Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 247 · 2015

Informatyka i Ekonometria 4

Grzegorz Kończak

University of Economics in Katowice
Faculty of Management
Department of Statistics
grzegorz.konczak@ue.katowice.pl

DATA VISUALIZATION IN THE RESAMPLING METHODS

Summary: Charts and graphical presentations are the examples of statistical tools. These tools are usually considered with the area of the descriptive statistics. However, many of graphics data visualizations can be successfully applied in the area related to inferential statistics. The paper presents a brief overview of the selected methods of the data presentation. The particular attention was paid to the new opportunities for graphic presentations that can be used in the method of statistical inference. Graphical methods for a long time have been successfully used in monitoring processes in statistical quality control. Due to the nature of the signals for determining the process deregulations it would be difficult, in this regard, to make a decision without the corresponding charts, and the decisions based solely on the results of numerical analyzes. The study also recalled the opportunities associated with the graphical presentation of the results of the statistical inference using resampling methods.

Keywords: charts, graphical presentations, resampling, bootstrap, permutation tests.

Introduction

The tools and methods used by statisticians are necessary for physicists, architects, engineers, doctors, psychologists, economists and for professionals in many other professions. A commonly used division makes distinction between descriptive statistics and inferential statistics. An inherently statistical research is related to the graphical presentation of data and analysis results. Data visualization is one of many statistical tools. This tool is usually associated with the issues of descriptive statistics. However, many graphics data visualizations can be successfully applied in the area related to statistical inference.

The paper presents a brief overview of selected methods of graphical data presentation. Particular attention was drawn to the new opportunities for graphi-

cal presentations that can be used in inferential statistics. For a long time graphical methods have been successfully used in monitoring production the processes in quality control. Due to the nature of the signals for determining the process deregulations it would be difficult, in this regard, to make decisions without the corresponding charts, which would be based solely on the results of numerical analyzes. The paper also describes the opportunities associated with the graphical presentation of the results of statistical inference using resampling methods.

1. Graphical methods in data analysis – books and software

There are many books and papers describing the role of charts and statistical graphics. Some of them describe historical graphics presentations [Tufte, 1983], design principles for the statistical graphs [Wilkinson, 2005] and the many other which present presents computer tools for designing graphics for data analysis [Sarkar, 2008; Kuhfeld, 2010]. There are also publications that focus on the selected aspects like graphical representation of multidimensional data [Young et al., 2006], qualitative data [Blasius and Greenacre, 1998; Friendly, 2000], presentations of large data sets [Unwin et al., 2006] or presentations of time series data [Aigner et al., 2011]. A special place is occupied by the reference books which deal with the possibility of a specific program or package [eg. Murrell, 2006].

There are many computer programs which help in preparing sophisticated graphics. Microsoft Excel can be used for creating standard graphics such as bar charts, pie charts, histograms, etc. IBM Statistics SPSS, Statistica and Mathematica have a huge set of various charts. One of the programs most frequently used by statisticians is R [www 4]. R is a free software environment for statistical computing and graphics. It can be run on Windows, Unix, Linux and MacOS. Around a hundred datasets are supplied with R in package datasets, and others are available in other packages. Graphical facilities are an important component of the R environment. It is possible to use the facilities to display a wide variety of statistical graphs and also to build entirely new types of graphs. The main graphic functions are in graphic and grid packages. Users can install other graphic packages that enhance the graphics capabilities of the program. The list of selected packages which expand the graphical possibilities of R are presented in Table 1.

2. Some new examples of graphical presentations in data analysis

There are many beautiful examples of graphical presentations of data in history. Tufte [1983] presents many historical examples of interesting data presentations such as:

− the inclinations of the planetary orbits as a function of time (s. 28),
− graphical train schedule for Paris to Lyon in the 1880s (s. 31),
− the terrible fate of Napoleon's army in Russia (s. 41),
− average size of deposits at the end of each month from 1876 to 1881 (s. 72).

