DOI: 10.15503/emet2015.58.68

The advantages and limitations of the statistical approach in scientific research

Monika Nawrocka
Dept. of Statistics, Methodology and Computer Science,
Academy of Physical Education them. George Kukuczki in Katowice,
Street Mikołowska 72A, Katowice, Poland

Adam Maszczyk
Dept. of Statistics, Methodology and Computer Science,
Academy of Physical Education them. George Kukuczki in Katowice,
Street Mikołowska 72A, Katowice, Poland

Artur Gołaś
Dept. Theory of Sport,
Academy of Physical Education them. George Kukuczki in Katowice,
Street Mikołowska 72A, Katowice, Poland

Agnieszka Dygacz
Dept. Theory of Sport,
Academy of Physical Education them. George Kukuczki in Katowice,
Street Mikołowska 72A, Katowice, Poland

Aleksandra Budzisz
Dept. of Pedagogy and Psychology, Academy of Physical Education them.
George Kukuczki in Katowice, Street Mikołowska 72A, Katowice, Poland
E-mail address: the.monkas@gmail.com

Abstract

Statistical approach is based on numbers. The justification of the results obtained has a strong foundation in probability equations, statistics and econometrics. The most frequently used aspects are: sample size, interval estimation or probability distributions. After the introduction of empirical data, it is possible to measure correlations and regressions, variances, statistical deviations and many other aspects. The application of statistics enables creating on the basis of the results obtained from the empirical data: histograms, graphs, diagrams used for constructing scientific reports and eliciting irrefutable conclusions.

Keywords: statistical research, technologies, economy, statistics, econometrics.

Introduction

Statistics is preoccupied with the analysis of a given statistical assembly in scientific research. The assembly consists of elements, events or people that have a common feature but they are not identical (the students of economy study together but they are not identical). There are two kinds of such an assembly: general population and sample. The general population occurs when every single element is analyzed and each of the elements, in the case of which general conclusions are amassed, is subject to the analysis. In a sample, a researcher chooses the subassembly of the general population that they want to analyze. The choice of a representative sample may be held in a purposeful way when the researchers themselves choose the cases or during the research (randomly). During the research, there is statistical data, that is, the properties which are taken into account while analyzing a population (students' grades). They are divided into constant data - the data that is common for all the members of population and variable data which is that of varying members of population. Another division of statistical features is the one into the measurable data (quantitative) - they can be measured and presented by means of units, "categorical" data (qualitative) - verbally described and the quasi-quantitative measures - verbally described but arranged (exam results arranged from the best to the worst).

Statistics is preoccupied with:

- The analysis of population structure - the analysis of the measure of location (the mean), the measure of variability (dispersion), the measure of asymmetry (skewness) and the measure of concentration
- The analysis of correlations and regressions - the analysis of the type of dependence
- The analysis of the dynamics of phenomena - the analysis of a phenomenon in time.

Statistical research is an essential element in the area of social sciences because with its help, it is possible to draw conclusions, evaluate research problems and make the right interpretation of the results obtained. Statistical procedure is mainly used by psychologists, economists, entrepreneurs, pedagogues, geographers and physicists. The source and the basis of the analysis are almost exclusively data from questionnaires and demographic as well as social data presenting social, economic and political aspects of human behavior (Szwed, 2008, s. 12). Therefore, it is possible to maintain that statistical approach is virtually an element present in every scientific discipline. Formulating a hypothesis and the effort of solving different research problems begins with no implementation of statistics. The stage of statistics starts with the appearance of empirical data during the research. While specific aims are known, statistical methods and the methods of statistical inference are used. The crucial phase of the research is the term of interdependence between given phenomena. Enumerating the cases of interdependence has its application with regards to a great number of coincidences in particular. The justification of this approach is the lack of ability to establish the interdependence between two features in a single case. The term of dependence is specified by the analysis of correlations and regressions.

The correctly obtained data are interpreted in order to make decisions. Consequently, the conclusions needed for statistical analysis are drawn on the basis of the data. Therefore, it should be contended that all the aspects, attitudes and behavior describe the reality in the form of numbers with the use of statistics.

