Bulletin of the *Transilvania* University of Braşov
Series V: Economic Sciences • Vol. 10 (59) No. 2 - 2017

Using the Boxplot analysis in marketing research

Cristinel CONSTANTIN[1]

Abstract: *Taking into account the needs of decision makers inside the companies, marketing research is meant to provide the best information that really can to help the adoption of the best decisions. In this respect a lot of methods of data analysis can be used but the researcher has to choose those results that minimize the errors. This paper proposes an instrumental research regarding the using of Boxplot analysis to identify certain outliers that can alter the information. In order to attain the objective of this research, we exemplified the Boxplot analysis on data related to the GDP recorded in 2014 by Romanian counties.*

Key-words: *Boxplot analysis, marketing research, outlier identification, GDP analysis, instrumental research, extreme values analysis*

1. Introduction

Data analysis is a one of the most important part of every research as far as the models used have to provide useful information for decision makers. A deep analysis using different tools can help researchers to compare the results and to avoid the errors that could be encountered due to poor analysis or bad methods. Starting from this problem, in this research we emphasised the utility of using Boxplot analysis especially when data are not normally distributed. Starting from an example based on quantitative data regarding the Romanian GDP recorded at territorial level we explain the benefits of using the Boxplot method for data analysis.

2. The Boxplot analysis

Boxplot analysis is considered a very popular and easy method used for univariate data analysis when an unimodal continuous variable is computed. Such variables are measured with interval or ratio scales, which allow calculating the mean and the variation statistics (variance, standard deviation etc.). But the value of these indicators could be negatively influenced by extreme values, which are not

[1] Transilvania University of Braşov, cristinel.constantin@unitbv.ro

consistent with the majority of data, called outliers (Carter, Schwertman and Kiser, 2009). The outliers can move the mean to the left or to the right and this parameter will conduct to an erroneous interpretation of the variable's central tendency.

The method (proposed by Tukey in 1977) takes into consideration the median as central tendency indicator and the quartiles instead of sample mean. The boxplot contains a box, which is delimited in the bottom side by the first quartile (Q1) and by the third quartile (Q3) in the upper side. For the median value (Q2) is drawn a line inside this box. The points outside this box are considered potential outliers and are drawing as whiskers in both side of the box. The whiskers stretch between two fences ($f_1$and f_2) that are calculated based on the interquartile range –IQR (Hubert and Vandervieren, 2008).

$$IQR = Q_3 - Q_1 \quad (1)$$

$$f_1 = Q_1 - 1.5\ IQR \quad \text{and} \quad f_2 = Q3 + 1.5\ IQR \quad (2)$$

An example of a boxplot with all the above-mentioned parts is presented in Figure 1.

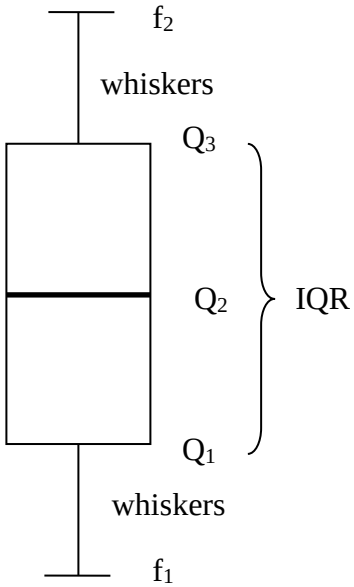

Fig. 1. *An example of standard boxplot chart*

Taking into account the values presented above, the Boxplot analysis is known as the *5-number summary*, which contains the maximum and minimum values (fences), the lower and upper quartiles (Q_1, Q_3) and the median (Q_2). If there are outliers that exceed the fences, they are represented individually by different symbols (Potter, 2006).

