# INTRODUCING DATA ANALYSIS IN A STATISTICS COURSE IN ENVIRONMENTAL SCIENCE STUDIES

C. Capilla
Technical University of Valencia, Spain
CCAPILLA@EIO.UPV.ES

*Education in methods of applied statistics is important for students who will be involved in management and decision-making processes. This paper discusses issues related to the teaching of statistics to students enrolled in an undergraduate environmental science degree course. The aim is to describe the teaching of graphical and numerical methods for summarizing and exploring data obtained in environmental studies. The application of descriptive and exploratory methods provides useful information regarding the distribution of the data at hand and of its patterns and associations. These methods are presented at the beginning of the course, following an introduction to the steps involved in the process of learning from data through the use of statistics. Students are instructed in the reading and interpretation of graphic and numeric data summary techniques. The importance of visualizing the main patterns and associations in the data is emphasized using environmental examples.*

## INTRODUCTION

Statistical literacy is an important aspect in the training of Environmental scientists. In the particular case of studying environmental changes and associated problems, there is a need for professionals who are capable of identifying, analyzing and answering scientific questions related to environmental systems. The scientific approach to any environmental issue requires the correct application of statistical methodology to ensure well-conducted data collection, analysis and interpretation (Barnett, 2004). Training in applied statistical techniques is very useful for professionals engaged in environmental problems and management.

In recent years, universities have increased the number of places in Environmental Science degrees. In 1997, the Technical University of Valencia (Spain) established a Bachelor of Science (BSc) in Environmental Science. According to the statute this is a two-year specialization course. The entry to the course is gained on the basis of the marks obtained in previous studies and the minimum completion time is two years (four regular terms) for students who have completed the first three academic years of an engineering or science degree.

The main goal of the Environmental Science course is to prepare students for careers in environmental management and decision-making. In order to meet this requirement, a statistics course is required in their program of study. The statistics course was specifically developed for the environmental science degree and is taught within that specialization during the first term. The pedagogical approach and learning activities follow the proposals made by mathematicians and statisticians during the reform movement of mathematics education, which took place in the last two decades. The aim of this paper is to describe the teaching of data summary and exploratory analysis methods contained in this course. The contents, methodology and the difficulties faced by students are also discussed.

## ENVIRONMENTAL STATISTICS AND EDUCATION

The application of statistics in areas such as economics, industry or social science is well established. However, over the last fifteen years, the field of environmental statistics, also known as “environmetrics” (Hunter, 1994), has become one of the most rapidly growing areas of statistical applications (Guttorp, 2003). This has inspired an increasing number of activities in the area, has promoted an increment in the number of publications as well as journals and conference sessions devoted to environmental statistical applications and it has also contributed to improving specialization courses on the subject. Piegorsch and Edwards (2002) discuss the development of such courses. Their argument is based on their experience with the design of a graduate environmental statistics course. They indicate that the selection of material for an environmental statistics undergraduate course does not pose a problem. Although there is no consensus on core material, they recommend that an undergraduate course in basic statistics, for undergraduate

students whose interests include environmental problem-solving should include topics such as random sampling, basic summary statistics, basic probability and statistical distributions, confidence intervals, significance testing, correlation and regression. As they comment "the challenge in developing an environmental statistics course is to truly engage the practitioner/non-statistics students." An effort to reach a consensus about the syllabus of courses is only beginning in environmental statistics and more needs to be done.

Furthermore, the reform movement of mathematical sciences teaching during the last two decades has influenced statistical education. There has been a review of contents, pedagogy and technology at every level of education (Moore, 1997). Changes in these three domains have been proposed and adopted as means of continuously improving the teaching of statistics. The recommended changes may be summarized as follows:

- Highlighting connections between statistics and other sciences
- Understanding and using students' prior conceptions
- An emphasis on analyzing and interpreting data
- More active participation on the part of the students
- Solving real-world problems
- Small-group cooperative learning
- More technology and communicating skills regarding data and chance

The statistics education community has been working quite hard to develop new pedagogical approaches and activities to accommodate these changes (Garfield, 1995; Scheaffer *et al*., 1996; Moore, 1997).

## THE STUDENTS

Each year, approximately 70 students take the environmental statistics course. As they have already completed three academic years prior to their enrolment, they are more mature and responsible than students in their first year at university. They have taken other courses where they have been engaged in active learning, so they readily accept this teaching methodology. The majority of the students (over 60%) have a degree in agricultural technical engineering. The other students have a degree in civil or industrial technical engineering, or have completed the first cycle of the Biology or Chemistry degrees. Furthermore, about 44% have previously worked as technical engineers, outside the university. In their professional lives some of the students are involved in environmental issues, which require the application of statistical methods. These students contribute greatly to the course, sometimes giving examples of the environmental problems encountered during their work, and they also tend to participate more in the discussions.

