TREATMENT OF THE DATA FILES WHICH DO NOT MEET BASIC CONDITIONS OF SELECTED STATISTICAL METHOD

Vladimíra Hovorková Valentová – Kateřina Gurinová

Abstract

The authors of the paper try to point out that there is one important problem in the basic statistical methods teaching at faculties of economics. The basic courses of statistics usually contain just knowledge about the statistical methods which use is connected with some requirements, especially with the data normality or homoscedasticity etc. But many students are surprised when they work with real data files and they find out that the data have "unsuitable" characteristics in relation to the requirements of selected methods. The aim of this paper is to present possible ways in such cases when some conditions of selected methods cannot be satisfied. We focused on the methods which are usually taught in faculties of economics in the frame of two-semester course of statistics. We mean the parametric hypotheses testing and analysis of variance above all and other methods. At the same time, we also discuss the assumptions of the methods use which can replace the basic methods. The truth is that these substitute methods are also connected with some assumptions.

Key words:analysis of variance, homoscedasticity, Kruskal-Wallis test, Mann-Whitney U test, normal distribution

JEL Code: C18, C81

Introduction

Application of many often used statistical methods is possible when some assumptions are fulfilled. It means that some methods were derived under some basic assumptions which the treated data files should have. One of very often occurring assumption is the normality of the frequency distribution. Another one is homoscedasticity or sample size determination, absence of outliers and absence of correlation among sample units (Meloun, Militký 2004).

Before carrying out of a statistical analysis it is necessary to explore the data set and look for its possible special features and to check existing conditions to be able to use the correct statistical methods for treating them then.The exploratory data analysis is a valuable

tool for it. It can provide much important information about the data file, alert on potential problems and be helpful when choosing the most suitable method for the data analysis. In fact, the incorrect realization of the analysis can totally depreciate the whole process outcomes.

The exploratory data analysis is based on graphical methods in a large extent. Their most valuable benefit is their clearness. Their biggest disadvantage is impossibility to decide exactly about rejection or non-rejection of some hypothesis in comparison with statistical tests. The analyst by himself has to determine how big the distance between the theoretical and empirical distribution is. On the other side, the methods of exploratory data analysis allow discover sources of the difference between the empirical and theoretical distribution (e. g. influence of a skewness, kurtosis, outliers and it is also possible to detect a mixture of various distributions etc.). Some authors assert that graphical methods are more sensitive than usually used tests which detection ability depends on power of a test $(1-\beta)$ (Meloun, Militký 2004). The best way how to explore the data is to combine the exploratory data analysis methods and numerical methods, e. g. tests of normality, tests of independence etc.

1 The Data Assumptions Verification

It is possible to assert that classical statistical analysis is based on the assumption of empirical distribution normality. It is necessary to realize that considerable departure from normality can disallow using of e.g. arithmetic mean as the estimator of expected value of the population, using of standard procedures for interval estimation etc. As was mentioned above, the commonly used methods of statistical inference also require independence of statistical units, sample homoscedasticity and appropriate sample size for a calculation of point or interval estimators with determined reliability and maximum error of estimation.

Therefore, the data assumption verification is extremely important activity which should be done before the analysis. In case that one or more assumptions violation is detected, it is necessary to use other statistical procedures which allow treat the data files correctly. But the situation is often complicated by the fact that some methods used for the data assumption verification have their own assumptions of use. And the violation of these assumptions can have a negative impact on the process. Therefore, it is necessary to combine the methods and ensure the logical sequence of the procedures.

The assumption of the distribution normality can be verified either by the graphical methods or by tests of normality. The most popular graphical methods are e. g. frequency

histogram, quantile-quantile plot and the density trace. There are many tests of normality, e. g. Shapiro-Wilk test (Shapiro, Wilk, 1965), chi-square test, D´Agostino´s K-squared test (D´Agostino, 1971), Anderson-Darling test, Jarque-Bera test (Jarque, Bera, 1987), Komogorov-Smirnov test and others.