Boxplot analysis is also useful in bivariate analysis, as it allows comparisons between different group categories that results from population's characteristics (gender, age, income, education etc.) or other grouping variables (Zikmund and Babin, 2010). Moreover, the information systems used for data analysis (e.g. SPSS) compute very easy this method. Thus, the boxplot is considered a useful tool due to some advantages such as: the compact display of results; easy to compute and understand by people that are not specialists in statistics; possibility to compare groups of data etc. (Arroyo, Maté, Roque, 2006).

Boxplot analysis can be computed very easy by using the IBM SPSS Statistics system. The "Explore" function provides the information regarding the *5-number summary* and identifies the outliers. The system also draws the Boxplot chart, which presents graphically al all the above-mentioned values. The outliers are labeled with the numbers that indicate the case position in the database or a labeling variable can be used.

2. Example of using the Boxplot analysis

In the followings we present an example of using the Boxplot analysis for a metric variable measured with the ratio scale. This variable is the GDP recorded in 2014 by every county of Romania and the main goal is to identify the 5-number summary and the outliers. This objective rose from one of the main economic problems that reveals a poor convergence at county and regional level in Romania.

For the beginning, the values of GDP recorded in 2014 by Romanian counties (including the capital city) have been considered for the Boxplot analysis using the "Explore" function in SPSS system. Among the provided information is the table of percentiles (see Table 1).

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average	GDP	4808,8	5790,4	6737,8	10437,1	15608,0	29795,1	33706,7
Tukey's Hinges	GDP			6748,9	10437,1	15289,6		

Table 1. *The percentiles for Romanian GDP at county level in 2014 (millions Lei)*

In the above table we can see the Tukey's Hinges, which represent three of the 5-number summary: the lower and upper quartiles (Q_1, Q_3) and the median (Q_2). Based on these values we can compute the interquartile range (IQR) and the fences. The outliers could be also identified like in Figure 1.

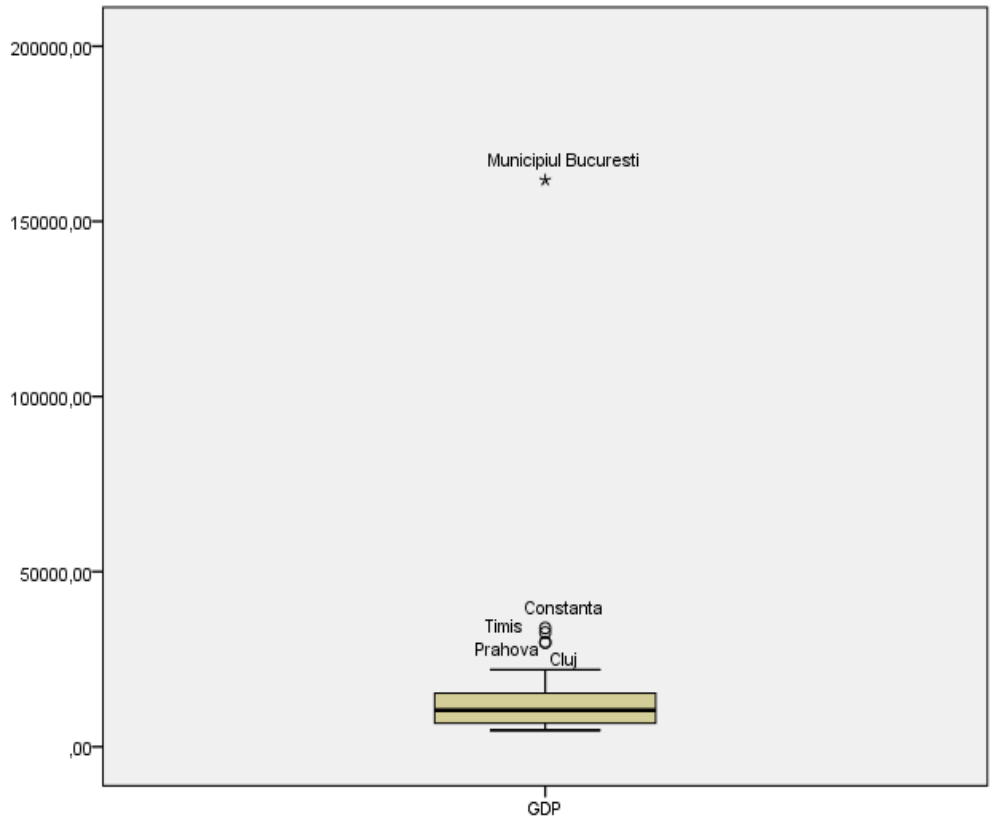

Fig. 1. *The Boxplot chart for the GDP recorded in 2014 by Romanian counties (millions Lei)*

