distribution to the normal distribution. Figure 4 also shows the approximation of the sample distribution to the normal distribution, but the deviation is clearly still present.

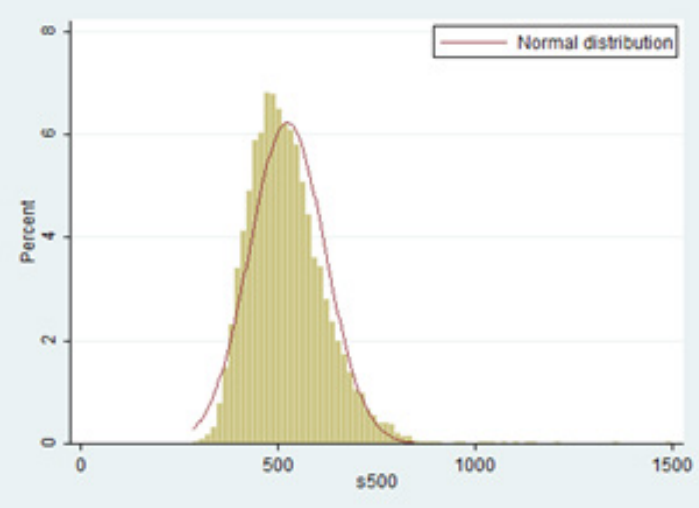
Figure 4. Q-Q diagram of the sample distribution on a sample of size 30 relative to the normal distribution



Source: Authors

To determine the adequate sample size for the application of parametric techniques, the simulation was also performed for a sample of size 500, this implies that for 10,000 simulations it is necessary to simulate 5,000,000 values

Figure 5. Sample distribution of health insurance costs for sample size 500 (n = 500)



Source: Authors

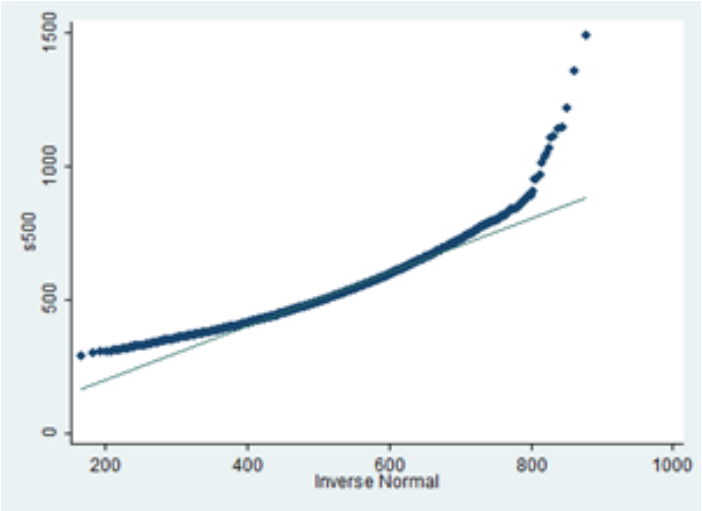
The mean of this distribution is 522.69; the standard deviation / standard error is 95.78; the skewness coefficient is 0.96, and the kurtosis coefficient is 5.99. Due to the still-present positive skewness, 2.5% of the value is less than the standard deviation z = -1.57, instead of -1.96; while 2.5% of the value is greater than the standard deviation of 2.24; instead of 1.96. Figure 5 confirms that the skewness is significantly reduced compared to the sample size 30.

Figure 6 shows that there is still a discrepancy between the quantiles of the sample distribution for samples of size 500 and the normal distribution. These results are also confirmed by normality tests, which say that with a risk of less than 0.0001 it is possible to reject the hypothesis that this sample distribution is equal to normal.

For a standardized deviation value of -1.57, it is really 2.5% of the smaller values, while a normal distribution would suggest that it is 5.78% smaller. The standard deviation value of 2.24 is larger,

according to the sample distribution, 2.5%, and a normal distribution would suggest that 1.23% are larger. Based on the previous one, we can conclude that the error in p-value at positive deviation is relatively small, but nevertheless, it is significant at negative deviation.

