STATISTICAL GRAPHICS AND EXPERIMENTAL DATA

E. Nikolić-Đorić, K. Čobanović, and Z. Lozanov-Crvenković
University of Novi Sad, Serbia and Montenegro
emily@polj.ns.ac.yu

Statistical graphs are the usual form of visual communication. In the utilization of statistical graphs the emphasis is focused either on the presentation of data or their analysis. The primary function of graphs was the visual presentation of data. But, with the development and application of electronic computers and software the analytical function of graphs has an increasing importance (Schmid, 1983). Today's graphical techniques allow comparisons between groups of data. Box-and-whisker plot is a simple graphical method that is introduced to students in teaching descriptive statistics as a useful exploratory data analysis tool for studying main characteristics of distribution, detecting outliers and extreme values. This paper deals with application of this diagram and its categorized form (trellis diagram) in analysis and modelling data from designed experiments in agriculture.

INTRODUCTION

The use of visualization is an essential way of teaching statistics. Preliminary data analysis using graphs should be emphasized in teaching of statistics. Students could very easily, using software possibilities in graphical presentation, get different information and draw conclusions concerning the study matter. The aim of this paper is to propose a way of presenting to students some methods of visualization in the analysis of real data. We will present some examples of the use of Box-Whiskers diagrams based on real data from agricultural practice.

Statistical methods, particularly inferential methods, are usually based upon rigidly defined conditions that do not hold in the most practical situations. The most common probabilistic models are based on the often wrong assumption that errors about the deterministic model are normally distributed.

Exploratory Data Analysis (EDA) is an alternative approach for data analysis, which employs a variety of techniques, mostly graphical, for data analysis. For EDA the focus is on data, its structure, outliers, extreme values and models suggested by the data. The new developments in computer hardware and software offer many possibilities in graphical presentation of multivariate data by employing sight, sound, colours and movement. The box-and-whisker (box-plot) diagram, introduced by Tukey in 1970, has a useful role in many situations when it is necessary to make preliminary analysis of data (*STATISTICA* 7.0, 2004).

Very often there exists the need of analyzing data sets according to many different characteristics. The data presented by box plots allow making comparisons between groups or treatments, comparing series of data according to the measures of central tendencies and dispersion, testing hypothesis about treatment means or totals, etc. Concerning the interaction test (Steel and Torrie, 1960), the existence of interaction between different groups or treatments, may be explored by box- plot diagrams. Box plot diagrams may be also used for exploring the normality of data set and the stationarity and seasonality of time series. Box plot diagrams allow ranking of many analyzed sets of data. When a great number of samples or treatments are analyzed, box-plots are very suitable for displaying results of multiple comparison tests, etc.

Many authors are emphasizing the need of treatment comparisons by different statistical tests (Montgomery, 1997; Hadživuković and Čobanović, 1994; Čobanović *et al.*, 2003). In such cases, box-plot diagrams can be very useful in the analysis of data results and in making decisions and conclusions about the tests of hypothesis.

Box-plot diagrams, particularly box-whiskers diagrams, are very convenient for the use in the teaching process of basic statistics. The statistical software *STATISTICA* 7.0 allows for different box-plots diagrams like: Box-Whisker, Whiskers, Boxes, Columns, High-low close. The authors suggest the use of box-whisker diagrams on the base of median, upper and lower quartiles and interquartile range for exploratory data analysis, as they convey information on the central tendency, the spread of the values and tails of distribution. In the case that the data are normally distributed, box-plots on the base of mean value and standard deviation may be used. If the aim is comparison of mean values of several groups, standard error instead of standard deviation is usually preferred.

Box-plot may be divided into tiles (known as trellis) and each subset of data may be presented on one tile. Although trellis was developed initially in the context of large data sets it is also useful for modelling data from designed experiments, even small experiments and it is very powerful tool for revealing the structure of interactions (Cleveland and Fuentes, 1997).

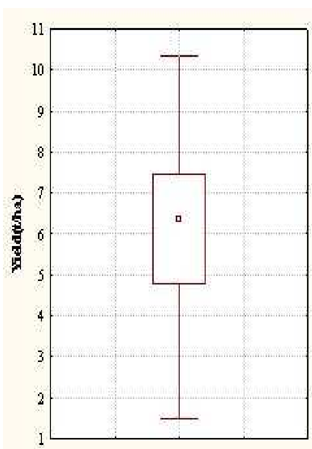
Data analysis using graphical methods in teaching Statistics could be very useful for those students who would continue studies in the research work.

The paper offers some examples of a possible use of box-plots diagrams in the teaching of analysis of experiments to students of agriculture.

ILLUSTRATIONS

We consider the results of field experiment conducted at the Institute for Field and Vegetable Crops in Novi Sad in the period 1994-1998 with three fertilizers (nitrogen, phosphorus and potassium) in three repetitions with nine variants of wheat. In the experiment four quantities of each fertilizer were applied (0, 50, 100, 150 kg/ha) at plots of the same size in 20 out of 64 possible combinations and the yield of wheat (t/ha) was measured.

