# The use of statistics in social sciences

**Article** *in* Journal of Humanities and Applied Social Sciences · November 2019

DOI: 10.1108/JHASS-08-2019-0038



**1 author:**

Petros Maravelakis
University of Piraeus
**57** PUBLICATIONS **1,181** CITATIONS

SEE PROFILE

# The use of statistics in social sciences



Petros Maravelakis
*Department of Business Administration, University of Pireaus, Piraeus, Greece*

Received 30 August 2019
Accepted 30 August 2019

## Abstract

**Purpose** – The purpose this paper is to review some of the statistical methods used in the field of social sciences.

**Design/methodology/approach** – A review of some of the statistical methodologies used in areas like survey methodology, official statistics, sociology, psychology, political science, criminology, public policy, marketing research, demography, education and economics.

**Findings** – Several areas are presented such as parametric modeling, nonparametric modeling and multivariate methods. Focus is also given to time series modeling, analysis of categorical data and sampling issues and other useful techniques for the analysis of data in the social sciences. Indicative references are given for all the above methods along with some insights for the application of these techniques.

**Originality/value** – This paper reviews some statistical methods that are used in social sciences and the authors draw the attention of researchers on less popular methods. The purpose is not to give technical details and also not to refer to all the existing techniques or to all the possible areas of statistics. The focus is mainly on the applied aspect of the techniques and the authors give insights about techniques that can be used to answer problems in the abovementioned areas of research.

**Keywords** Time series, Statistics, Sampling, Social sciences, Count data, Non-parametric, Multivariate, Statistical modeling, Bayes

**Paper type** General review

## 1. Introduction

According to Dodge (2008), Statistics:

> [. . .] is made up of a set of techniques for obtaining knowledge from incomplete data, from a rigorous scientific system for managing data collection, their organization, analysis, and interpretation, when it is possible to present them in numeric form.

The purpose of this paper is to try to review the statistical techniques in the field of social sciences in other words social statistics.

Social statistics is the field of statistical science that deals with the study of social phenomena and in particular human behavior in a social environment. Such phenomena are any kind of human activities, including activities of groups of people like households, societies and nations and their impacts on culture, education and other areas. Generally, we can say that social statistics deal with the application of statistical methodology in areas like



Journal of Humanities and Applied Social Sciences
Vol. 1 No. 2, 2019
pp. 87-97
Emerald Publishing Limited
2632-279X
DOI 10.1108/JHASS-08-2019-0038

survey methodology, official statistics, sociology, psychology, political science, criminology, public policy, marketing research, demography, education, economics and others.

Due to the nature of social sciences it is common to study indicators that cannot be measured directly. Moreover, data that is unobservable, informal, illegal or "too personal" are often studied in this area (Lovric (2011)). For example, a social researcher may be interested on the data (answers) of the question "Do you participate in illegal gambling?". Other similar questions may be asking, for example, about the sexual behavior of a respondent, possible addictions etc.

In this paper, we review some statistical methods that are used in social sciences and we draw the attention of researchers on less popular methods. Our purpose is not to give technical details and also not to refer to all the existing techniques or to all the possible areas of statistics. We focus mainly on the applied aspect of the techniques and we give insights about techniques that can be used to answer problems in the abovementioned areas of research.

The remaining of the paper is organized as follows. In Section 2, we refer to the sampling issues which are of fundamental importance to gather data that are representative of the population under study. Section 3 gives an overview of descriptive statistics, regression analysis and analysis of variance. Since regression is a method heavily used to model quantitative data it is of great importance in social sciences. Analysis of variance can be used to identify if one or more categorical variables have a statistically significant effect on a continuous (dependent) variable. Parametric models are presented in Section 4. It is practically impossible to cover all the available models but a number of important models are outlined. Non-parametric methods are given in Section 5. These methods are particularly important because assumptions in parametric models are frequently questionable in practice. Some multivariate techniques are described in Section 6. Since most of the data contain several different variables that are related multivariate techniques are a necessary tool. Usual practice in questionnaires is to collect categorical data with inherent order and without order. In Section 7, we describe some of the popular techniques that analyze such data. Some of the time series methods to analyze data that are dependent over time are given in Section 8. In Section 9, we present the data mining methods that can be used to identify patterns in large data sets. Finally, we give some conclusions and recommendations.