In the above analysis, a variable named "County name" was used in "Explore" function in order to label cases. Otherwise the system identifies cases by numbers according to their position in the database. The line inside the box represents the median and the lower and upper borders are the quartiles Q1 and Q3. On the chart are also mapped the outliers that exceed the values of fences. In such a situation we can find 4 counties (Constanta, Timis, Prahova and Cluj) and the capital city (Municipiul Bucuresti).

We can observe a big difference that exists between the capital city and the rest of counties, which is beyond the "outer fence", which is calculated by adding 3xIQR to the Q3. Such cases are considered extreme outliers (Carter, Schwertman and Kiser, 2009). If we look at the capital city, the value of GDP recorded in 2014 was of 161,772.4 Lei million, which is more than 10 times higher than the value of the Q3. Thus, this case could be considered a very extreme outlier. It underlines the huge

discrepancy that exists between the capital city and the rest of counties in terms of economic development.

In the context of Boxplot analysis, such extreme outliers can determine distortions of the charts and would be better to exclude those cases from the analysed variable. Thus, after we excluded the capital city, the boxplot chart looks like the figure below.

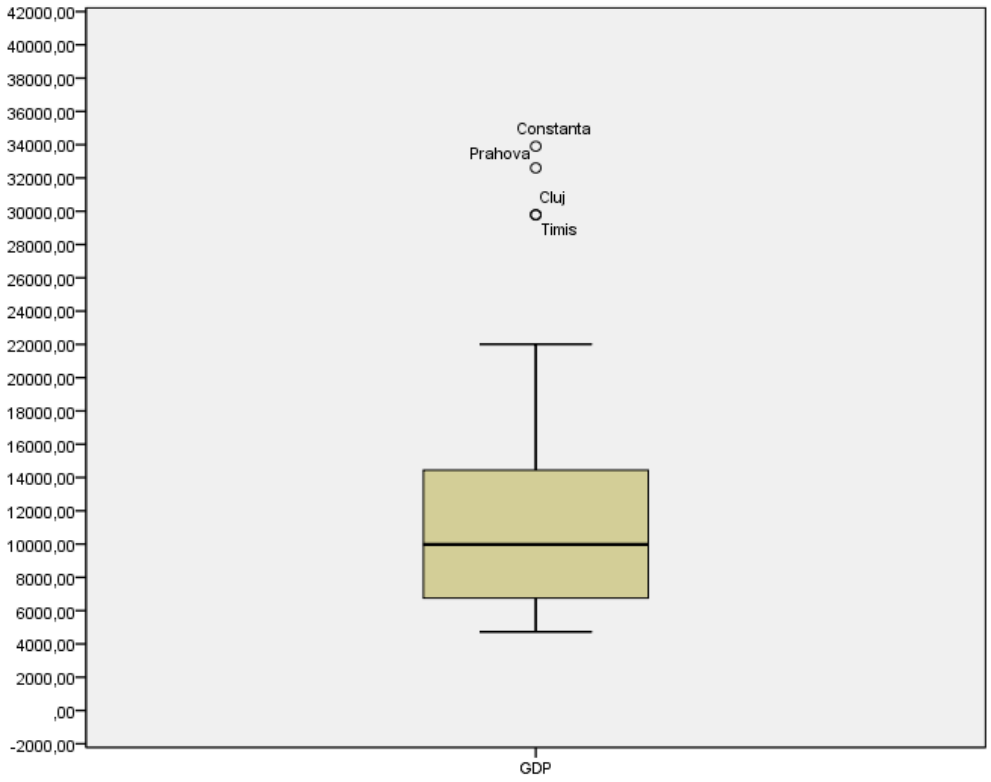

Fig. 2. *The Boxplot chart for the GDP recorded in 2014 by Romanian counties (Bucharest excluded)*