A box-and-whisker diagram on the base of all 2120 experimental data (Figure 1) shows the main characteristics of its distribution. It can be seen that there exists some negative asymmetry as $Q_1 - Q_2 > Q_2 - Q_3$, where $Q_i, (i = 1,2,3)$ is i-th quartile. Also, we see that there are no outliers in the data set. At this point it should be stressed that the box-plot cannot be used for proving normality, but can be used to detect violation of the normality assumptions. Although useful in detecting outliers, the box-plot can hide multimodality. To avoid misinterpretation, it should be explained that the condition $Q_1 - \min X > \max X - Q_3$, (as met in this case), does not mean that there are more data in the lower part of data when compared to the upper part.



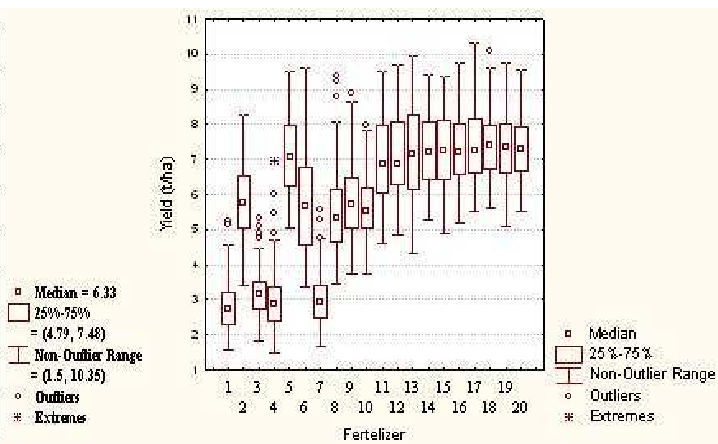


Figure 1: Box-whisker plot of yield   Figure 2: Box-whisker plots of yield data for levels of first factor

Box-plot for each level of a particular variable helps us to make comparisons between subsets of data. Box-plots made for each variant of fertilizer (Figure 2), for all varieties of wheat (Figure 3) and for each year (Figure 4) help us to consider the influence of factors on the experimental units. From these graphs it can be easily seen that the yield of wheat depends on the variant of fertilizer, on the genetic factor and on the weather conditions. The importance of box-and-whisker plot increases when it is used to compare multiple data sets as it allows an easy comparison between the distributions of the data. Box-and-whiskers plots are comparable if subsamples are of the equal size. They may be used for examination if ANOVA assumptions are satisfied (i.e. if groups are roughly symmetric and are of similar spread). It may be seen that the assumptions are not grossly violated although variants 1,3,4,7 exhibit relatively less spread (Figure 2), and data groups in Figures 3 and 4 exhibit a slightly negative asymmetric distribution.

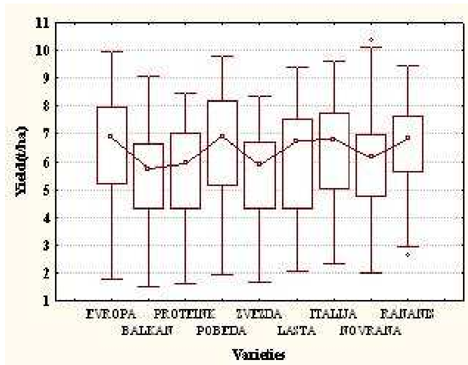


Figure 3: Box-whisker plots of yield data for levels of second factor

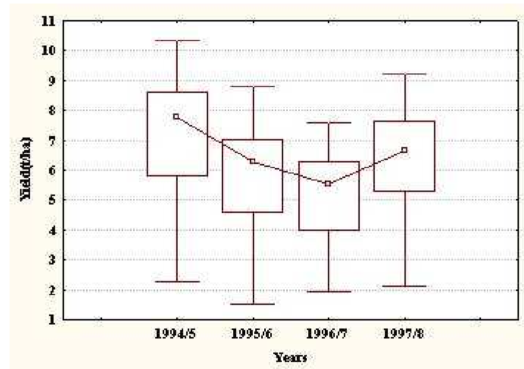


Figure 4: Box-whisker plots of yield data for levels of third factor

Categorized box-plot (trellis diagrams), obtained by arranging plots according to specified levels of a given categorical variable, helps in displaying interaction between factors. Taking variety as the variable of interest, categorized box-plots have been created as shown in Figure 5.