Table 1. Selected packages to extend the graphical possibilities of R

Package	Description
graphics	Base graphic package
lattice	Lattice Graphics [Sarkar, 2008]
latticeExtra	Extra Graphical Utilities Based on Lattice
ggplot2	An Implementation of the Grammar of Graphics [Wilkinson, 2005]
ggvis	Interactive Grammar of Graphics
GGally	Extension to ggplot2
Grid	The Grid Graphics Package
gridBase	Integration of base and grid graphics
ggparallel	Variations of Parallel Coordinate Plots for Categorical Data
GrapheR	A multi-platform GUI for drawing customizable graphs in R
Vcd	Categorical data visualization
tableplot	Represents tables as semi-graphic displays
tabplot	Tableplot, a visualization of large datasets
tabplotd3	Tabplotd3, interactive inspection of large data
RGraphics	Data and Functions from the book R Graphics [Murrell, 2006]
maps	Draw Geographical Maps
mapStats	Geographic Display of Survey Data Statistics
googleVis	R Interface to Google Charts
RGoogleMaps	Overlays on Google map tiles in R
gRBase	A package for graphical modelling in R
plotrix	Various Plotting Functions
rggobi	Interface between R and GGobi
gplots	Various R Programming Tools for Plotting Data
prettyGraphs	Publication-quality graphics
animation	A gallery of animations in statistics and utilities to create animations
iplots	Interactive graphics for R
scatterplot3d	3D Scatter Plot
misc3d	Miscellaneous 3D Plots
RColorBrewer	ColorBrewer Palettes
colorspace	Color Space Manipulation

Source: Own preparation based on: [www 1].

Historical graphical examples presented in the book mentioned above are very sophisticated. Nowadays, computer programs limit their users to a certain type of graph. At the same time, they allow for the construction of new interesting presentation graphics that can be used in data analysis. Some of them include: WordCloud (or TagCloud), TimeLine Plot and Gapminder (TrendAnalyser).

2.1. WordCloud

One of the new graphical visualizations is WordCloud. WordCloud is a visual representation of text data. This tool can be used for books, articles and internet websites. WordCloud shows the importance of each tag (word) in the analyzed text. There are many internet sites where users can make an on-line WordCloud for given a set of text data. An example of the WordCloud for the website dedicated graphical data analysis is presented in Figure 1.

Fig. 1. WordCloud from the website „Milestones Project"

Source: [www 2].

2.2. TimeLine plot

A TimeLine plot is a chart that describes how a set of events is used over time. One of the popular types of timeline is the Gantt chart. An example of the TimeLine plot which presents a milestones in graphical presentations is shown in Figure 2.

Fig. 2. Milestones in graphical presentations

Source: [www 2].

2.3. Gapminder (Trendalyzer)

Trendalyzer (http://gapminder.org) is an interactive presentation tool that is based on a typical scatter plot. Two variables are mapped onto the Cartesian co-ordinate system. The size of bubbles represents the third variable and the color represents the fourth variable. Additionally, the animation is used to step through time. Figure 3 presents the Trendalyzer window.

The package googleVis allows creating the similar to Gapminder visualizations in R program. A set of examples (see Figure 4) of interactive visualizations with data from Local Data Bank [www 6] is presented at http://stat.ue.katowice.pl/gVis.

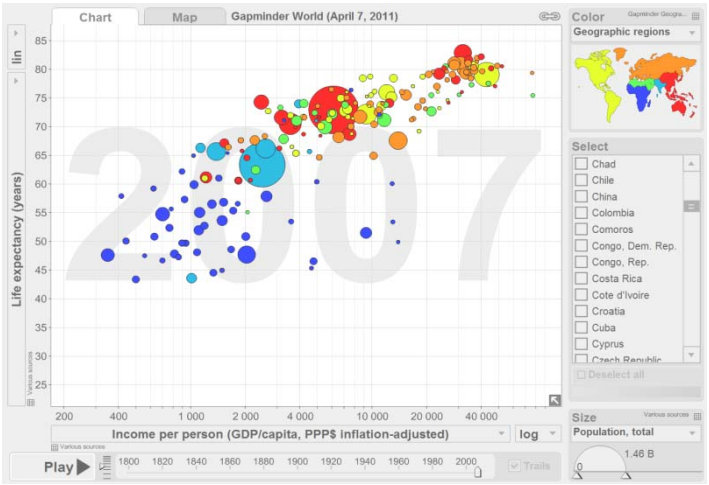

Fig. 3. Animated data visualization in Gapminder

Source: [www 3].

