THE ROLE OF STATISTICAL SOFTWARE IN TEACHING DATA ANALYSIS

Carmen Capilla[1]

[1]*Polytechnic University of Valencia (SPAIN)*

Abstract

Teaching statistics at the university level is constantly changing due to the influence of modern technology. Statistical software for computations and visual representations, enable students' active knowledge constructions by "doing" and "seeing" statistics. In this paper a case study of teaching statistical data analysis using software is described. The environmental science degree at the Polytechnic University of Valencia (Spain), includes a statistics course whose contents mainly are data analysis tools. Solving environmental problems requires the proper application of data analysis methods. This paper presents the methodology used to teach this subject. The use of software is an important part of the course. The ten computer laboratory activities of the course are described and some examples are given. Assessment results of these classes are analyzed. The comparison of students' average marks in the evaluations shows that they are significantly smaller in the sessions devoted to univariate descriptive analysis, discrete probability methods, and introduction to advanced data analysis. There are significant differences between the average marks of students. They are heterogeneous in their backgrounds which is an important factor to influence the teaching-learning process of the course.

Keywords: Data analysis, statistical software, environmental data, undergraduate students.

1 INTRODUCTION

The Polytechnic University of Valencia (PUV) created the environmental science degree in 1997. Since then students with different backgrounds have taken the courses. The degree program is interdisciplinary and is associated to the Civil Engineering School. A compulsory statistic course is included. The course is taught in the first semester of the fourth degree year. The total number of hours to teach theory and practice is 60.

The statistics course main objective is to teach students methods to obtain and efficiently analyze environmental data. Present global and local environmental problems require the application of scientific methodology to take appropriate decisions [1]. Environmental statistics has extensively been developed during the last decades [2]. The great demand of professionals in this field has made universities to design and offer degrees of specialization on the subject. Several authors discuss which contents should be included in these courses for undergraduate and graduate levels [3]. A scientific approach to environmental issues requires properly applied statistical methodology to ensure well-conducted data collection, analysis and interpretation. An undergraduate course in basic statistics for students whose interests include environmental problem solving should cover such topics as random sampling, basic summary statistics, basic probability and statistical distributions, confidence intervals, significance testing, correlation, and regression.

Teaching statistics has to be designed to highlight its connections with other sciences. Statistics education researchers ([4], [5]) stress the importance of understanding and using students' prior conceptions, encouraging their active involvement and small group cooperative learning, and using real problems to emphasize the analysis and interpretation of data. Students' technology and communicating skills regarding data and chance have to be improved. Software adequate features for supporting learning and doing statistics in courses, are easy-to-use and learn menu-driven program with high-quality graphics, implementing the basic statistical methods and simulation tools [6].

This paper describes the experience of using statistical software for teaching basic statistical methods in the environmental science degree of the PUV. The next section explains the course computer laboratory sessions and gives some examples of the in-class activities. The assessment results are analyzed and the final section contains finally some concluding remarks.

2 COMPUTER LABORATORY CLASSES

2.1 CONTENTS OF THE CLASSES

The course is taught during the 15 weeks of the first semester. There are lecture classes of three hours every week. In these sessions the basic concepts are explained using real data examples, and exercises are solved to illustrate the methods application. There are also ten sessions during the term that take place in the computer laboratory. Approximately 40 students follow the course. The group is divided in three subgroups. Two of them have the practical classes on Tuesday morning and the third one on Friday afternoon. In these classes they usually work in two people teams. The contents of the ten sessions are given in Table 1.

Table 1. Statistical concepts applied in the computer laboratory classes

Class	Contents
1	Unidimensional and bidimensional frequency tables
2	Univariate descriptive analysis for quantitative variables
3	Bivariate descriptive analysis for quantitative variables
4	Discrete probability models: Binomial and Poisson distributions
5	Continuous probability models: Normal, lognormal, extreme value distributions
6	Inference on the mean and standard deviation of a normal population, and on the correlation coefficient
7	Comparison of means and standard deviations of two normal populations
8	Analysis of variance: effects of two factors and application in bietapic sampling
9	Linear regression models
10	Introduction to descriptive analysis of time series and principal components analysis

The exercises have to be solved using the two statistics programs available at the PUV campus: Statgraphics and SPSS. In the first part of the class the teacher introduces the software options to make the graphics and computations, using an example already explained in the lectures. Emphasis is placed on the interpretation of the software outputs. In the second part of the session the teams have to answer the questions of a real data case study. They have to elaborate a written report. The deadline to present the report is 15 days after the session date. The reports are corrected by the teacher and returned to the students, to give them a feedback of their work. The data used in the activities are air quality observation of the city of Valencia, meteorological observations, water quality data, and exercises which appeared in some of the books recommended to the students to follow the course ([7], [8]). The examples allow making connections with other subjects of the degree. The course materials (powerpoint presentations, activities, exercises) are available in the PoliformaT platform of the PUV created with the Sakai project.

