Statistical Techniques for Research

Jose Carlos Casas-Rosal
https://orcid.org/0000-0002-6954-6517
Universidad de Córdoba, Spain

Carmen León-Mantero
https://orcid.org/0000-0002-2870-8807
Universidad de Córdoba, Spain

Noelia Jiménez-Fanjul
https://orcid.org/0000-0002-5728-8725
Universidad de Córdoba, Spain

Alexander Maz-Machado
https://orcid.org/0000-0002-4112-4363
Universidad de Córdoba, Spain

INTRODUCTION

Data analysis of a study and its subsequent statistical processing represent the base of numerous research works carried out in any scientific field. The wide variety of techniques allow an in-depth analysis of any data set depending on the desired goal – such goal has to be taken into account at the beginning of the study and it has to assume central importance during the research design, as it requires the verification of a set of hypotheses. The refutation of one or more hypothesis needed for the application of a particular statistical technique may lead to inconsistent or invalid results, which would in turn invalidate the whole research.

The vast majority of techniques applied in research is aimed at inferring the results of a sample to the target population of the study, which at the same time entails a certain degree of uncertainty – this degree of uncertainty can be controllable and measurable as long as the application conditions of the techniques are met. Its refutation may involve from a slight increase of the degree of uncertainty to its immeasurability. It is crucial for the researcher to possess a proper knowledge of the technique so as to obtain the results and the conclusions of his investigation.

However, and due to the great variety of techniques, its misuse in research is very common, either because of the refutation of any of the hypothesis set or because of the nature of the data that is being analysed. Because of that, the results obtained from its application cannot be taken into consideration due to its lack of validity.

Some of the main statistical techniques used in different research fields are presented in this chapter. A brief definition of each technique, together with the specification of its utility is included as well, in order to determine accurately its purpose. The required hypothesis regarding data and regarding experimental design are listed as well. Besides, some of the computer programs with which these techniques can be used are listed – we emphasize in R, a free and open-source software with multiple packages that allow to apply these techniques in an effective and simple way.

DOI: 10.4018/978-1-7998-3479-3.ch044

The analysis of the great amount of statistical techniques that exist would require a much larger extension than that available in this chapter, so the most used ones have been selected. Some of the simplest ones have been included (such as descriptive analysis, graphic analysis or parametric and non-parametric hypothesis testing for statistical inference), as well as some multivariate analysis techniques, such as principal component analysis, factor analysis, structural equation models, regression analysis and causal models.

BACKGROUND

The word *Statistics* comes from the Italian word *Statista*, meaning "Statesman" – referring to its first meaning as a science used by the Government to collect socio-demographic information from its population. From its introduction, it has experimented several extensions in its definition. During the nineteenth century, its field of action was enlarged to the "recollection of information from any data set", and therefore, its application to other disciplines was extended as well. Later, the analysis and interpretation of data was also considered as a part of the discipline.

However, even though the origin of the word dates from the eighteenth century, Statistics had already been used by the Babylonian and Egyptian civilizations in the elaboration of frequency tables and population censuses. Several authors consider four fundamental periods during the development of Statistics: a period up to 1750 is characterized by the development of probability and the exposition of non-probabilistic methods of data analysis; a period from 1750 to 1820 when inference and mathematical statistics were introduced with works from Laplace and Gauss; a period from 1820 to the early twentieth century with works from Galton, Pearson and Fisher, when statistical inference, correlation and statistical models were developed; and a fourth period, during the last third of the twentieth century, with a strong evolution of Computer Science (Fienberg, 1992). Since then, the amount of techniques developed has been significantly extended due to the calculation capacity of computers.

Therefore, descriptive analysis and the elaboration of representative graphics can be framed within the beginning of Statistics as a discipline, while regression analysis and statistical inference were developed during the nineteenth century, whose most relevant researchers in the field are shown in Figure 1. Other techniques were developed during the twentieth century. And most of those were developed during its second half, after the revolution that took place due to the calculation capacity of computers. Factor analysis was developed by Charles Spearman during 1904 and later on by Louis Leon Thurstone. Due to the evolution of this methodology, together with Biometrics, Econometrics and Computing, structural equation models were born during the 70's. Principal component analysis is due to Hotelling (1933). And by the end of the twentieth century, artificial neural networks started to be applied to numerous statistical problems such as estimation or classification.

This great development has favoured the appearance of a large number of research works that use statistics to obtain important results that permit correct decision-making in important fields such as Health Sciences, Engineering or Social Sciences. Kaplan-Meier estimator, for instance, was developed in 1958 (Kaplan & Meier, 1958) and has been used in Engineering, Medicine and Economy to estimate the survival of the components of a machine, the survival of patients of a certain disease or the time it takes for a person to return to the active market after losing his job.

Each and every one of the existing statistical techniques are based on a mathematical basis that allows to demarcate with exactitude the conditions of applicability of the technique and its degree of uncertainty, if any. The application of a technique under different conditions of those for which it was designed is