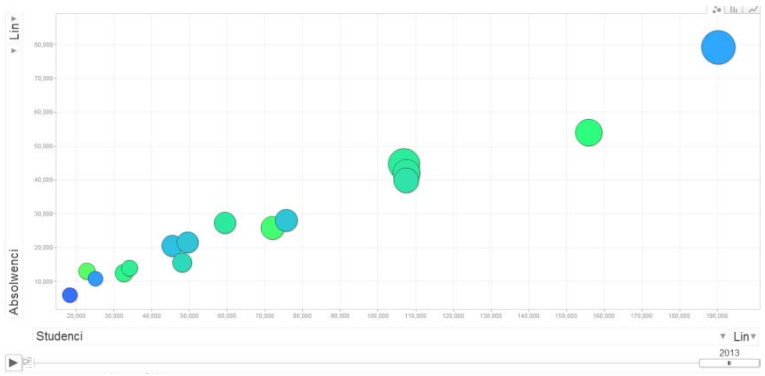

Fig. 4. Like Gapminder data visualization

Source: [www 5].

3. Graphical presentation and statistical inference

Charts and graphical presentations are usually used to describe the analyzed data. However, the graphs are often an important part of a statistical inference. For example Pawłowski [1976] used scatter plot in variance analysis. Bellow, there are some examples of the use of charts in the area of statistical inference.

3.1. Graphical methods in quality control procedures

Control charts are the tools most often used in monitoring processes. These tools were proposed in 1924 by W.A. Shewhart. The sequential plans are often used in acceptance sampling. The results of a sequence plan are often presented on a chart.

The Shewhart's control chart is a graphical view of a sequence of statistical tests. In practice there are no results of hypothesis testing (the value of the test statistic, the critical value) but only the point is plotted on the chart. There are three lines on the chart: the upper control limit, the lower control limit and the central line (sometimes additionally an upper warning line and a lower warning line). On the basis of a graphical view the decision is made if the process is on or off-control.

There are many types of signals which could be interpreted as information on the monitored process deregulation. The most often used signals are following [Montgomery, 2009]:

− one or more points outside the control limits,
− two of three consecutive points outside the two-sigma warning limits,
− a run of eight consecutive points on one side of the center line,
− six points in a row steadily increasing or decreasing.

It could be very difficult to control such various possible signals without graphical presentation on a control chart. A typical *X*bar control chart is presented in Figure 5.

Control charts are used in monitoring processes. Statistical tests are used in some other area of statistical quality control i.e. acceptance sampling. The test most often used in acceptance sampling is a sequential test. Sequential tests show the probability of rejection and acceptance as each sample is controlled. A typical graph used by sequential tests in acceptance sampling is presented in Figure 6.