## 2. Sampling

The first issue before the use of any statistical method is the collection of the sample. We may say that sampling is a collection of techniques on how to select a number of individuals from the target population to estimate certain characteristics of the population that we want to study. There are two ways to select a sample, either using a probability or a nonprobability sample. In probability sampling every unit of the population has a chance of being selected in the sample. Moreover, this probability is greater than zero and it can be computed. The methods that are used are simple random sampling, systematic sampling, stratified sampling, cluster sampling and multistage sampling (or combinations of these methods). On the other hand, in nonprobability sampling some of the units of the population have zero probability of being selected or the probability of being selected cannot be computed. The most known non-probability sampling methods are intentional sampling, convenience sampling, quota sampling and snowball sampling.

The main difference between probability and non-probability sampling is the fact that with probability sampling we draw a random sample. This assertion is extremely important due to the fact that using statistical theory we can extend the results from a random sample to draw conclusions about the population. This is not allowed in non-probability sampling.

The way to select a random sample is apparently very important. However one may observe that in papers published in social sciences this fact is overlooked. Specifically, emphasis is given to the conclusions of the research effort and not on the way data was gathered. This fact is surely a very important factor on the reliability of the results. In the same issue, another statistical question that needs special attention is the determination of the sample size. In many published papers the authors just refer to the size of the sample without referring to the technical details of its computation. This is another serious problem that affects the credibility of the results in a survey.

These two problems are not the only that appear in surveys in social sciences but surely they are the most usual. Other issues that occur are blurred definition of the population, problems during the collection of the data that are rarely mentioned (for example replacement of selected units), non-sampling errors (for example non-response, over or under coverage) and others.

Generally, we can say that in many cases sampling is treated with less attention than what is needed. Researchers in the field and practitioners may refer to the classical book by Kish (1995). Other useful references are Cohen (1988), Joseph *et al.* (1997), Lenth (2001) and Shuster (1990).

## 3. Descriptive statistics, regression and analysis of variance

The first step in any statistical analysis is the use of descriptive statistics to present the data and try to identify any kind of trends, relationships or abnormal behavior. Analysis based on descriptive statistics or exploratory data analysis usually makes no stochastic assumptions. A first approach in parametric tests is to use the classic hypothesis tests and confidence intervals. Apart from that there are also other statistical methods that can be employed in social sciences. Regression analysis and analysis of variance (ANOVA) are some of the classical methods.

### *3.1 Regression*

Regression is one of the most known methods used for analyzing relationships between variables. The main objectives of a regression analysis is to check if there is an association between variables, to identify the strength of this relationship and to conclude to a regression equation that is used to describe this relationship.

There are several forms of regression modeling, for example, linear regression, logistic regression and regression discontinuity. There are also other aspects of the regression methodology but we confine ourselves to these cases. All these methodologies have been extensively used in real cases of social sciences.

Linear regression is the simplest of these methods since it is used to model the relationship between one dependent variable and one or more explanatory variables. In this methodology, we try to find a function to fit the values of the explanatory variable that vary linearly with the target variable. Linear regression is particularly useful since it is able to predict the value of the dependent variable given the value of the explanatory variable or variables. We have to stress that in this method the target variable (dependent variable) is continuous.

In logistic regression we want to obtain a nonlinear curve to fit the data when the target variable is discrete. This methodology is particularly useful in modeling a target variable having value for example Yes (0) or No (1). More formally we can say that the target variable is binomial. Our aim is to find an equation that functionally connects the values of the explanatory variables to the values of the target variable. The explanatory variables can be either continuous or categorical. Since the range of the explanatory variables can be

between $-\infty$ and $\infty$ a proper transformation is applied to the target variable. If we transform the target variable to the logarithm of the odds of its values then the transformed target variable is linearly related to the explanatory variables. For more details about the linear and the logistic regression the interested reader can refer to Kutner *et al.* (2005).