A student's previous knowledge is one of the most important factors "to influence learning" (Batanero *et al*., 1994). All the students enrolling for the statistics course have studied calculus and algebra in their first terms at the university. However, the statistical backgrounds of the students range from no statistical experience whatsoever, to those who have completed an introductory statistics course in their previous studies (approximately 65% of the students). In order to assess the students' previous statistical knowledge, a multiple-choice test is given at the start of the course. The questions are selected and adapted from the Statistical Reasoning Assessment test described in Garfield (2003). The questionnaire provides scores to facilitate an evaluation of the students' reasoning on several basic statistical methods and concepts. Batanero *et al*., (1994) note that "the identification of errors and difficulties which students display is needed in order to organize statistical training programs and to prepare didactical situations which allow the students to overcome their cognitive obstacles."

The test results reveal that some students have difficulties and misconceptions when reasoning about the basic statistical concepts, which they ought to know following the completion of an introductory course. The most common difficulties students face when confronted with basic summary statistics (centre, spread and position measures) and with descriptive methods to study association (two-way tables and correlation) are the following: 31.6% of them think the average is the most common observation; 73.7% compare groups based on their averages. A similar percent do not understand sampling variability. A frequent error (71.05%) is to reason that small samples should resemble the population from which they are sampled, and to use small

samples as a basis for inference and generalizations. Finally, 55.3% do not correctly interpret two-way tables.

Taking into account these results, when developing the statistics course for the BSc in Environmental Science no assumption is made regarding previous statistical knowledge and experience. The course is designed with the aim of preparing the students "to use statistical thinking and reasoning" (Garfield, 1995), and to be able to interpret and evaluate environmental data analyses, rather than placing the emphasis on mathematical proofs and calculations. Derivations and hand calculations are replaced by the widespread use of a statistical computing packages and instead, the emphasis is placed in the understanding of concepts and having a hands-on approach to data analysis.

## THE PROCESS OF DISCOVERING KNOWLEDGE

The official program of the environmental science degree in Spain gives the following keywords for the compulsory statistics course: :statistical distributions, sampling, estimation, hypothesis testing, regression and correlation." The statistics course at the Technical University of Valencia (Spain) was designed to include these topics, as well as taking into consideration the student's background and the recommendations given by Moore (1997) about teaching methodology and by Piegorsch and Edwards (2002) about the contents of the course.

The methods and concepts are introduced using environmental examples. These examples provide a rich source of situations to help illustrate both the need for and the use of tools in the discovery of observation-based knowledge. During the introduction to the course, the phases of learning from data are explained: a clear statement of the problem, data collection and initial exploration, model formulation, tentative model fitting and inference, validation of assumptions and the use of the model to make predictions and decisions regarding the problem. The important role of the data behind the sequential learning process is stressed, with an alternation of the induction and deduction steps in the statistical modelling approach. As Professor Box (1999) states, "Discovery of new knowledge requires the use of the scientific paradigm in which the model is continually changing."

The goal is for the students to understand that a model is a useful approximation to reality, but should never be considered as the definitive answer. Another important aspect that should be emphasized to students when dealing with specific statistical methods during the course is that some statistical problems can be viewed from different perspectives and that there is not always one right answer. On the other hand, the course is also designed with the aim of motivating the students to use a variety of statistical techniques to solve problems in-class activities, as well as in assignments, thereby, developing connections between the different topics.

## TEACHING DESCRIPTIVE METHODS

Descriptive statistical methods are studied in the first module. The subject is taught during five weeks of the term. The teaching takes place on different days of the week and is divided into three types of sessions. First, there is a two-hour large group lecture per week where the teacher explains the basics of the topic, motivating the need to learn new concepts. The methodology used is the study of a real problem, drawn from the environmental science area and that stimulates the active participation of the students in the discussion. At the end of the lecture the students are assigned to a formal team of two or three students, with the sole objective of reviewing the concepts just explained in class. The teams are kept for cooperative working in the three sessions during the term. The idea behind teamwork is that students understand the benefits of working together with others. Group work will also play an important role in their professional lives after leaving the university.