While testing normality, it is necessary to take into account another factor – the sample size. The problems arise in case of small sample sizes (less than 10) and large ones (more than 2000). Tests of normality for small sample sizes show a low power of a test. It means that in case when the null hypothesis is rejected, we can be quite sure that the data do not come from the normal distribution. If the test does not lead to the null hypothesis rejection, we do not have enough proofs to be able to reject it. When we test the normality in large samples, even small departures from the normality are considered to be statistically significant because of a high power of a test. In such cases we can recommend using graphical methods.

The normality assumption verification is often connected with an activity which leads to outliers identification. Some graphical method is usable for the preliminary outliers identification, e. g. box plot, symmetry plot, Q-Q plot (Meloun, Militký, 2004). When testing outliers, it is possible to use e. g. Grubbs´ test for outliers or Dixon´s Q test (Dixon, 1950). When some outlier is identified, it is suitable to use some of the robust methods for the data analysis because these methods are not responsive to outliers. Most of them are based on percentiles.

The sample units independence is another important assumption of the data of a high quality. We can test it with the help of e. g. von Neumann test or Wald test (Meloun, Militký, 2004). For the time independence of sample units check can be used e. g. Durbin-Watson test or sign test etc.

The testing of homoscedasticity is possible with the help of e. g. Bartlett´s test (Bartlett, 1937), Levene´s test (Gastwirth, Gel, Miao, 2009), Cochran´s C test or Hartley´s test. Especially the homoscedasticity assumption is very often occurring requirement for using some methods. Analysis of variance is best known one. Commonly used Bartlett´s test is quite sensitive to the normality assumption violation which appears in case of small sample sizes. In such situations it is better to use Levene´s test which is robust.

Results coming from the analysis of a small sample are always affected with a large amount of measurement uncertainty. Therefore, it is suitable use such data just in necessary cases when it is not possible to reach larger sample size. For small samples (from 4 to 20

units) we can use Horn procedure (Horn, 1983) to the measures of central tendency and variation estimations.

2 What to Do when the Data Normality Assumption Is Broken?

If a great departure from normality is detected in a data set, a problem with the choice of statistically correct procedure for data analysis arises. One of the possibilities is using nonparametric methods. Another possibility is transformation of the data in such way to become normally distributed or to be very close to normal distribution.

2. 1 Nonparametric Methods

The choice of a procedure depends on various factors. Firstly, it is necessary to choose a suitable nonparametric method for a certain situation because we have usually more possibilities what to do.

When we investigate independence of numeric variables with the help of standard analysis of variance, the normality assumption is crucial. If it is not possible to consider the data set to be normally distributed, we can apply nonparametric possibility which is Kruskal-Wallis test (Kruskal, Wallis, 1952).It is a nonparametric test for a simple sorting and its usage is possible when we consider both balanced design and unbalanced design. When the test leads to the null hypothesis rejection (the null hypothesis assumes equality of medians in all the groups), we can recommend Nemenyi test for balanced designs (Marcinko, 2014).

When the normality assumption is violated in case of the hypothesis test for μ, it is suitable to use the sign test or Wilcoxon signed-rank test. In case of comparison of two means when we consider independent samples, we can apply Wilcoxon signed-rank test or Mann-Whitney U test (Wilcoxon, 1945) instead of parametric test for $\mu_1 - \mu_2$. If we compare paired samples and the normality assumption is not held, we can use the sign test, Wilcoxon signed-rank test or McNemar´s test.

2. 2 Data Transformation

In case when we decide to transform data if the assumption of data normality was violated, it is necessary to make some of the non-linear transformations. Many tools for the data transformation exist – e. g. logarithmic transformation, power function transformation, exponential transformation or Box-Cox transformation (Box, Cox, 1964). Their aim is to approximate the data to normality as much as possible.

These transformations are especially suitable for asymmetric unimodal distributions. They come from the hypothesis that the data are non-linear transformation of the random variable which is normally distributed. Correct transformation leads to the variance stabilization, approximation to normality or sometimes to normality of the data distribution (Meloun, Militký, 2004).