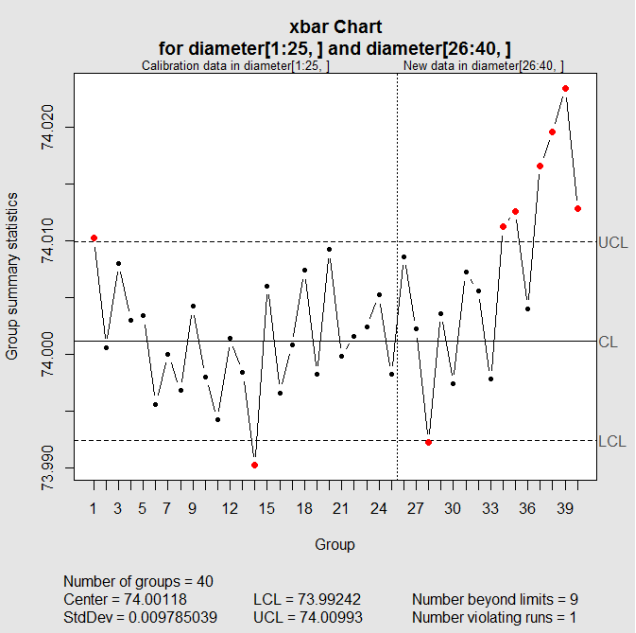

Fig. 5. The *X*bar control chart

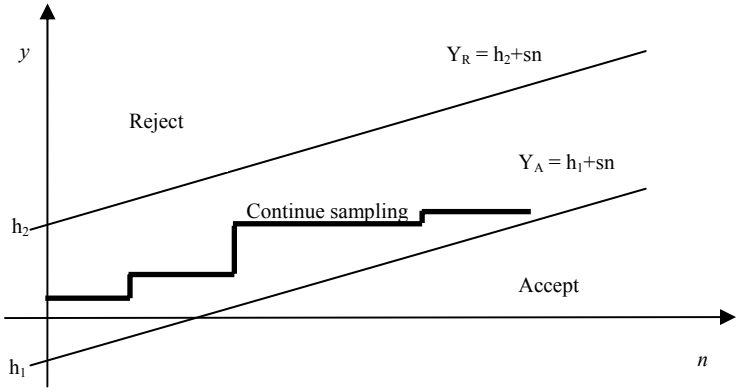

Fig. 6. Sequential sampling – procedure used in acceptance sampling

3.2. Normality testing − qqplot

Various statistical tests are used for testing normality. Shapiro-Wilk test, Kolmogorow-Smirnof test and Lillieforse test are the ones which are used most often. In normality testing a chart called qqplot could help. This plot is used to check the validity of a distributional assumption (for example normal) for a data set. If the data indeed follow the assumed distribution, then the points on the plot will fall approximately on a straight line. An example of qqplot used for multivariate normality testing is presented in Figure 7.

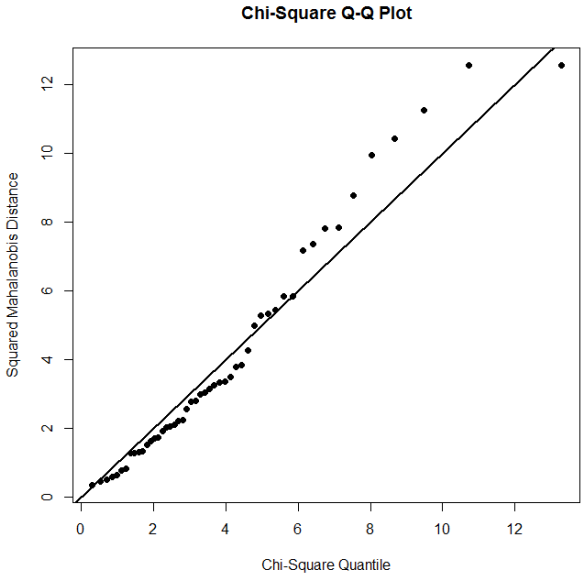

Fig. 7. Testing normality − qqplot

3.3. Graphical method in categorical data analysis

Zeileis et al. [2005] introduced several extensions to residual-based shadings for a mosaic-plot. This plot provides a method for visualizing contingency tables. For two-way contingency tables a mosaic plot fits the model of independence. Residuals for the independence model (differences between the observed and the expected values) are shaded as in Figure 8. The greater differences between the observed and the expected values the darker the correspondent area on the plot.

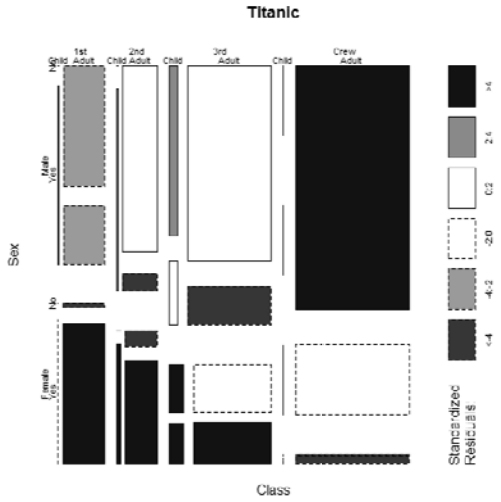

Fig. 8. Mosaic plot with residual shading

4. Resampling methods

Statistical methods are called resampling methods if they do one of following:

− estimate the precision of sample statistics by using randomly drawing subsamples (with or without replacement),
− exchange labels on data points when performing statistical tests,
− validate models by using random subsets.