The functions and methods of statistics

Statistics, which is a scientific discipline, is preoccupied with the ways (methods, apparatuses) of gathering and describing quantity related data as well as eliciting conclusions (procedures present in mass processes) on their basis (Szwed, 2008, s.18). According to the subjective definition, statistics is the information specifying the observed assemblies (such data as: an average annual rainfall or about whether 66 000 married couples divorced in 2013). The apparatuses and statistical methods have a foundation in the scientific research related to the processes, structure and socio-economic phenomena as well as establishing connections and interdependence between each other. The picture below portrays the functions of statistics in the case of scientific research.

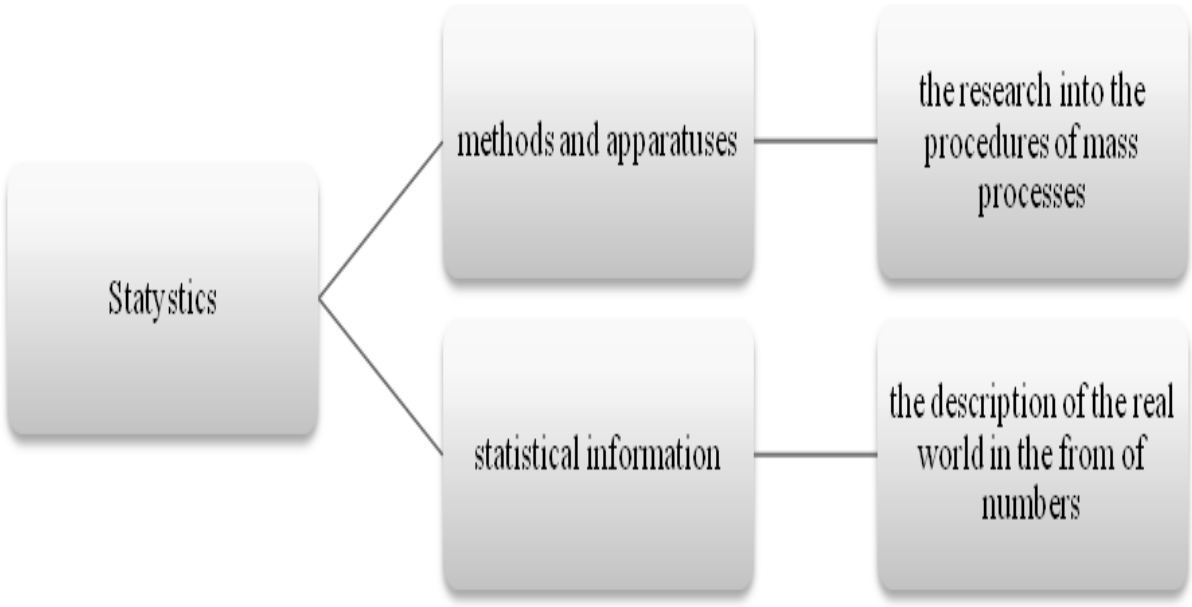

Fig. 1. The functions of statistics in research.
Source: a single-handedly created image.

Statistical methods provide the procedures that become an incredibly useful apparatus if skillfully used and thanks to which we generate concise, accurate and incredibly helpful information in eliciting the right conclusions (Szwed, 2008, p. 17).

The process of research in statistics

The procedure of statistical research is based on a few stages. As soon as the statistical research is initiated, the first stage is the formulation of a research problem. Stating a research problem is not an element of statistics. Subsequent-

ly, a researcher has to gather data in order to solve the problem. If the data are not sufficient, the new primary (the information collected especially for the purpose of solving a specific problem: the coincidence or observation method) or secondary data (the information which already exists and is gathered in order to attain other goals) are added. The stage initiates descriptive statistics. As soon as sufficient data are gathered, they are classified and grouped by means of graphical elements (diagrams, graphs, tables). Spreadsheet programs and other software for managing information are used then. While using the information derived from general population, the researcher can attempt to analyze a research problem and make a decision on the basis of the data obtained (Park et al., 1992; Pandy, 2001; Paradysz, 2005; Ostasiewicz et al., 2006). If the data used are derived from partial observation, then the implementation of the procedure of statistical reasoning via the method of inductive statistics is necessary. The parameters used are: cross-sectional estimation and statistical hypothesis testing.