The second session lasts an hour and is taught once a week. It involves group participation in activities or discussion of environmental case studies introduced by the teacher using statistical software. There is a personal computer in the classroom with connection to the university network, which has access to the statistical software and video projection. These are valuable tools since they easily display graphs and the results of calculations that help students to intuitively understand the theoretical concepts explained.

In the descriptive analysis module, there are three sessions of the third type. Each one lasts two hours. The large group is divided into six subgroups that occupy the computer laboratory with different schedules. The subgroup size is smaller than 12 students. The teams carry out practical tasks using statistical software to solve problems. A written report must be prepared at the end of this session, including a brief summary of what they have understood about the main concepts involved in the laboratory task, and the interpretation and discussion of the analysis.

Students use in the three sessions teaching materials that contain the educational modules, the activities for the classroom and the computer laboratory tasks. There is no need for extensive copying. These materials have been prepared by the teachers involved in the course and are available on the class web page. Students may study in depth specific statistical techniques using additional references, e.g., Gilbert (1987), Manly (2001), and Berthouex and Brown (2002) that give detailed treatment of statistics applications in environmental systems.

## DESCRIBING AND EXPLORING UNIVARIATE DATA

Descriptive and exploratory data analysis methods are useful for revealing the main characteristics of the data distribution and its patterns and associations. The course starts with univariate summary and exploratory methods. Concepts such as population, random sample, data types or variability are introduced. Different types of observations are discussed: quantitative or qualitative, discrete or continuous, the importance of recognizing these and selecting the appropriate method for summarizing the sampled information.

Univariate frequency and relative frequency tables are the first techniques to be applied. Their graphic representation using bar charts or pie charts, when the data are qualitative, is studied. The histogram is introduced as the plot of the frequency table of quantitative observations. Students use histograms, interpreting the information contained in these concerning the data distribution: position, spread and shape. The application of histograms to detect outliers, mixtures of populations or other patterns in the sample, which may affect the conclusions, is discussed.

Using examples of environmental quantitative data, measures of central tendency, such as the mean and median, are introduced. It is discussed when these measures are more appropriate for use and when they effectively reflect the central tendency, and the problems of the mean when outliers are present in the sample. The concept and use of the trimmed mean when outliers are present in the data is included in this module. The quartiles and quantiles of a sample are also studied. Percentiles are introduced using examples of environmental standards formulated in reference to these.

Using several examples to illustrate the point, it is shown that it is also useful to calculate spread measures in order to complete a data summary. The range is introduced as a simple measure of variability for small samples without outliers. The sample variance and standard deviation are discussed and found to be appropriate in those situations where the sample mean satisfactorily represents the centre of the distribution. A study of the interquartile range as a more robust measure of variability is undertaken. The coefficient of variation to evaluate the relative dispersion, which allows a comparison of variability of data with different units, is introduced. In the last part of the sampled parameters, the two shape measures - skewness and kurtosis are studied. Their interpretation and limitations when characterizing sample distribution shapes are discussed.

The box-whisker plot is used as a tool to chart some of the central tendencies and spread measures, and to detect outliers. The importance of interpreting the information that the plot contains and relating it to the position, spread and shape measures is emphasized. Many students have difficulties in reading and recognizing the characteristics of a distribution from a plot. Therefore, several activities are developed in the classroom and in the computer lab to improve these skills. Data sets are used to illustrate the application of these methods in the comparison of data distributions.

Some examples of univariate time series and their descriptive analysis are also discussed. The time sequence plot is studied as a useful method for visualizing temporal variations and data

components (trend, seasonal and irregular components). The seasonal subseries plot is introduced to facilitate the analysis of data exhibiting seasonality.

## EXPLORING ASSOCIATIONS IN BIVARIATE DATA

Association and correlation concepts are discussed in the bivariate descriptive analysis module. Using several examples, it is shown during the interpretation of these concepts that they do not necessarily indicate causation (Batanero *et al*., 1994; Garfield, 1995). In the case of bivariate qualitative variables, the two-way frequency table is applied. Students have difficulties when reading and interpreting the conditional relative frequencies in these tables. Several analyses are performed during the activity hours to facilitate the students' understanding of this tool. When analyzing quantitative bivariate variables, the covariance and correlation measures are applied. The information contained in these is related to the interpretation of the scatter plot, a useful tool for the clear demonstration of linear and nonlinear relationships, and outliers or other anomalies in data.