Regression discontinuity is used to compute the effect of an intervention. This methodology is able to give unbiased estimates of this intervention. In a regression discontinuity design we use a rule to assign the intervention to a unit. This methodology is extensively used in education. Specifically, a scoring rule is used after a test is given in a class to select the students that need more effort on the specific course. Students with scores below a cutoff value are assigned to the group that will spend more time studying and students with scores above the cutoff value are assigned to the comparison group, or vice versa.

The effect of the intervention is estimated as the difference in the mean outcome of the treatment group and the comparison group. A regression line or curve is estimated for the two groups (treatment and comparison groups), and the difference in the mean of these regression lines at the cutoff value of the measured variable is the estimate of the effect of the intervention. We conclude that there was an effect of the intervention if a "discontinuity" appears between the two regression lines at the cutoff value. A detailed description of regression discontinuity is given in Riley-Tilman and Burns (2009) and Jacob and Zhu (2012).

### *3.2 Analysis of variance*

Analysis of variance (ANOVA) is a well-known method used to compare several means at the same time using a fixed confidence level. The data used are the results of an experiment. There is a continuous dependent variable, and one or more qualitative independent variables (categorical or nominal variables). The design of the experiment must be done in such a way that it will not affect its results. For example, a completely randomized experiment does not affect the output of the experiment. However, the choice of the design of the experiment affects which analysis of variance method will be used. There are a lot of different designs of experiments and analysis of variance methods for several different cases.

Regression analysis and analysis of variance are closely related. If we use dummy variables as independent variables in analysis of variance then the analysis becomes regression analysis. However, there is a serious difference between the two methods. In the analysis of variance if the design of the experiment is properly done, we may conclude that there is causality (the independent variable has a causal effect on the dependent variable). On the other hand, in regression analysis a statistically significant effect may mean causality or not (a statistically significant result does not necessarily mean causal effect).

The analysis of variance tests the independence of the response and explanatory variables. If we decide that there is this type of dependence then we have to do extra analysis to identify which means are different and to what extent.

The analysis of variance assumes that the samples in the groups (categories of the independent variable) are independent. This means that each group has a different sample of subjects. However, there are cases where each group has the same sample of subjects. Apparently, the samples are then dependent and of course we have to take this fact into consideration to reach credible results. This case is called repeated measures analysis of variance.

For more information on this topic, see Agresti and Finlay (2009) and Cohen and Lea (2004).

## 4. Parametric methods

Assume that a researcher wants to use the ANOVA and apart from the dependent variable and a categorical variable (factor), data for one or more quantitative variables measured on each experimental unit are available. Then, if these variables have an effect on the outcome of the experiment, they can be used in the model as independent variables. Such variables are called covariates or concomitant variables. The analysis involving all these variables is called analysis of covariance. Although the model is more complex by including the extra variables, the profit is that the error variance is reduced.

Another very useful class of models is mixed models. Mixed models contain both fixed and random effects. They are particularly useful in social sciences when we have repeated measurements. Moreover, in the case of missing data, which are very common in sample surveys, mixed models offer a strong alternative to methods like ANOVA for repeated measures. Their drawback is that estimation is more difficult along with the fact that we end up to have a more complex model.

A useful class of models is also the semiparametric models (or even better the semiparametric regression models). These regression models include both parametric and nonparametric components. They are used when the usual parametric models do not have a satisfactory performance. More about nonparametric methods are given in Section 5.

Another very useful method is robust regression. Keeping in mind the usefulness of linear regression, its wide applicability and acceptance between the researchers it is natural to propose a method that overcomes the difficulty to fulfill its assumptions. Robust regression is used to avoid the effect of outliers. One approach is to use the M-estimators and another one is to replace the normal distribution in the assumptions with a heavy-tailed distribution.

Undoubtedly methods like linear regression and ANOVA have been used to an enormous extent in social sciences but many times without the proper accuracy in the details. We believe that much of the work done could be improved using the more advanced models presented in this section. For more details the reader could refer to Christensen (2011) and Rencher and Schaalje (2008). For robust regression a useful reference is Rousseeuw and Leroy (1987).

## 5. Nonparametric methods

In social statistics the vast amount of research is based on parametric methods. However, many parametric methods are based on strong assumptions that are disregarded most of the times. This has serious effect on the justification of the results.