Cross-sectional estimation - defined as the confidence interval. It is calculated for the expected value, variance, standard deviation and frequency.

$P\{f1(Tn) \leq 0 \leq f2(Tn)\} = 1 - \alpha$

Statistical hypothesis testing - the quality of the evaluation of specific structural parameters is checked while using e.g. the relevance of structural parameters (Ryguła, 2000; Rószkiewicz, 2002; Pułaska-Turyna, 2005; Piłatowska, 2006).

The results of statistical research are described by means of the parameters of population structure analysis:

- The measure of location- the mean, dominant (mode), quartiles and median. The dominant (mode) is the mean position measure which appears and dominates most frequently in a series. The first thing that needs to be done is to find the interval containing the mode, that is, the interval in which there is the greatest number (ni). Consequently, the value is calculated according to the following formula (Wierzbiński, 2006):

$$Mo = xl + \frac{nd - nd - 1}{(nd - nd - 1) + (nd - nd + 1)} * id$$

xl - the lower limit of the interval containing the mode
nd - the number of the interval containing the mode
nd-1 - the number of the interval preceding the one containing the mode
nd+1 - the number of the interval followed by the one containing the mode
id - the span of the interval containing the mode
The other essential measure is the median - the middle value of the data analyzed.

- The measure of variability (dispersion) - there are classical (variance, standard deviation, mean deviation, coefficient of variation, classical and typical range of variability) and position measures (quartile deviation, position coefficient of variation and position range of variability). In the case of the research into classical measures, the standard deviation, which provides the information about how much the values of data deviate from the value of the average of

all values on the average, is most frequently encountered. The standard deviation is calculated from the variance. On the other hand, the variance is not subject to interpretation. The interpretation of mean deviation is the same as in the case of the standard deviation. However, one should remember that the values of these measures will be different in the case of the same series. The standard deviation will be always of greater value.

The variance of a detailed series

The pointwise variance

$$S^2 = \frac{1}{n}\sum(xi - \bar{x})^2$$

The standard deviation

$$S^2 = \frac{1}{n}\sum xi^2 * ni - (\bar{x})^2$$

The index that provides the information about what is typically included

$$S = \sqrt{S}^2$$

within the limits and what is not typical beyond the limits is a typical range of variability. It shows that everything that is included within the limits of the standard deviation is typical. It is calculated according to the following formula:

On the other hand, the use of the position measures of interquartile range

$$\bar{x} - S(x) \leq xtyp \leq \bar{x} + S(x)$$

gives the information about the maximal divergence. Another important position measure is the quartile deviation which specifies how much the mean values of data deviate from the middle value (the median) (Sobczyk, 2002; Sej-Kolasa et al., 2004; Snarska, 2005; Sokołowski et al., 2013).

$$Q=Q3-Q1/2$$

The quartile deviation

- The measures of asymmetry - asymmetry is characterized by a different direction, strength and degree. The measure of direction is *(x- Mo)*. If *x-Mo < 0*, then the asymmetry is negative (left-sided) and indicates that fewer members analyzed centers around greater data analyzed. If *x - Mo > 0*, the asymmetry is positive (right-sided) and indicates that a greater number of members analyzed centers around smaller data analyzed. There may be an equation in which *x - Mo = 0*. In such a case, the asymmetry is characterized neither by negative nor positive direction, the series is symmetrical (a central tendency). The measurement of the strength of asymmetry requires calculating the skewness.

  - it is the mixed measure (the classical position one)
- The measure of concentration - there are the concentration coefficient (the

$$As = x - Mo/S(x)$$

kurtosis) and the coefficient of Lorenz concentration. Both of these are symmetrical, but they differ with regards to their degrees. It is the measure of gathering individual observations around the mean. It informs that the greater the value of the coefficient is, the higher the curve of population is (the curve is smoother).

While being in possession of the results, one should move on to the interpretation of the results obtained and finding a solution to the research problem.