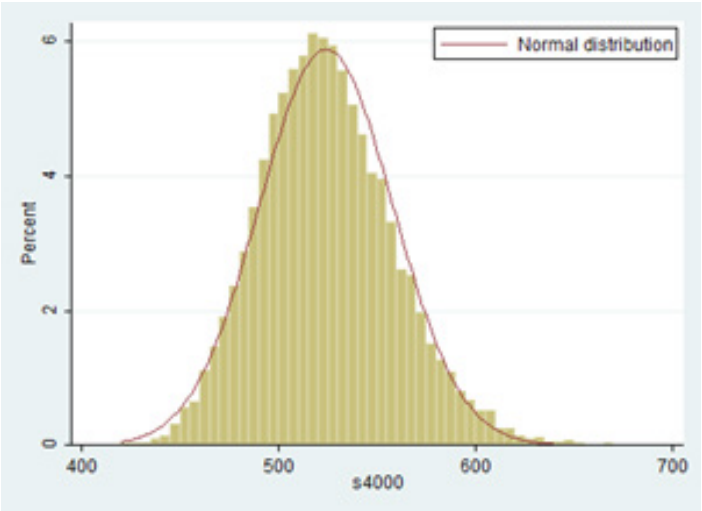
Figure 6. Q-Q diagram of the sample distribution on a sample of size 500 relative to the normal distribution



Source: Authors

Figure 7 shows that the distribution of 4,000 achieves a smoothness of distribution with still slight skewness to the right. This distribution is characterized by a mean of 523,96; standard deviation / standard error 33.92; the skewness coefficient is reduced to 0.34; and the kurtosis coefficient at 3.17. The skewness and kurtosis coefficients are now very close to the values 0 and 3, respectively, which characterize the normal distribution. However, both normality tests (Jarque - Bera and Kolmogorov - Smirnov) reject the hypothesis of distribution normality. However, it should be borne in mind that the number of simulations/replicates is relatively large - 10,000 and that-tests applied to such large samples have high power and they are able to identify small deviations from the normal distribution, and those small deviations do not have to have a major impact on statistical inference.

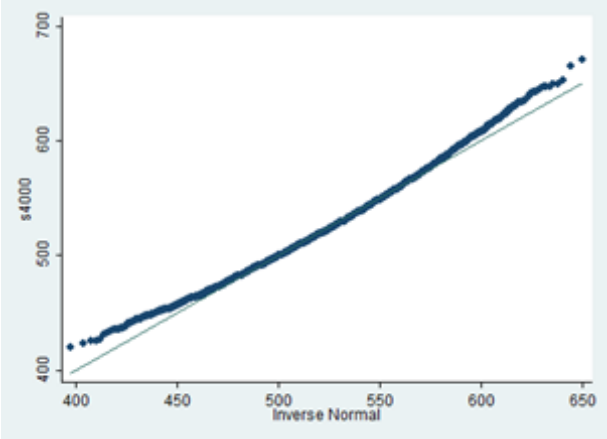
Figure 7. Sample distribution of health insurance costs for sample size 4,000 (n = 4,000)



Source: Authors

Analyzing the tails of the distribution, we find that 2.5% of the value is less than the standard deviation of 1.80 (by a normal distribution of 3.6%) and that 2.5% of the value is larger than the standard deviation of 2.15 (by the normal distribution of 1.5 %). The above points out that, with a negative deviation, the p-value would be overestimated - for a real standardized deviation of 1.8 and p-value of 0.05 (double value of 0.025 due to two-sided inference), the estimated p-value based on the normal distribution would be overestimated at the level of 0.07 (double value of 0.035). Here, we observe that the error in the p-value is relatively reduced, and taking a lower level of p-value would ensure that the risk of the first type error is not too large (resulting in a decrease in the power of the statistical test, that is, an increase in the error of the second type).
Figure 8 shows that the quantile degree of empirical distribution and normal distribution is very high.

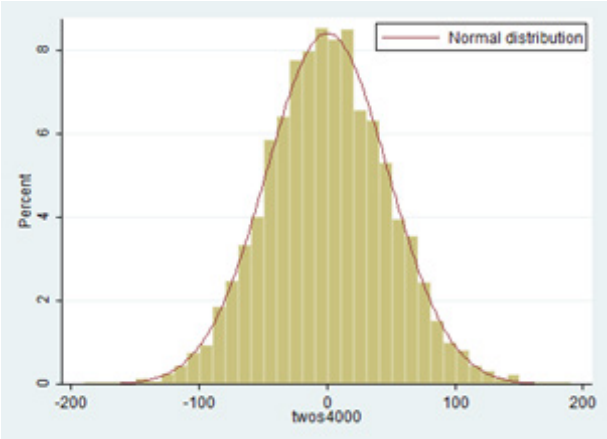
Figure 8. Q-Q diagram of the sample distribution on a sample of 4.000 in relation to the normal distribution



Source: Authors

However, when applying the two-sample t-test, the sample distribution of only one sample is not relevant, but rather the relevant sample distribution is the difference in means for the two samples. In order to derive the sample distribution of the differences of 10,000 means of the samples in Figure 7 are divided into two groups of 5,000, between these two groups the differences of means are calculated, and the distribution of 5,000 differences in means is obtained.