The alternative in this case is to use nonparametric statistical methods. Nonparametric statistics do not rely on a specific family of probability distributions and there is no assumption about the probability distributions of the variables used. Therefore it is an ideal collection of methods for handling real data that most of the times fail to follow these strong assumptions of parametric inference.

There is a number of techniques that are already popular among the researchers in social sciences. Such techniques are certain hypothesis tests like Wilcoxon Signed rank test, Mann-Whitney test and Kruskal-Wallis tests. Other used techniques are the Spearman correlation coefficient, the runs test and normality tests. For a detailed review of such techniques the interested reader can refer to Corder and Foreman (2009).

However, there is a number of other nonparametric methods that have been developed and are already famous among statisticians that have not gained much attention between the researchers in social sciences. Such methods are the jackknife and the bootstrap methods. Jacknife can be used to compute the bias and the variance of an estimator whereas

bootstrap estimates the variance and the distribution of a statistic or it is used to construct confidence intervals. It must be noted that both these methods are computationally demanding. Nevertheless, they can be very useful in social sciences especially in the cases of complex estimators of parameters that need to be further studied.

Another useful method is nonparametric regression. The usual linear regression is a heavily used method in social sciences. However, its assumptions are very rarely referred due to the fact that they rarely hold. Nonparametric regression is a solution in that case. It is able to answer the initially stated problem that led to regression with flexibility in terms of the assumed model. Other interesting nonparametric methods are the ones used for density estimation like cross-validation and density estimation. These methods estimate the probability distribution function using just the data. They can be used in cases where the distribution of the data is unknown and difficult to be computed analytically. If a researcher is able to compute the distribution function of the variable or variables under study then he/she can obtain statistical methodologies like confidence intervals or hypothesis testing making the decision process easier and credible. For more details about these methods the interested reader could refer to Wasserman (2006).

## 6. Multivariate methods

Usually in social sciences and generally in real problems more than one variable is involved. These variables need to be considered together since most of the times they are related. Several methods have been developed for the analysis of such data. These methods include among others cluster analysis, correspondence analysis, principal component analysis and factor analysis.

One of the main goals of multivariate analysis is classification. Cluster analysis is a method of classification which aims to group individuals (objects) so that those allocated to a particular group are, in a way, considered to be close together. The data used in cluster analysis are a data matrix where the columns are used for the objects and the rows for the attributes that describe the object. The output of a cluster analysis is the clusters that are used to characterize objects as similar or not. In hierarchical cluster analysis the clusters appear as a tree (they have hierarchy). In nonhierarchical cluster analysis, the number of clusters are determined by the researcher which have to be less than the number of objects. Both of these techniques are processed through statistical software. The allocation of people in similar groups is very important for a social scientist since it gives him the ability to pin point the special characteristics of these groups.

Correspondence analysis is an exploratory technique that helps a scientist to analyze multi-way frequency tables. Its main goal is to plot the data using less dimensions to identify their key features. The data used in this method have to be nonnegative and they should appear in a data table. Correspondence analysis aims to display data tables in two-dimensional spaces, called maps. The idea behind this method is that the model must follow the data, and not the opposite. In its simplest form we have a variable that we want to model and several explanatory variables. All these variables are frequencies appearing in one or more contingency tables. We use cross-tabulation for each of the explanatory variables and the variable we want to model to identify the level of their association. A technique which is also used, is to stack the tables before the application of correspondence analysis to reveal the relationship of the variable we want to model with the explanatory variables in the same map.

Principal component analysis is used to summarize $p$-correlated variables by a smaller number of uncorrelated variables. These variables contain most of the information that exist in the original set of variables. Keeping in mind the vast amount of data a social scientist has

at hand today, we may conjecture that this technique is very important. The fact that we end up with a smaller number of variables, demands less computational power to perform the analysis of the remaining variables. Moreover, the fact that the variables are uncorrelated makes the analysis easier since the techniques used do not have to consider a relationship between the variables used. However, there a number of drawbacks. First of all, the fact that a piece of information is lost may affect the conclusions of the analysis. Moreover, if we begin with thousands of variables (which is not rare today) we may have to work with a lot of variables even after the application of principal component analysis to retain most of the information in the data.