The interdependencies of mass phenomena

The statistical description of the interdependencies of mass phenomena or cases is used when there is more than one or at least two kinds of data analyzed at the same time. The cause-effect interdependence must be held between the analyzed data. The nature of the analyzed data may be different depending on a need (measurable - measureable or immeasurable - measurable). While analyzing the interdependencies of phenomena, one may have a pointwise, weighted or span series. The systems used in the case of the analysis:

- Two kinds of measurable data together with a great number of variants.
- Weighted or span pointwise series, there may be two kinds of span data or one kind of span data and two kinds of pointwise data. There are boundary series which are related to y or x coordinate without taking into account such an aspect as the variant of the other data or conditional series related to y or x coordinate provided that the other data adopts a given variant.
- Two "categorical" kinds of data together with many variants which appear as pointwise series.
- Measurable data with the "categorical" data provided that the measurable data is treated as the "categorical" data.
- Two immeasurable kinds of data but the ones that have only two variants and appear as weighted series.

The research into the interdependencies of different data is analyzed according to correlations and regressions. There are also two kinds of interdependence: functional (the change in one value influences the change in the value of the other variable) and stochastic (the change in one value influences the change in the probability distribution of the other variable). The stochastic variable is called a correlation. The correlation establishes the direction and the strength of the interdependence of the data analyzed by means of different correlation coefficients because of the fact that there are various systems in existence. The direction may be positive or negative. The positive one is when there is one direction of the changes of data (x ↑ y↑), whereas the negative one is when the changes of the data take place in different directions. The strength informs a researcher about the numerical value of the correlation coefficient in a range (-1 ; +1) (Brandt, 1998). The basic measure is the Pearson product-moment correlation coefficient:

$$ryx = \sum(yi - \bar{y})(xi - \bar{x})/n * S(y)S(x)$$

Another measure of researching the interdependencies of data is regression.

The mechanism of connections between the data analyzed is determined with its use. To this end, the functions of regression are used: the even ones, the linear ones, the increasing or odd ones, the nonlinear ones and the decreasing ones. The adjustment of the calculated function of regression may be evaluated by the implementation of adjustment measures or statistical tests. The main measure of adjustment is the coefficient of determination $R^2=1-\varphi^2$ and the coefficient of convergence $\varphi^2=\Sigma(yi-y)^2/\Sigma(yi-\bar{y})^2$. The coefficient of determination determines the degree of the adjustment of a model to the whole, that is, the influence of exogenous variables (x) on an endogenous variable. The degree may be expressed as a percentage and is included in the interval <0,1>. The other measures of adjustment are: the residual variance (it measures the degree of the diversity of the squares of remainders ; without interpretation), the standard deviation of the change, and the average parameter estimation errors (it informs about how much a researcher is mistaken).

The procedures in statistical methods

The observations of different cases are characterized by the mass and correctness of main (systematic) and extraneous (random) reasons. The systematic reasons decided are decisive factors in the case of making a decision about the essence and shape of the correctness of population. They constantly influence a phenomenon and the general whole of the part of the population researched. The extraneous reasons emerged as a result of the influence of random factors. However, it can also happen that there is a single case in observations. In such a case, a researcher is not able to discover whether the reason is main or random. During the research, there may be a limitation in which a large number of cases will not constitute a given phenomenon. This will cause the fact that the nature of regularity will not be systematic. It means that statistical estimations will not be true and another type of research will be needed. There must be a great number of cases for the regularities and observations to be reliable.

The methods of statistical observations in scientific research are divided into:

- Full methods - statistical assembly or general population. The given populations constitute the characterization. Conducting an analysis of the population researched is possible with the use of descriptive statistics. The methods of statistical material description, the measures of variability, asymmetry and concentration are used in particular.
- Partial methods (incomplete, not full) - they become effective when there are no conditions that allow the implementation of full observation. The choice of members (sample) out of the whole general population is random. The sample specifies the series and cannot greatly differ from the general population.

The research into general population cannot always be conducted because of high costs. A part of the general population is chosen in such a case. A series is specified as a subset because its internal structure does not differ from the general population. After the analysis, the collected statistical data do not always reflect the reality. They can be inflated or underestimated.