After the exclusion of the extreme outlier, two quartiles have been translated with one case towards the lowest value and 5-numbers summary is as follows:

1. Q_1= 6748.90 Lei mil. (Vrancea county)
2. Q_2 = 9980.60 Lei mil. (Gorj county)
3. Q_3 = 14445.90 Lei mil. (Mures county); IQR = Q_3 - Q_1 = 7697 Lei mil.
4. f_1 = - 4796.6 Lei mil.
5. f_2 = 25991.4 Lei mil.

It can be observed that the lower fence has a negative value due to a great dispersion of the individual values that lead to an interquartile value (IQR) higher than Q_1. To avoid such distortions, the SPSS system cuts the whiskers at the level of the minimum value, which in our case is Covasna county (4723.70 Lei mil.) The position of the median line shows also a higher dispersion of the values in the upper side of the box than in the bottom side. This heterogeneity is emphasised by the outliers that are plotted beyond the whiskers.

The "Explore" function of SPSS also offers the possibility to obtain the highest and the lowest five values by using the option "Outliers" from "Statistics". We obtained in our analyse the table below, even if the extreme values are not truly outliers.

Extreme Values

			Case Number	County	Value
GDP	Highest	1	21	Constanta	33901,50
		2	30	Prahova	32602,70
		3	3	Cluj	29805,60
		4	40	Timis	29770,50
		5	8	Brasov	22014,40
	Lowest	1	9	Covasna	4723,70
		2	34	Mehedinti	4748,30
		3	23	Tulcea	5151,60
		4	6	Salaj	5766,80
		5	26	Calarasi	5845,40

Table 2. *The counties with the highest and the lowest values of GDP in 2014 (millions Lei)*

We can identify the Brasov county among the counties with the five highest values even if its value is lower than the upper fence (f_2). The five cases with the lowest values are all between the fences so that they are not truly outliers according to the Tukey's rule.

The Boxplot analysis can also use an additional variable, called "Factor variable", in order to group the analysed values in certain categories that could be different each other. For our example, we used as factor variable the four Romanian macroregions. The counties have been associated with these macroregions according to their membership. The results are presented in Figure 3.