As we already stated in the introduction sometimes in social science research, we cannot measure the variable or variables that we are interested in a direct way. These variables are called latent variables and commonly they are called factors. An example of a latent variable is human intelligence. In Factor analysis we try to relate the observable to the unobservable variables by a probability model to make statistical inference. The main objective of the analysis is to select the number of the latent variables that have to be used to explain the correlations between the unobservable variable and to interpret them. Another objective is to predict the values of the latent variables that produced the observable variables. In factor analysis the researcher regresses each of the observed variables on the set of the latent variables. Usually after the computation of the factors a social scientist tries to "name" them based on the numerical findings. However, since there is not a specific way to perform this action, the result of this step is sometimes not properly elaborated. For all the above methods indicative references are Everitt (1993), Greenacre (2007), Jolliffe (2002) and Bartholomew and Knott (2011).

Apart from these well-known methods there are also some other methods equally important but less used. These methods are path analysis, structural equation modeling and multilevel modeling.

Path analysis is concerned with causation. Specifically it uses regression methods to identify patterns of causation in networks. In the beginning path analysis starts with a network of variables to specify the paths of causation. Usually, a cause and effect relationship assumes that there are a number of relationships and some variables that are believed to be caused by others, appear to affect other variables. A regression model cannot identify such a case because it can merely use one dependent variable. In path analysis all the necessary regression models considered, account for all the relationships needed.

Structural theory tries to give the structural relationships between constructs. This theory is represented by a structural model using a number of equations. These equations are usually accompanied by a proper diagram indicating the relationships. In other words, structural equation modeling is a method that tries to estimate the relationship between latent variables. This relationship can be linear or non-linear. The advantage of this method is that it allows us to test hypotheses on the relationships between observed variables and latent variables and also between the latent variables themselves.

Multilevel modeling is used to analyze data involving clusters. Specifically, in social research we are often concerned with the relationship between individuals and the groups they belong. This relationship actually leads to nested data, that is individuals nested within groups. For example in education students are nested within schools. The performance of a student in a series of exams could be affected by both characteristics of the student and of the school he/she attended.

For path analysis, structural equation modeling and multilevel modeling, the interested reader can refer to Agresti and Finlay (2009), Bartholomew *et al.* (2008) and Timm (2002).

## 7. Categorical data

Usually in social sciences researchers have to analyze categorical data. A categorical variable can take a limited number of specific discrete values. Usually such values occur for example when respondents are assigned in groups or when a property holds or not. In social sciences the different categories of a categorical variable often measures attitudes and opinions.

Categorical variables with a natural ordering are called ordinal variables. Categorical variables without ordering are called nominal variables. Methods designed for ordinal variables cannot be used with nominal variables due to the fact that nominal variables do not have ordered categories. Methods designed for nominal variables can be used with nominal or ordinal variables, since they only require a categorical scale.

The most famous models for analyzing categorical data are logistic regression models. Logistic regression can be used with continuous and discrete predictors (Agresti (2007)). Loglinear models are used to analyze associations among multiple categorical response variables. A log-linear model can be transformed using logarithm to a polynomial function of the parameters of the model. This is very helpful since the researcher can use linear regression (Azen and Walker (2011)).

A broad class of models is the generalized linear models. These models are a generalization of ordinary linear regression in the sense that it allows the distribution of the error to be different from the normal distribution. Another class of models is those that are used to analyze repeated measures data or longitudinal data. That kind of data is repeated observations of the same variables over several periods of time. One feature that must be taken into consideration is that data are correlated since the same subjects are measured over time (Lawal (2003)).

We may say that in general researchers in social sciences could rely more on the abovementioned models for the analysis of categorical data. These models are not very popular among researchers who tend to rely more on descriptive measures. We believe that the practitioners in the area could benefit a lot from the already developed methods.

## 8. Time series

Time series is a sequence of observations on a variable of interest with chronological order. That kind of data is quite natural in some of the fields in social sciences like economics. The observations in a time series are considered dependent. Time series analysis is a collection of techniques for the analysis of this time dependence.