The concepts of association and correlation are extended to the multivariate case. In this module, the emphasis is placed on graphical data presentation methods to analyze patterns and relationships between variables. Visualization is critical in the understanding of environmental data and the results derived from their statistical analysis. Moreover, as Weldon (2002) points out: "While multivariate analysis has formidable mathematical problems associated with it, some very simple graphical strategies are possible to convey in a first course. Use of these plotting methods in an early service course may be one way to alert students to the very common multivariate nature of real-life data."

## SOME ACTIVITIES

During the course students conduct the activities in small groups of two or three people. These activities may or may not require the use of statistical software. In the first instance, the work takes place in the computer room, and the groups have to write a summary giving the main conclusions of their analyses. All the activities are related to authentic environmental examples or to data generated by the students. Some examples are given below.

Example 1: This activity has been built up around one proposed in Scheaffer *et al.* (1996). At the beginning of the course, students are asked to answer an anonymous questionnaire with the aim of obtaining their personal information. This contains variables such as age, gender, weight or height, or their opinion about which environmental problem is more critical, the environmental factor which they consider more important where they live, etc. Data files are created using their responses. A discussion takes place regarding whether or not the students can be considered a representative sample for the questions posed and, in addition, they are asked to recognize the different types of data involved. The file is used in several activities in the computer lab. For example, after the two-way tables have been introduced in the classroom, the students have to answer an exercise in which they are required to study if a relationship between gender and the chosen critical environmental problem exists.

Example 2: The Technical University of Valencia is a member of an international network of environmental groups, universities and other educational establishments, who in turn work with local groups and individuals around the coast of Europe. Common aims are: Protection and sustainable use of our coastal resources; informed public participation in environmental planning and management, including coastal zone management. Volunteers across Europe carry out an annual survey on the coast. The database with that information is used in several examples to motivate the application of descriptive methods.

Example 3: There is a monitoring network in Valencia, which continuously measures several atmospheric variables (pollutants and climatological parameters). Data are available on the internet. Firstly, students are asked to download the data. Then, they develop a project working in small groups. A group may have to descriptively analyze the relationship between ozone and nitrogen oxides and meteorological parameters such as solar radiation, wind speed and temperature. To provide the answer, they have to apply several graphical and numeric descriptive tools, recognizing which ones are more appropriate and, finally, interpret the results. Examples of graphical tools they have to apply are the coplot, the bubble chart or scatterplot matrices.

The analysis of real data helps the students to view statistical methods as valuable tools for decision making in environmental systems. These activities also allow the development of connections with other compulsory subjects of the Environmental Science degree, such as Atmospheric Pollution or Meteorology and Climatology. They are carried out under the guidance of the teacher who adopts the role of "facilitator" (Moore, 1997) and motivates statistical concepts, activities and problems.

ACKNOWLEDGEMENTS

The author would like to thank the R+D+i Linguistic Assistance Office at the Technical University of Valencia and Dr. Liliana González, for their help in revising the English language. The author is also indebted to Prof. Joachim Engel for constructive comments, which have contributed to improving this paper.

REFERENCES

Barnett, V. (2004). *Environmental Statistics*. Chichester, UK: Wiley.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., and Holmes, P. (1994). Randomness, its meanings and educational implication. *International Journal of Mathematical Education in Science and Technology*, 29, 113-123.

Berthouex, P. M. and Brown, L. C. (2002). *Statistics for Environmental Engineers* (Second edition). Boca Raton, FL: CRC Press.

Box, G. E. P.(1999). Statistics as a catalyst to learning by scientific method. Part II- A discussion, *Journal of Quality Technology*, 31, 16-29.

Garfield, J. B. (1995). How students learn statistics, *International Statistical Review*, 63, 25-34.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistical Education Research Journal*, 2, 22-32.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring.* New York: Van Nostrand Reinhold.

Guttorp, P. (2003). Environmental statistics- a personal view. *International Statistical Review*, 71, 169-179.

Hunter, J. S. (1994). Environmetrics: An emerging science. In G. P. Patil and C. R. Rao (Eds.). *Handbook of Statistics XII: Environmental Statistics*. Amsterdam: North Holland.

Manly, B. F. J. (2001). *Statistics for Environmental Science and Management.* Boca Raton, FL: CRC Press.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, 65, 123-137.

Piegorsch, W. W. and Edwards, D. (2002). What shall we teach in environmental statistics?. *Environmental and Ecological Statistics*, 9, 125-150.

Scheaffer, R. L., Gnanadesikan, M., Watkins, A., and Witmer, J. (1996). *Activity-Based Statistics.* New York: Springer-Verlag.

Weldon, K. L. (2002). Advance topics for a first service course in statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.