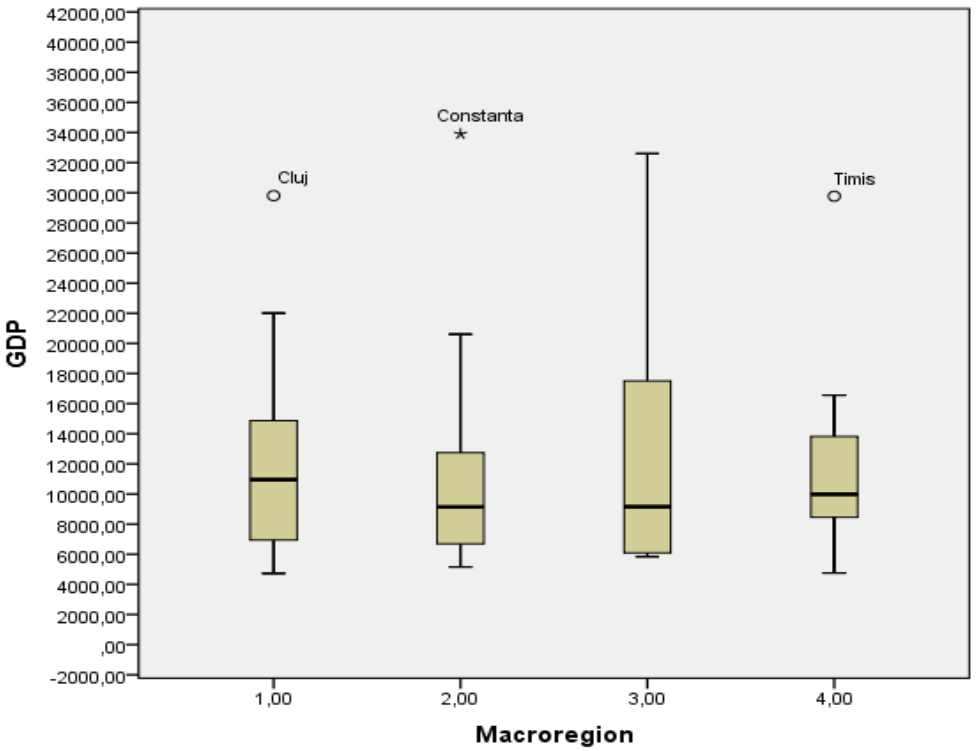

Fig. 3. *The Boxplot chart for the GDP recorded in 2014 by macroregions (Bucharest excluded)*

The results reveal a difference between the macroregions concerning the homogeneity of the results obtained by their members. Thus, while the boxes of three macroregions are quite compact, the Macroregion 3 has the biggest interquartile range and a high dispersion of the values higher than median. Due to this heterogeneity, the upper whisker covers the value recorded by Prahova county, which is not considered outlier in this case even if it is on the second place after Constanta county. In the same time, Constanta county becomes an extreme outlier in Macroregion 2 where the lowest values were recorded in comparison with the other macroregions.

3. Discussions and conclusions

Boxplot analysis provides researchers with a friendly tool meant to easy understand the distribution of numerical data, especially when the distribution is not symmetric

and the mean become unrepresentative for the analysed variables. The outliers could be also identified and analysed separately from the rest of cases in order to avoid the bias of results, especially when we use data for predictions. When certain groups with different behaviours are identified, like in the above analysis, the Boxplot analysis offers the opportunity to plot charts for every group and to make comparisons.

From the marketing research point of view, if we look at the results obtained for the GDP analysis at territorial level, one can find opportunities to develop business in the most developed counties or macroregions but also to explore new markets in regions with less development. On another hand, for the central and local authorities there is enough information about the counties that need a special attention in order to obtain the economic convergence.

Taking into account the above mentioned this instrumental research could be useful both for practitioners and academic researchers. The Boxplot analysis is easy to be computed in SPSS system and we can also use data obtained from surveys. Thus, the variables measured with a numerical scale with 5 or 7 levels could be introduced in analysis in order to find the 5-numbers summary and identify possible outliers that need a separate treatment. In other marketing researches, like the experiments, Boxplot analysis can be used successfully to compare the results before and after the experimental treatment. In conclusion, in our opinion this instrumental research represents an additional effort to help researchers to obtain good results in their activities.

6. References

Arroyo, J., Maté, C., Roque, A.MS., 2006. Hierarchical Clustering for Boxplot Variables. In: V. Batagelj, H.H. Bock, A. Ferligoj, A. Žiberna (eds.). *Data Science and Classification. Studies in Classification, Data Analysis, and Knowledge Organization*, 59-66. Springer, Berlin, Heidelberg.

Carter, Nancy J., Schwertman, Neil C. and Kiser, Terry L., 2009. A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. *Statistical Methodology*, 6 (6), pp. 604-621.

Hubert, M. and Vandervieren, E., 2008. An adjusted boxplot for skewed distributions, *Computational Statistics and Data Analysis*, 52, pp. 5186-5201.

Potter, K., 2006. Methods for presenting statistical information: the box plot. In: Hagan, H., Kerren, A. and Dannemann, P. (eds.). *Visualization of Large and Unstructured Data Sets*, GI-Edition Lecture. Notes in Informatics (LNI).

Zikmund, W. and Babin, B., 2010. *Exploring marketing research*, 10th ed. South-Western/Cengage Learning, Mason, Ohio, Australia.