There are a lot of different approaches to handle time series data. A first approach is to use the autoregressive models or the moving average models. The autoregressive model (AR) assumes that there is linear dependence of the variable we study with its own previous values. The moving average (MA) model is a linear regression of the current value of the series against current and previous (in terms of time) error terms.

Another class of models are the autoregressive moving average (ARMA) models. We use the notation ARMA($p$, $q$) to define a model with $p$ autoregressive terms and $q$ moving-average terms. A generalization of this model is the autoregressive integrated moving average (ARIMA) model. This model is generally referred as ARIMA($p$, d, $q$) where parameters $p$, d, and $q$ are non-negative integers that refer to the order of the autoregressive, integrated and moving average parts of the model respectively. All the above mentioned models (AR, MA, ARMA, ARIMA) form among other techniques the Box-Jenkins method for modeling time series. For more details Box *et al.* (2008).

Another class of time series models, especially useful in econometrics, are the autoregressive conditional heteroskedasticity (ARCH) models. In ARCH models we assume

that the variance of the current error term is a function of the actual sizes of the previous time periods' error terms. ARCH models have been extensively used to model financial time series. A generalization of the ARCH models is the generalized autoregressive conditional heteroskedasticity models (GARCH). In GARCH models we assume that the error variance is modeled by an ARMA model. There are a number of newer model proposals based on ARCH and GARCH models. The interested reader about ARCH models can refer to Xekalaki and Degiannakis (2010).

Another interesting characteristic in time series is forecasts. Apparently, it has attracted the interest of researchers in various fields. Several techniques on this very interesting issue have been proposed. Methods and examples of applications are given in Bisgaard and Kulahci (2011).

The research in social sciences, using the already stated models for time series, mainly appears in economics and marketing. We strongly believe that researchers in other areas of social sciences could benefit from these models also.

## 9. Data mining

Data mining is a collection of techniques used to find patterns in a set of data. They are extremely important in the analysis of large data sets of social phenomena. Other names that refer to the same collection of techniques are machine learning and predictive analytics. During the last years there is an increasing interest in these techniques although most of them are known for decades. We have to note here that the use of a computer is compulsory to run these techniques and moreover that if we have large data sets the larger the amount of data the more computational power we need.

The computational methods that comprise the field of data mining derive from the areas of statistics and artificial intelligence. These techniques are used to find meaningful associations between related variables usually between a large number of variables. These structures help the practitioner to draw useful conclusions about his/her research questions.

An important feature that is one of the objectives of a data mining analysis is the generalization of the results. To be more specific, if after an analysis of the data at hand using data mining techniques we conclude that there are some important patterns, then we would also like to find that these patterns exist and in the data that we will gather in future. This generalization is very important for drawing conclusions that are irrespective of the collected data the specific time we run the analysis.

If we consider the predictive dimension of data mining we can refer to the two important conclusions of such an analysis. The first conclusion is that after we reach a useful and meaningful model we can use it to predict the variable under study using some or all the remaining variables. Obviously, such a conclusion gives the researcher the ability to compute the values of the dependent variable given the values of the independent variables. The second conclusion is that the researcher is able to comment about the relationship between the dependent and the independent variables.

Another characteristic we need to highlight is the need to know as much as possible about the data and the process. The definition of the variables, the way they are measured and their interrelation in terms of the case studied are extremely important to the researcher to assist him reach a meaningful conclusion. Additionally, since the data are most of the times in vast numbers there is the need to store, process and compute them. Therefore, it is highly probable that knowledge of databases and parallel computing will be compulsory for the application of data mining techniques.

Keeping in mind the vast amount of social data that are gathered in today's world using classical ways (e.g. questionnaires) along with the use of mobile technologies, social

networks, texts, photographs, videos and all the different types of human activities that are transformed to data we can easily conclude that it is not a rare event to have to analyze thousands of variables with many cases in each of them. In such cases we can say that we end up with big data (data with high volume, high velocity and high variety). This fact highlights the need to use data mining techniques that can handle such amount of data. More information and detailed representation of data mining techniques can be found in Hastie *et al.* (2009) and Azzalini and Scarpa (2010).