The resampling methods used most often are:

− bootstrap,
− cross-validation,
− dependent random groups,
− independent random groups,
− balanced half-samples,
− permutation tests,
− Monte Carlo methods.

4.1. Bootstrap

Bootstrap was introduced by Efron [1979]. The primary goal of statisticians is to summarize a sample and generalize results to the parent population. The main idea of bootstrap is to consider the sample data as a whole population data and to resample with replication N times from the sample data with the same sample size. Bootstrap is typically used for:

− variance estimation of complex estimators,
− testing hypothesis,
− interval estimation of population parameters,
− prediction of time series (moving block bootstrap).

Davison and Hinkley [1997] present examples of graphical tests. They suggest that for qqplot a "probable envelope" could be obtained by computer simulation, to which the original data plot could be compared. Figure 9 presents the qqplot for sample data of the size $n = 12$. The null hypothesis is that the sample data were taken from an exponential distribution. The added "probable envelope" (right panel) helps to make a decision that there is no reason to reject the null hypothesis.

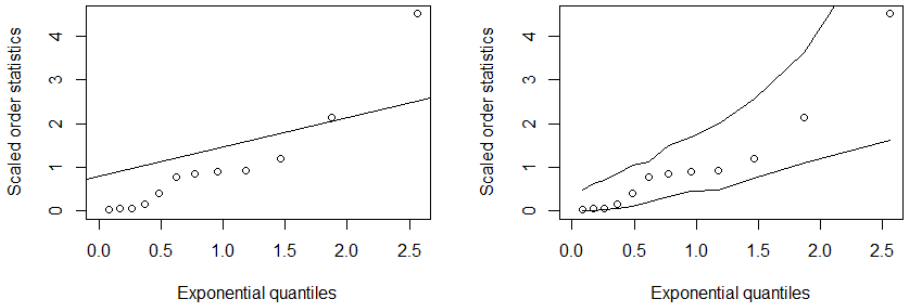

Fig. 9. Testing distribution – qqplot (left) and qqplot with „probable envelope" (right)

4.2. Permutation tests

Permutation tests were proposed by Fisher and Pitman [see Good, 2005] in the mid-thirties of the twentieth century. These tests are computer-intensive statistical methods so the role of these tests has increased rapidly in 21st century. For a permutation test the only requirement is that the labels can be permuted. In these tests instead of comparing the observed value of the test statistic to a known standard distribution the reference distribution is generated from the sample data.

Good [2005] indicates 5 following steps in permutation testing:
1. Identify the null hypothesis and the alternative hypotheses.
2. Choose a test statistic T.
3. Compute the test statistic (T_0).
4. Determine the frequency distribution of the statistic under the null hypothesis.
5. Make a decision using this distribution as a guide.

To make a decision the achieving significance level (*ASL*) should be calculated. *ASL* is the probability of observing a value at least that large as a T_0 value when the null hypothesis is true. In the case of a right sided critical area the *ASL* is expressed by [Kończak, 2012]:

$$ASL = P_{H_0}(T \geq T_0) \tag{1}$$

The value of ASL is unknown and it could be estimated as follows:

$$\hat{ASL} = \frac{card\{b : T(b) \geq T_0\}}{B} \tag{2}$$

The smaller the value of *ASL*, the stronger evidence against the null hypothesis. For a given significance level α we reject the null hypothesis if *ASL* is less or equal to α.

A typical empirical distribution in permutation testing is presented in Figure 10. The black dot represents the T_0 value in permutation test. If the point is to the right from the histogram, it leads to the rejection of a null hypothesis. If the black dot is in the middle of the histogram, then there are no reasons to reject the null hypothesis.