Objectivity in statistical approach

Statistical approach is objective. It is this way because the foundation in statistics is numbers. Quantitative analyses guarantee the researchers the precision of results as well as drawing clear and irrefutable conclusions.

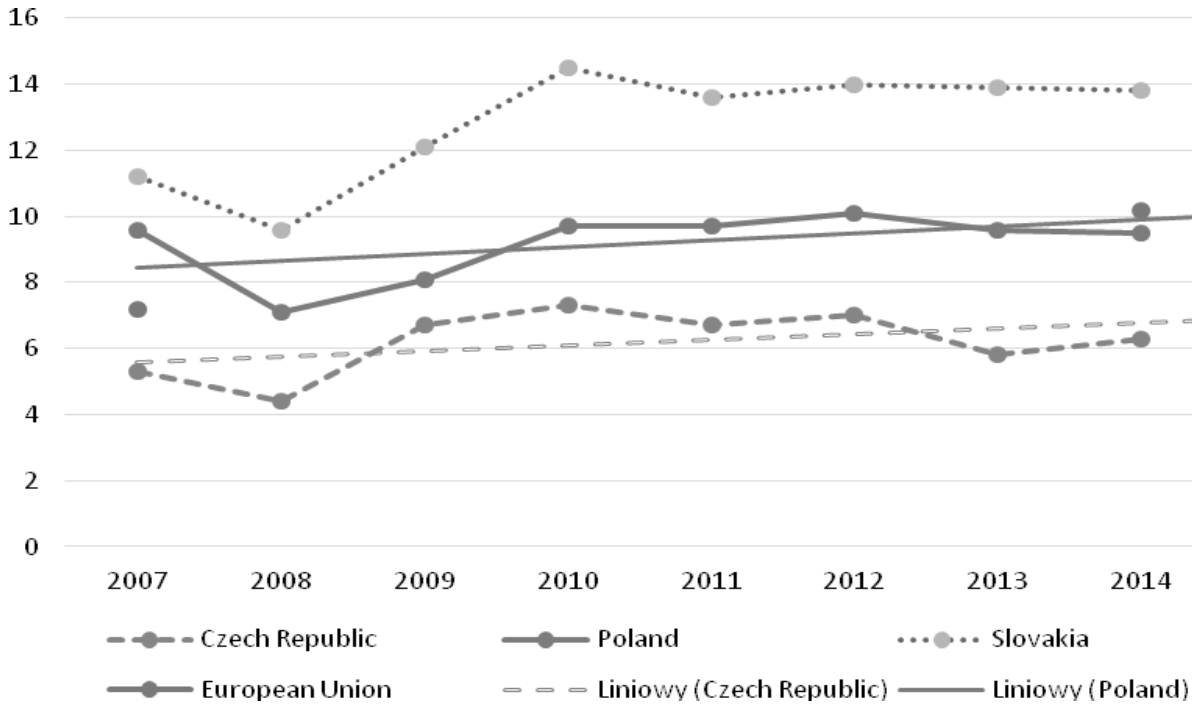

Fig. 2. An example of the results of the research into the unemployment rate in Poland, Czech, Slovakia and European Union (n=8).
Source**:** own compilation based on the database obtained from websites: www.eurostat.com

According to Eurostat the unemployment rate in Poland, a decrease to 7,1% was estimated from 2007 to 2008, whereas it stays on the level of 8,5% in the current year. The reading of the unemployment rate in The Czech Republic (6,3%), Slovakia (13,2%) and European Union (10 %) has been attached to the picture. The trend line for future periods for Polish and the Czech Republic reached a growth. The use of a linear regression analysis was used to predict future values of the unemployment rate. The average unemployment rate in Poland is lower than in the chosen countries of European Union (Fierla, 2007). It is debatable that the percentage of young people immigrating, inter alia, to the presented countries is still growing. Therefore, statistical data are not always the basis for evaluating what happens in the world.