3 What to Do when the Other Assumptions Are Violated?

As was written above, the assumption of data normality is not the only one requirement but there are many other assumptions. But we will focus on some of them only in agreement with the aim of this paper. We mean the methods which are included in basic courses of statistics in faculties of economics.

Parametric analysis of variance requires not only the normally distributed data set but also the assumption of homoscedasticity must be held. When we prove heteroscedasticity in a data set, it is possible to use as same procedure as in case of not normally distributed data – we can recommend Kruskal-Wallis test, more is mentioned e. g. in (Kruskal, Wallis, 1952). If the random samples are not independent, we can use e. g. Friedman´s test (more in Friedman, 1937). It is useful for testing equality of population medians. If the null hypothesis is rejected, it is possible to use Friedman´s test also for post-hoc analysis.

If we use Student´s t-test for independent samples, the assumption of normality and homoscedasticity must be held. If the assumption of normality is not rejected but heteroscedasticity was proved, we cannot use Student´s t-test of the equality of population means. It is suitable to apply e. g. Welch´s t-test which is insensitive to equality of variances, more in (Welch, 1951).

The chi-square test of independence in contingency table is another often used method. Before using it, it is necessary to verify assumptions about sample size, respectively about size of expected frequencies in each group. If the size of expected frequency in a group is less than the requirement, it is possible to merge neighbour columns or rows of a contingency table but this activity has to be logical not mechanical. It can happen that the merging of columns or rows does not lead to solving the problem with a small size of expected frequency in some groups. Then it is not possible to test independence of following variables in this way.

The chi-square test also belongs to the group of goodness of fit tests. This test requires a large sample size. The sample size is large enough when expected frequencies in each group

are equal or greater than 5.Or some authors prefer softer rule – see Table 1. If this requirement is not held, we can merge the groups or use Kolmogorov-Smirnov test which is suitable also for small sample sizes.

Tab. 1: Summary of procedures suitable in cases of violation of selected methods assumptions

Method	Assumptions	Verification	Procedure when the assumptions are not held
Analysis of variance	Normality	Shapiro-Wilk test, D´Agostino´s K-squared test, Anderson-Darling test etc.	Kruskal-Wallis test
	Homoscedasticity	Bartlett´s test, Levene´s test	Kruskal-Wallis test
	Independence	x	Friedman´s test
Test for μ	Normality	Shapiro-Wilk test, D´Agostino´s K-squared test, Anderson-Darling test etc.	Sign test, Wilcoxon signed-rank test
Test for $\mu_1 - \mu_2$ (independent samples)	Normality	Shapiro-Wilk test, D´Agostino´s K-squared test, Anderson-Darling test etc.	Mann-Whitney U test
	Homoscedasticity	Bartlett´s test, Levene´s test	Welch´s t- test
Test for $\mu_1 - \mu_2$ (paired samples)	Normality	Shapiro-Wilk test, D´Agostino´s K-squared test, Anderson-Darling test etc.	Sign test, Wilcoxon signed-rank test, McNemar´s test
Chi-square test of independence	Sufficient size of expected frequencies	All expected frequencies are greater than 1 and 80 % of expected frequencies are equal or greater than 5 (soft assumption).	Merging of similar groups
Goodness of fit test – chi-square test	Sufficient size of expected frequencies	All expected frequencies are greater than 1 and 80 % of expected frequencies are equal or	Merging of similar groups, Kolmogorov-Smirnov test

		greater than 5 (soft assumption).	

Source: own

Conclusion

Basic courses of statistics in faculties of economics include almost commonly used statistical methods which require some assumptions to be held. In practice there is no exception that the data do not meet assumptions required which can be a problem for writing diploma works. In case when students do not know suitable method for their data analysis, they usually use the method which they know without reference to the assumptions violation. It is naturally problem from statistical point of view.

In this paper the summary of methods useful when some assumption of commonly used methods are violated is presented. We mentioned the assumptions of each method, ways how to test the assumptions and possible ways of solving the problem of the assumption violation.

It is obvious that it is more suitable to use parametric tests then nonparametric ones because they have bigger power function. In cases when these parametric tests are not usable from the reason of their assumptions violation, it is necessary to apply nonparametric ones. We can find out that their power function declines from the reason of information loss which results from the replacement of original data with ranks. Their advantage is their insensitivity on the shape of the distribution.