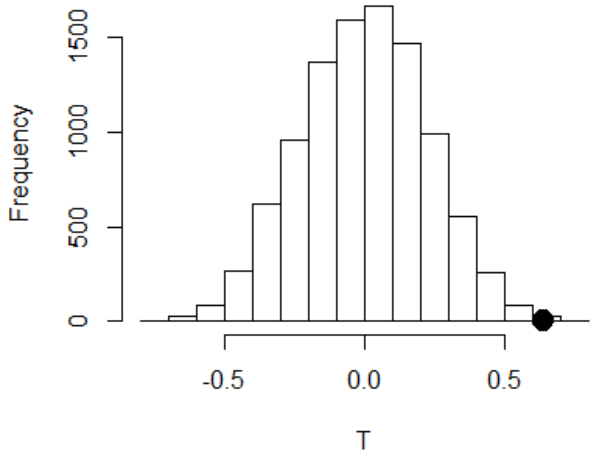

Fig. 10. Typical empirical disribution in permutation testing with T_0 value (*ASL* = 0.0014)

Stelmach [2012] describes the case of using permutation test for correlation, where the occurrence of the outlier can disrupt the shape of the histogram of empirical values of *T* statistic. The graphical view of the empirical distribution is quite different in the case of an outlier existence (right) than in the case of no outlier (left).

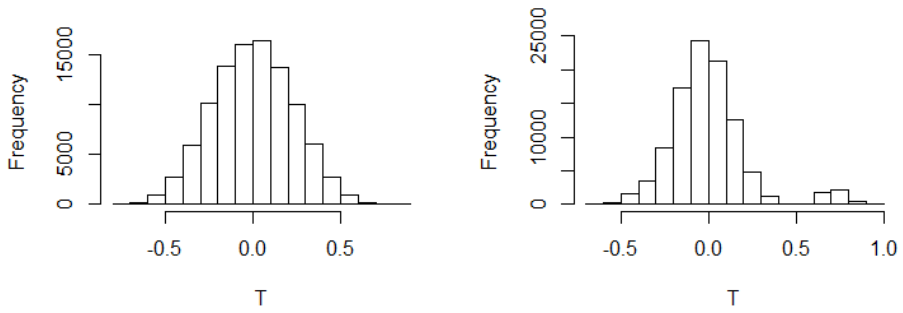

Fig. 11. Empirical distribution of *T* statistics for original data (left) and with added one outlier (right)

The data used for testing the correlation with permutation test are presented on a scatterplot in Figure 12 (left – no outlier, right – with one outlier).

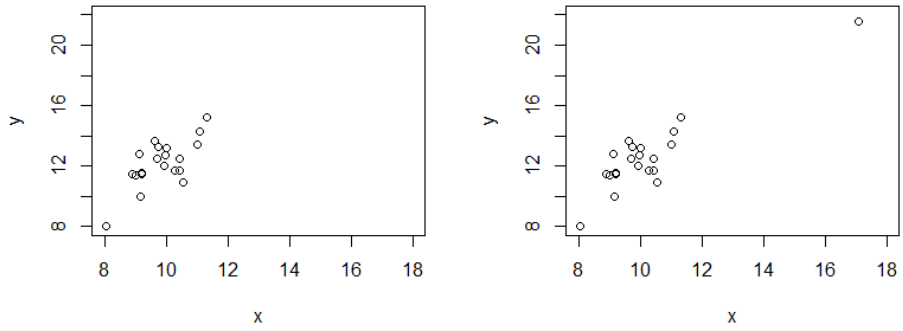

Fig. 12. Scatter plot of data used in permutation test for correlation – original data (left) and with added one outlier (right)

Concluding remarks

Graphical methods are often used in many statistical procedures. These methods are often connected to the area of descriptive statistics. The most interesting part of statistics is inferential statistics. This part is a set of formal techniques which are usually not linked to graphical methods. However, some graphical methods are closely related to statistical tests and other inferential methods. The examples of graphical methods which are connected to inferential statistics are presented in the paper.

References

Aigner W., Miksch F., Schumann H., Tominski C. (2011), *Visualization of Time-Oriented Data*, Springer-Verlag London.

Blasius J., Greenacre M. (1998), *Visualization of Categorical Data*, Academic Press Limited, London.

Davison A.C., Hinkley D.V. (1997), *Bootstrap Method and Their Applications*, Cambridge University Press, Cambridge.

Efron B. (1979), *Bootstrap Methods: Another Look at the Jackknife*, "Annals of Statistics" 7, No. 1, 1-26.