The objective results of statistical research may be distinguished in analyses like (Szwed, 2008):

- *The evaluations of the level of phenomena* (the percentage of the number of household users with access to the Internet in Poland was 67 percent in 2013 - as it results from the all-Poland research Social Diagnosis 2013).
- *The evaluations of the course of socio-economic processes* (the number of immigrants that have lived in Poland for more than three months is 40097 on the

whole according to Central Statistical Office, Lower Silesian - 3558, Kuyavian- Pomeranian - 1022, Lublin - 1014 and Łódź - 3069).

- *The analysis of the dynamics of phenomena in time* (the ecology; the sport science - analysis of the dynamics of volatility athletics results with prediction (Osowski, 2000; Osiński et al., 2001; Savaglio, 2001; Maszczyk, 2013; Nawrocka, 2014))
- *The research into the structure of socio-economic processes* (according to Central Statistical Office, at least 36 000 inhabitants of our country announced their departure from Poland in 2013).
- *Determining the connections between different phenomena* (there is a strong connection between revealing aggression behind the wheel and being a victim according to the research of Gallup, Inc. from 2003; 70% of the people that had experienced other drivers' aggressive behavior admitted to similar behavior in other situations).
- *Determining the influence of different factors on the course of socio-economic processes* (according to French doctors' research, a lifestyle - 52%, genetic factors and the state of the natural environment - 20% each, as well as health care - 8%, exert the greatest impact on population health; the predominant causes of death are circulatory diseases - 54%, and neoplasms - 24%) (Stanisz, 2007).

Conclusion

The statistical approach is indeed a precise way of conducting scientific research and obtaining honest results in order to draw exact conclusions. The reality presented by means of statistics is very scientific. Nevertheless, it is worth taking into consideration the fact that wrongly chosen numbers, methods or apparatuses mislead due to unawareness or researcher's aware activity. The manipulation of statistical data may have revolutionary as well as destructive results. However, such conduct is not acceptable. Therefore, the right knowledge and experience is needed in order to make use of statistical methods.

The limitations which a researcher has to face are high costs due to which conducting the research into general population is usually impossible. The statistical data collected from the subset of the general population do not always reflect the reality after the analysis. That is why statistical data are not always the basis for evaluating what happens in the world. There is a danger of the manipulation of the statistical data or the inappropriate use of apparatuses. However, the critical parameters at solving many complicated optimization problems are often time constraints or a hardware limitatiom (Mańdziuk, 2000).

The benefits are the opportunity to use the statistical approach in many scientific disciplines (medicine, marketing, economics, psychology etc.). Statistical research has its application to general population as well as partial population. All the stages of the statistical research may be easily transferred into a graphical format (diagrams, graphs and tables). Graphical presentations are more readable when a huge data set is concerned. The biggest advantage is a huge range of statistical methods. The methods chosen in the article serve the purpose of

analyzing population, statistical inference procedure and correlational relationships between variables. The statistical approach gathers very many other significant measures for the purpose of research. The calculus of probabilities is worth mentioning. In the classical sense, the probability of the specific random event A is the relation of the number of events conducive to this event being held to the number of all the possibilities of elementary events (equally likely and excluding each other) (Szwed, 2008, p.99). The literature of Wojna proved that the set of methods is used to operations research and optimization. I shall reply to questions related to management decision-making and economic and social phenomena (Wojna, 2007). While the modeling and programming linear suited in determining the predictors, the formation of linear trends, and study the relationship between variables.

In the end, I will add that without statistics, there would be no opportunity to evaluate global phenomena and conduct an analysis of the processes taking place in the whole world. One can repeatedly encounter different data in the form of voting results, the mean of social processes which have the foundation in numbers. While speaking of phenomena, the mass ones usually require calculating the data with the help of statistics. One of the good examples may be conducting an analysis of human migrations in Europe. The population is so highly dense that this phenomenon cannot be "counted on one's fingers".

References

Brandt, S. (1998). *Analiza danych* [Analysis of the data]. PWN: Warsaw.

Fierla, I. (Ed.) (2007). *Geografia ekonomiczna Unii Europejskiej* [Economic Geography The European Union]. PWE: Warsaw.