## 10. Conclusions

In this paper, we reviewed some statistical methods useful in the area of social sciences. Sampling techniques, regression analysis, analysis of variance, parametric and nonparametric models along with multivariate methods were presented. Categorical data analysis techniques, time series methods and data mining were also presented. Indicative references in all of these areas are also given.

Statistical methods have played a very important role in social sciences. In every applied research effort statistical techniques are compulsory to reach a non-questionable conclusion. We strongly believe that advanced statistical methods can be employed heavily in this area. It seems that researchers rely more on classical statistical methods although they could benefit from the use of newer and advanced techniques.

## References

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, John Wiley, New York, NY.

Agresti, A. and Finlay, B. (2009), *Statistical Methods for the Social Sciences*, 4th ed., Pearson/Prentice Hall, NJ.

Azen, R. and Walker, C.M. (2011), *Categorical Data Analysis for the Behavioral and Social Sciences*, Routledge, New York, NY.

Azzalini, A. and Scarpa, B. (2010), *Data Analysis and Data Mining*, Oxford University Press, New York, NY.

Bartholomew, D.J. and Knott, M. (2011), *Latent Variable Models and Factor Analysis*, 2nd ed., Vol. 7, Kendall's Library of Statistics, Arnold.

Bartholomew, D.J., Steele, F., Moustaki, I. and Galbraith, J.I. (2008), *Analysis of Multivariate Social Science Data*, 2nd ed., CRC Press, New York, NY.

Bisgaard, S. and Kulahci, M. (2011), *Time Series Analysis and Forecasting by Example*, John Wiley, New York, NY.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008), *Time Series Analysis: forecasting and Control*, 4th ed., John Wiley, New York, NY.

Christensen, R. (2011), *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer.

Cohen, J. (1988), *Statistical Power for Behavioral Sciences*, Lawrence Erlbaum Assoc., Mahwah, NJ.

Cohen, B.H. and Lea, R.B. (2004), *Essentials of Statistics for the Social and Behavioral Sciences*, John Wiley, New York, NY.

Corder, G.W. and Foreman, D.I. (2009), *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, John Wiley, New York, NY.

Dodge, Y. (2008), *The Concise Encyclopedia of Statistics*, Springer.

Everitt, B. (1993), *Cluster Analysis*, Arnold, London.

Greenacre, M.J. (2007), *Correspondence Analysis in Practice*, Chapman and Hall, Boca Raton.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: data Mining, Inference, and Prediction*, Springer, New York, NY.

Jacob, R. and Zhu, P. (2012), *A Practical Guide to Regression Discontinuity*, mdrc.

Jolliffe, I.T. (2002), *Principal Component Analysis*, 2nd ed., Springer, New York, NY.

Joseph, L., Burger, R.D. and Belisle, P. (1997), "Bayesian and mixed bayesian/likelihood criteria for sample size determination", *Statistics in Medicine*, Vol. 16 No. 7, pp. 769-789.

Kish, L. (1995), *Survey Sampling*, John Wiley, New York, NY.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005), *Applied Linear Statistical Models*, McGraw-Hill/Irwin, New York, NY.

Lawal, B.H. (2003), *Categorical Data Analysis with SAS and SPSS Applications*, Lawrence Erlbaum Associates, NJ.

Lenth, R.V. (2001), "Some practical guidelines for effective sample size calculations", *American Statistician*, Vol. 55 No. 3, pp. 187-193.

Lovric, M. (2011), *International Encyclopedia of Statistical Science*, Springer.

Rencher, A.C. and Schaalje, G.B. (2008), *Linear Models in Statistics*, John Wiley, New York, NY.

Riley-Tilman, T.C. and Burns, M.K. (2009), *Evaluating Educational Interventions*, The Guilford Press, New York, NY.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley, New York, NY.

Shuster, J.J. (1990), *Handbook of Sample Size Guidelines for Clinical Trials*, CRC Press, Boca Raton, FL.

Timm, N.H. (2002), *Applied Multivariate Analysis*, Springer, New York, NY.

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer, New York, NY.

Xekalaki, E. and Degiannakis, S. (2010), *ARCH Models for Financial Applications*, John Wiley, New York, NY.

**Corresponding author**
Petros Maravelakis can be contacted at: maravel@unipi.gr