Friendly M. (2000), *Visualizing Categorical Data*, SAS Press, Cary, NC.

Good P. (2005), *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Science Business Media, Inc.

Kończak G. (2012), *On Testing Multi-directional Hypothesis in Categorical Data Analysis* [in:] Proceedings of COMPSTAT 2012, eds. A. Colubi, E.J. Kontoghiorghes, K. Pokianos, G. Gonzalez-Rodriguez, p. 427-436.

Kończak G. (2014), *Rola graficznych prezentacji danych w popularyzacji statystyki*, „Wiadomości Statystyczne", nr 7, t. LIX, p. 49-61.

Kuhfeld W.F. (2010), *Statistical Graphics in SAS. An Introduction to the Graph Template Language and the Statistical Graphics Procedures*, SAS Institute Inc., Cary, North Carolina.

Montgomery D.C. (2009), *Introduction to Statistical Quality Control*, John Wiley & Sons, Inc. New York.

Murrell P. (2006), *R Graphics*, Chapman & Hall/CRC, Boca Raton.

Pawłowski Z. (1976), *Statystyka matematyczna*, Państwowe Wydawnictwo Naukowe, Warszawa.

Sarkar D. (2008), *Lattice. Multivariate Data Visualization with R*, Springer Science+Business Media, New York.

Stelmach J. (2012), *Using Permutation Tests in Multiple Correlation Investigations*, Acta Universitatis Lodzienzis Folia Oeconomica, Vol. 269, p. 73-81.

Tufte E.R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.

Unwin A., Theus M., Hofmann H. (2006), *Graphics of Large Datasets*, Springer, New York.

Wilkinson L. (2005), *The Grammar of Graphics*, Springer, New York.

Young F.W., Valero-Mora P.M., Friendly M. (2006), *Visual Statistics. Seeing Data with Dynamic Interactive Graphics*, A John Wiley & Sons Inc., New Jersey.

Zeileis A, Meyer D., Hornik K. (2005), *Residual-based Shadings for Visualizing (Conditional) Independence*, Report 20, Department of Statistics and Mathematics, Wirtschaftsuniversitat Wien, Research Report Series.

[www 1] http://cran.r-project.org/.

[www 2] http://www.datavis.ca/.

[www 3] http://www.gapminder.org.

[www 4] http://www.r-project.org.

[www 5] http://stat.ue.katowice.pl/gVis.

[www 6] http://stat.gov.pl/bdl.

WIZUALIZACJA DANYCH W METODACH WYKORZYSTUJĄCYCH REPRÓBKOWANIE

Streszczenie: Statystyka należy do bardzo wyjątkowych specjalności naukowych. Narzędzia i metody wykorzystywane przez statystyków są niezbędne w pracy fizyka, architekta, inżyniera, medyka, psychologa, ekonomisty, a także dla specjalistów wielu innych zawodów. Powszechnie stosowany podział prowadzi do wyróżnienia statystyki opisowej i statystyki matematycznej. Nieodłącznie z badaniami statystycznymi jest związana graficzna prezentacja danych i wyników analiz. Wizualizacja danych jest zwykle kojarzona z zagadnieniami statystyki opisowej. Wiele rozwiązań graficznych można jednak z powodzeniem zastosować w obszarze związanym z zagadnieniami wnioskowania statystycznego.

W artykule przedstawiono zwięzłą charakterystykę wybranych metod graficznej prezentacji danych. Szczególną uwagę zwrócono na nowe możliwości w zakresie graficznych prezentacji, które mogą być wykorzystane we wnioskowaniu statystycznym. Metody takie od bardzo dawna są z powodzeniem wykorzystywane w sterowaniu jakością produkcji. Ze względu na specyfikę określania sygnałów rozregulowania procesów trudno w tym zakresie byłoby zrezygnować z odpowiednich wykresów, a decyzje oprzeć wyłącznie na wynikach liczbowych analiz. W opracowaniu odwołano się również do możliwości związanych z graficzną prezentacją wyników wnioskowania statystycznego z wykorzystaniem metod repróbkowania.

Słowa kluczowe: wykresy, prezentacje graficzne, repróbkowanie, bootstrap, testy permutacyjne.