Krzysztofiak, M., & Urbanek, D. (1981). *Metody statystyczne. Wyd. 4* [Statistical methods. Ed. 4]. PWN: Warsaw.

Mańdziuk, J. (2000). *Sieci neuronowe typu Hopfielda. Teoria i przykłady zastosowań* [Hopfield-type Neuron Networks. Theory and application examples]. EXIT: Warsaw

Maszczyk, A. (2013). *Analiza i predykcja dynamiki zmienności światowych wyników konkurencji lekkoatletycznych w latach 1946-2011* [Analysis and prediction of the dynamics of the volatility the world athletics of competition results in the years 1946-2011]. Katowice:AWF.

Nawrocka, M. (2014). Analysis of dynamic sports result changes in 20 km race-walking. In: I. Čillik (Ed.) *Atletika 2014 : recenzovaný vedecký zbornik z medzinárodnej konferencie* (pp. 276-288). Banská Bystrica : Univerzita Mateja Bela.

Osiński, Z., & Wróbel, J. (2001). Optymalizacja maszyn a optymalizacja matematyczna [Optimization of machines and mathematical optimization]. *Scientific Papers of Bialystok Technical University. Construction and Operation of Machines, 8,* 384-395.

Osowski, S. (2000). *Sieci neuronowe do przetwarzania informacji* [Neural networks to process information]. Oficyna wydawnicza Politechniki Warszawskiej: Warsaw

Ostasiewicz, S., Rusnak, Z., & Siedlecka, U. (2006). *Statystyka. Elementy teorii i zastosowania* [Statistics. Elements of the theory and Applications]. Wydawnictwo Akademii Ekonomicznej we Wrocławiu: Wroclaw.

Pandy, M.G. (2001). Computer modeling and simulation of human movement. *Annual Review of Biomedical Engineering, 3,*245-73.

Paradysz, J. (2005). *Statystyka* [Statistics]. Wydawnictwo Akademii Ekonomicznej w Poznaniu: Poznań.

Park, D.J., Jun, B.E., & Kim, J.H. (1992). *Novel fast training algorithm for multilayer feedforward neural network.* Korea Advanced Institute of Science and Technology. Daejeon: South Korea.

Piłatowska, M. (2006). *Repetytorium ze statystyki* [Repetytorium of statistics]. PWN: Warsaw.

Pułaska-Turyna, B. (2005). *Statystyka dla ekonomistów* [Statistics for economists]. Difin: Warsaw.

Rószkiewicz, M. (2002). *Metody ilościowe w badaniach marketingowych* [Quantitative methods in marketing research]. PWN: Warsaw.

Ryguła, I. (2000). *Narzędzia analizy systemowej treningu sportowego* [Tools of system analysis in sports training]. AWF Katowice.

Savaglio, S. (2001). Scaling laws and forecasting in athletic world records. *Journal of Sports Science*, *7(1),* 11-16.

Sej-Kolasa, M., Zielińska, A. (2004). *Excel w statystyce. Materiały do ćwiczeń* [Excel in statistics. Materials to exercise]. Wydawnictwo akademii Ekonomicznej we Wrocławiu: Wrocław.

Snarska, A. (2005). *Statystyka, Ekonometria, Prognozowanie. Ćwiczenia z Excelem [Statistics, Econometrics, Forecasting. Exercises with Excel]*. Placet: Warsaw.

Sobczyk, M. (2002). *Statystyka* [Statistics]. PWN: Warsaw.

Stanisz, A. (2007). *Przystępny kurs statystyki w oparciu o program STATISTICA PL* [Approachable course statistics based on the program STATISTICA PL]. StatSoft: Krakow.

Szwed, R. (2008). *Metody statystyczne w naukach społecznych. Elementy teorii i zadania* [*Statistical Methods in the Social Sciences]*. KUL: Lublin.

Wierzbiński, J. (2006). *Statystyka opisowa* [Descriptive Statistics], WZ: Warsaw.

Wojna, A. (2007). *Predykcja ekonometryczna i modelowanie stochastyczne. Cz. 1* [Econometric prediction and stochastic modeling. Part 1]. Wydawnictwo Uczelniane Politechniki Koszalińskiej: Koszalin.