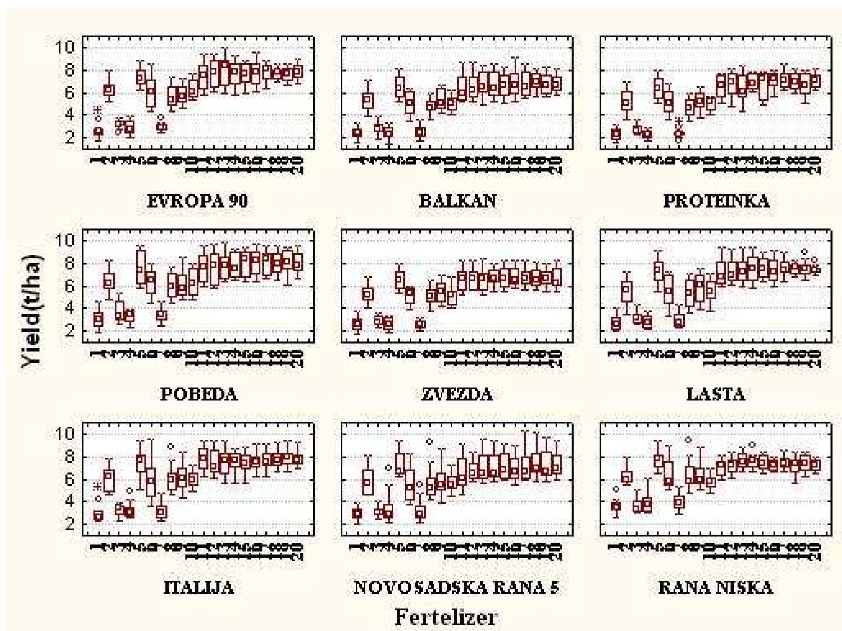


Figure 5: Trellis box-plots

It may be easily seen that the higher yields are obtained for variants 11-20 of fertilizer for all varieties of wheat. The stability of the yield for this variant depends on the variety; it is the least for Rana Niska and the highest for Novosadska Rana, so there exists an interaction between the fertilizer and the variant of wheat. In a similar way, the interaction between a fertilizer and years, years and varieties can be presented. The box-plot may be used for displaying conclusions of multiple comparisons. This is very useful when the number of comparisons is large. On the base of box-plots that contain mean value and corresponding 95% confidence interval for varieties of wheat (Figures 6) it may be concluded that the variants of wheat Balkan, Proteinka, Zvezda are homogeneous as their confidence intervals overlap. The same can be said for the variants Evropa, Italija and Rana niska. Standard errors are calculated using mean-squared error from ANOVA for three-factorial experiment. Similarly, the variants of fertilizer may be compared visually (Figure 7). In the case of unbalanced design or if the assumption of normality is not satisfied, the notched plot may be applied. It consists of notches that are drawn about median on both sides of box-plot and are extended to $\pm 1{,}58 \cdot \mathrm{IQR}/\sqrt{\mathrm{n}}$, where $IQR = Q_3 - Q_1$. The notches that do not overlap represent significant difference between medians.

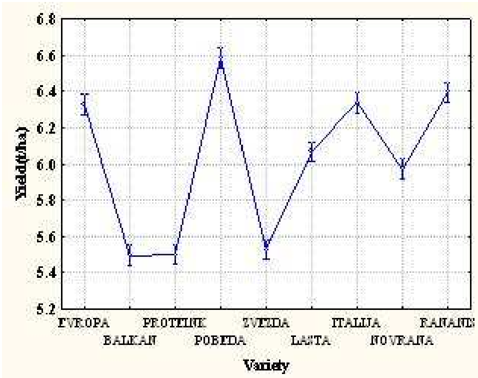


Figure 6: Multiple comparisons of varieties of wheat

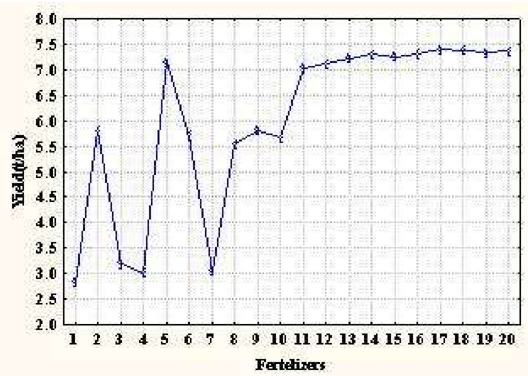


Figure 7: Multiple comparisons of fertilizers

CONCLUSION

The paper emphasizes that the box-plots diagrams are very useful in preliminary data analysis, since they are very illustrative and simple for interpretation. Therefore, they could be recommended in teaching statistics. We find that this way of presenting the multivariate data analysis helps in improving the quality of teaching.

REFERENCES

Cleveland, W. S. and Fuentes, M. (1997). Trellis display: Modeling data from designed experiments. Bell Labs Technical Report, 1-20.

Čobanović, K., Mutavdžić, B. and Nikolić-Đorić, E. (2003). Multiple comparison tests. *Yugoslav Journal for Agriculture Annals of Scientific Work*, 27(1), 66-73.

Čobanović, K. and Hadživukovič, S. (1994). Statistika: Principi i primena. *Poljoprivredni Fakultet,Institut za Ekonomiku Poljoprivrede i Sociologiju Sela.*

Montgomery, D. C. (1997). *Design and Analysis of Experiments* (4th edition). New York: John Wiley and Sons, Inc.

Schmid, C. F. (1983). *Statistical Graphics, Design Principles and Practices*. New York: John Wiley and Sons, Inc.

STATISTICA 7.0. (2004). *StatSoft,Inc*. University Licence

Steel, R. G. D. and Torrie, J. H. (1960). *Principles and Procedures of Statistics.* New York: Mc Graw-Hill Book Company, Inc.