Acknowledgment

This paper was written in the frame of work on the MŠMT project, OP Education for Competitiveness, section of the support 2.3, called Innovation of the Study Programme Economics and Management with the Focus on Knowledge Economics, No. CZ.1.07/2.200/28.0317.

References

Bartlett, M. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, 160*(901), 268-282. Retrieved April 1, 2015, from http://rspa.royalsocietypublishing.org/content/royprsa/160/901/268.full.pdf.

Box, G., & Cox, D. (1964). An Analysis of Transformation. *Journal of the Royal Statistical Society, 26*(2), 211-252. Retrieved April 2, 2015, from http://fisher.osu.edu/~schroeder.9/AMIS900/Box1964.pdf.

D'Agostino, R. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika,58*(2), 341-348. Retrieved April 9, 2015, from http://webspace.ship.edu/pgmarr/Geo441/Readings/D'Agostino 1971 - An Omnibus Test of Normality for Moderate and Large Size Samples.pdf.

Dixon, W. (1950). Analysis of Extreme Values. *The Annals of Mathematical Statistics,21*(4), 488-506. Retrieved April 14, 2015, from http://projecteuclid.org/download/pdf_1/euclid.aoms/1177729747

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association,32*(200), 675-701. Retrieved April 18, 2015, from http://sci2s.ugr.es/keel/pdf/algorithm/articulo/1937-JSTOR-Friedman.pdf.

Gastwirth, J., Gel, Y., &Miao, W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statistical Science,24*(3), 343-360. Retrieved April 11, 2015, from http://projecteuclid.org/download/pdfview_1/euclid.ss/1270041260.

Horn, P. (1983). Some Easy t Statistics. *Journal of Americal Statistical Association,78*(384), 930-936. Retrieved April 4, 2015, from http://www.jstor.org/discover/10.2307/2288206?uid=383234071&uid=3737856&uid=2&uid=3&uid=67&uid=382743071&uid=62&sid=21106520856883.

Jarque, C., & Bera, A. (1987). A test for normality of observation and regression residuals. *International Statistical Review,55*(2), 163-172. Retrieved April 12, 2015, from http://webspace.ship.edu/pgmarr/Geo441/Readings/Jarque and Bera 1987 - A Test for Normality ofObservations and Regression Residuals.pdf.

Kruskal, W., &Wallis, W. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association,47*(260), 583-621. Retrieved April 11, 2015, from http://homepages.ucalgary.ca/~jefox/Kruskal and Wallis 1952.pdf.

Marcinko, T. (2014). Consequences of Assumption Violations Regarding One-Way Anova. *The 8th International Days Of Statistics and Economics,* 2014, Prague, Czech Republic, 974-985. Retrieved April 2, 2015, from http://msed.vse.cz/msed_2014/article/342-Marcinko-Tomas-paper.pdf.

Meloun, M., &Militký, J. (2004). *Statistická analýza experimentálních dat* (2nd ed., p. 953). Praha, Czech Republic: Academia.

Shapiro, S., &Wilk, M. (1965). Ananalysis of variance test for normality (complete samples). *Biometrika,52*(3/4), 591-611. Retrieved April 10, 2015, from http://estadisticaccbas.uaa.mx/moodle/file.php/1/Lecturas/shapiro1965.pdf.

Welch, B. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika,38*(3/4), 330-336. Retrieved April 21, 2015, from http://www.soph.uab.edu/Statgenetics/People/MBeasley/Courses/Welch1951.pdf.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin,1*(1), 80-83. Retrieved April 16, 2015, from http://sci2s.ugr.es/keel/pdf/algorithm/articulo/wilcoxon1945.pdf.

Contact

Vladimíra Hovorková Valentová
Technical University of Liberec
Studentská 2, 461 17 Liberec
vladimira.valentova@tul.cz

Kateřina Gurinová
Technical University of Liberec
Studentská 2, 461 17 Liberec
katerina.gurinova@tul.cz