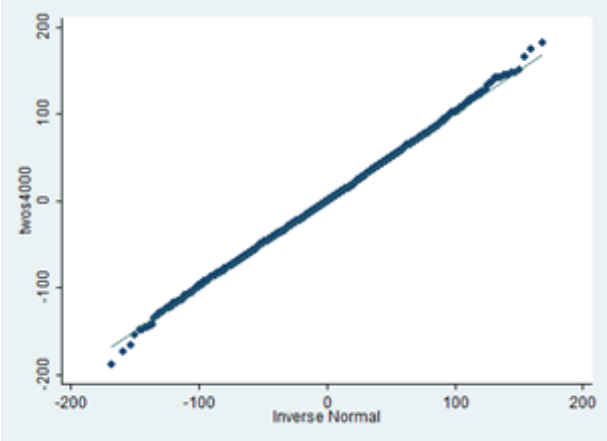
Figure 9. Sample distribution of differences between two groups of samples, where in both groups the samples are 4,000 (n = 4,000)



Source: Authors

Mean of distribution from Figure 4.10. is 0.28 which is approximately 0 given that the samples are independently generated, so the limes of the mean when the sample size tends to infinity is 0. The standard deviation / standard error is 47.51. The skewness coefficient is only 0.05; and the kurtosis coefficient is 3.09. It has previously indicated that this distribution has a distribution that is very close to the normal distribution.
Of the standard deviation value -1.92, 2.5% is smaller (2.6% according to the normal distribution), while 2.5% is larger than the standard deviation 2.01 (2.2% normal distribution). Which indicates that the true p value at negative deviations of 0.05 would be estimated at 0.052 with the help of the t-test. With positive deviations, the rating would be 0.044.
According to both tests of normality, the distribution of Figure 4.10 does not deviate from normal. Jarque - Bera shows a p value of 0.1513, and Kolmogorov - Smirnov 0.134.

Figure 10. Q-Q diagram of the sample distribution of differences of means, on samples of size 4,000 relative to the normal distribution



Source: Authors

Figure 10 shows a high degree of agreement between the quantiles of empirical and normal distribution.
Considering the statistical moments, deviation in p values, normality tests, and quantile agreement, we can conclude that the assumption of normality of distribution on the sample distribution of differences in means - is fulfilled.
If the population is divided into groups (based on sex, age, etc.), heterogeneous groups can be obtained (in terms of statistical moments), but on large samples (4,000 and more) this variability should not significantly compromise the quality of conclusions based on t-test, which will be analyzed below.
To evaluate the effect of skewness and kurtosis on arithmetic inference using the Wilcoxon - Mann - Whitney test, health insurance cost data analyzed the difference in average costs between men and women of all ages, on a sample size of 4,000. The average value of health insurance costs for men in 2017 was 536.03KM and for women was 536.37KM. Given that the standard deviation in males is 2.357 and in females 1.988 even with samples of 481.253 and 535.934, respectively, the standard error of both samples is greater than 2.5 (while the difference in means is 0.31) we can with high certainty conclude that this difference is not statistically significant. The cost skewness in men is 19.18 and in women 21.49. The prevalence in males is 1,125 and in females 1,309. From the previous statistical moments, we can conclude that the means are equal in gender (which is confirmed by the bootstrap method) and that the other moments are different between the sexes. Generation of 4,000 samples for both sexes and the calculation of the Wilcoxon - Mann - Whitney test were applied on such data. This procedure was repeated 10 times. In all ten replicates, the Wilcoxon - Mann - Whitney test rejected the null hypothesis at a risk less than 0.000000001 in all

cases, ie with a t statistic greater than 8 at absolute value. From the above, we can see that Wilcoxon - Mann - Whitney rejects the hypothesis of equality of distributions for both sexes, although the difference in means is not statistically significant. Thus, we can estimate that Wilcoxon - Mann - Whitney is not suitable for indirectly inferring the difference in means for two samples, in case of pronounced difference in skewness and kurtosis.

4. DISCUSSION

There is a widespread belief that the t-test and linear regression apply only in the case of variables with a normal distribution, however, according to the Central Limit Theorem, they can be applied to large samples of variables with data deviating from the normal distribution. As evidenced by the simulation of health cost data (Lumley, Diehr, Emerson & Chen, 2002). Fagerland&Sandvik (2009b) found that when the sample size is sufficiently increased, beyond 200, the t-test becomes robust even for pronounced asymmetric distributions.
The median sample size of empirical research published in The Lancet and BMJ journals increased from 33 and 37 from 1972 to 2007, respectively, to 3116 and 3104. Meanwhile, the incidence of Wilcoxon - Mann - Whitney testing increased from 11 to 27%, at the expense of the t-test whose use decreased from 44 to 26%, which is a paradox. One reason is that journals such as NEJM in the instructions to the authors suggest that non-parametric tests be applied to data with a distribution that deviates from normal. However, such guidance directs large-sample research authors to apply nonparametric tests unnecessarily (Fagerland, 2012).

CONCLUSION

As same to the normal distribution and parametric tests that were popular at the beginning of the 20th century, nowadays, there is increased usage of non-parametric tests and avoiding of parametric tests. For the distribution of the t-statistic, more relevant is the sampling distribution, rather than the distribution of sample data. More precisely, there is a relevant sampling distribution of two means difference. The normality of sampling distribution is a condition for t-test applications. Therefore, the normality of distribution may be overrated.
The paper was presented in the case of health insurance cost data. Health insurance costs have a right-skewed distribution with high peakedness. Following the Central Limit Theorem, as sample size increases, sampling distribution approaches to normal. For a sample size of 4,000, the sampling distribution of health insurance cost data is close to the normal distribution. But, the Jarque-Bera test of normality finds enough evidence to reject the null hypothesis of normality. As we have already said, we have to look at the sampling distribution for two means difference. For the size of 4,000, the sampling distribution for two means difference is almost normal. The Jarque-Bera test didn't find enough evidence to reject the null hypothesis. Therefore, we can conclude that the t-test is applicable for analysis in this case. This doesn't mean that a t-test is always applicable to health insurance cost data, or skewed distributions in general. Firstly, it is necessary to check if the sample size is enough that sampling distribution for two means different approaches to the normal distribution. On the other hand, there has been showed that the Mann-Whitney test could be sensitive to skewness differences between samples. It leads to a gap between the actual risk level of the first kind and nominal one.
The results of this paper claim that the t-test has better performance than the Mann-Whitney test, in an analysis of the difference between two groups of health insurance cost data if samples are large enough. This conclusion is still applicable to any other field where skewed distribution could be found. Further research could be oriented to techniques for checking if the sample size is enough large for the t-test application.

AKNOWLEDGEMENTS

The authors express their gratitude to the Health Insurance Fund of Republic of Srpska for giving data for the analysis.

REFERENCES

[1] Blanchet, J., & Lam, H. (2013). *A heavy traffic approach to modeling large life insurance portfolios.* Insurance: Mathematics and Economics, 53(1), 237-251. https://doi.org/10.1016/j.insmatheco.2013.04.011

[2] Chang, H. J., Wu, C. H.; Ho. J. F., Chen P. Y. (2008). *On sample size for using Central limit theory in various distributions,* International Journal of Information and Management Sciences, 19 (1).

[3] Cundill, B., & Alexander, N. D. (2015). *Sample size calculations for skewed distributions.* BMC medical research methodology, 15(1), 28. https://doi.org/10.1186/s12874-015-0023-0

[4] Fagerland, M. W. (2012). *t-tests, non-parametric tests, and large studies—a paradox of statistical practice?.* BMC Medical Research Methodology, 12(1), 78.https://doi.org/10.1186/1471-2288-12-78

[5] Fagerland, M. W., & Sandvik, L. (2009). *Performance of five two-sample location tests for skewed distributions with unequal variances.* Contemporary clinical trials, 30(5), 490-496. https://doi.org/10.1016/j.cct.2009.06.007

[6] Fagerland, M. W., & Sandvik, L. (2009). *The wilcoxon–mann–whitney test under scrutiny.* Statistics in medicine, 28(10), 1487-1497. https://doi.org/10.1002/sim.3561

[7] Fay, M. P., & Proschan, M. A. (2010). *Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules.* Statistics surveys, 4, 1. https://doi.org/10.1214/09-ss051

[8] Ловрић, М.,Комић, Ј., & Стевић, С. (2006). *Статистичкаанализа: методи и примјена.* Економски факултет Бања Лука, Бања Лука,(330-357).

[9] Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). *The importance of the normality assumption in large public health data sets.* Annual review of public health, 23(1), 151-169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

[10] Mann, H. B., & Whitney, D. R. (1947). *On a test of whether one of two random variables is stochastically larger than the other.* The annals of mathematical statistics, 50-60. https://doi.org/10.1214/aoms/1177730491

[11] Pandey, R. M. (2015). *Commonly used t-tests in medical research.* Journal of the Practice of Cardiovascular Sciences, 1(2), 185. https://doi.org/10.4103/2395-5414.166321

[12] Restrepo-Morales, J. A., & Medina Hurtado, S. (2012). *Estimation of operative risk for fraud in the car insurance industry.* Global Journal of Business Research, 6(3), 73-83. https://doi.org/10.1002/9781118387047.ch6