2.2 SOME EXAMPLES OF ACTIVITIES

2.2.1 Descriptive analysis of the weekly average level of a pollutant in an urban area

The objective of this activity is that students apply the univariate descriptive methods to analyze location, dispersion and shape of quantitative distributions at a descriptive level. They have to use graphical and numerical methods. Both programs, Statgraphics and SPSS have to be used. Weekly

observations of atmospheric sulphur dioxide (SO_2) in a Valencia monitoring site, are in a file they have to open in the data editors of the programs. The sampling period is a year. Several questions are then posed in relation with the problem.

Question 1: Compute the location, dispersion and shape parameters of the weekly concentration of SO_2.

The output that students obtain using Statgraphics is:

Summary Statistics for SO2

Count = 52
Average = 26,75
Median = 25,5
Variance = 48,6618
Standard deviation = 6,9758
Minimum = 12,0
Maximum = 44,0
Range = 32,0
Lower quartile = 22,0
Upper quartile = 30,0
Interquartile range = 8,0
Stnd. skewness = 0,797089
Stnd. kurtosis = 0,00403502

They have to make connections of these parameters with the information contained in the univariate descriptive plots: box and whisker plot and histogram.

Question 2: Interpret the information on the data sampling distribution that gives the Box and Whisker plot (Fig. 1), taking into account the parameters values.

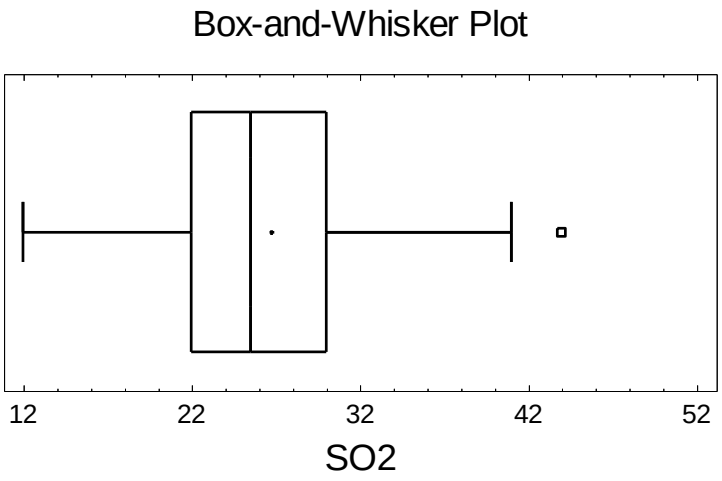

Fig. 1. Box and Whisker plot of SO_2 weekly concentrations.

Question 3: Represent the histogram of the data. Use an adequate number of intervals taking into account the sample size (Fig. 2).

Question 4: How is the sample distribution? Is it asymmetric? Are there outliers? What parameters and dispersion would you chose to summarize the data distribution?

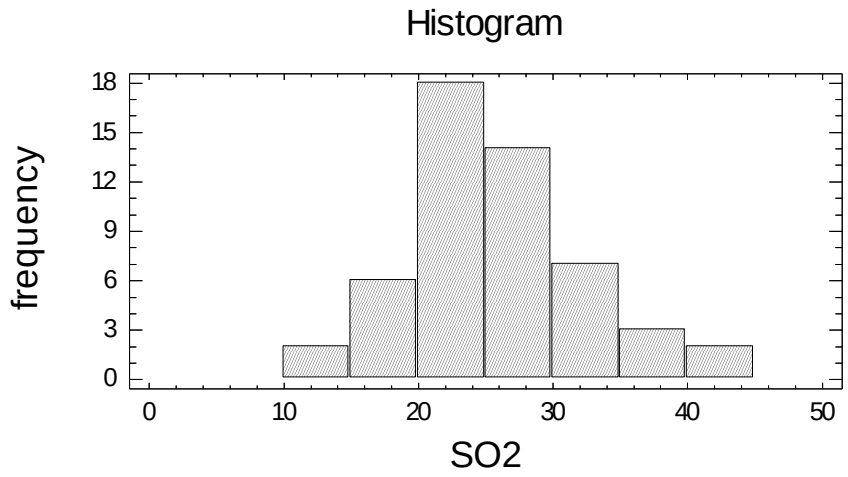

Fig. 2. Histogram of the SO_2 weekly data.

2.2.2 Probability distribution of survival times to exposure to a toxic product

The objective of this activity is that students use probability plots to identify the distribution model that best fits he data. The sample data are taken from [9]. They were obtained in an experiment in which 50 mice were exposed to a toxic product during 105 weeks. At the end if the experiment 32 mice had died and 18 survived. (censored data). The observations were:

50, 73, 76, 79, 81, 81, 83, 85, 85, 85, 87, 89, 90, 91, 91, 94, 94, 95, 95, 97, 98, 99, 100,100, 100, 100, 102, 102, 103, 103, 104,104

Question 1: Represent the probability plots of the data, and indicate in which plot the fit is better. What probability distribution can be used to model the variable?

Fig.3 shows the normal probability paper of the sample, after indicating that there are 18 data above the maximum value.

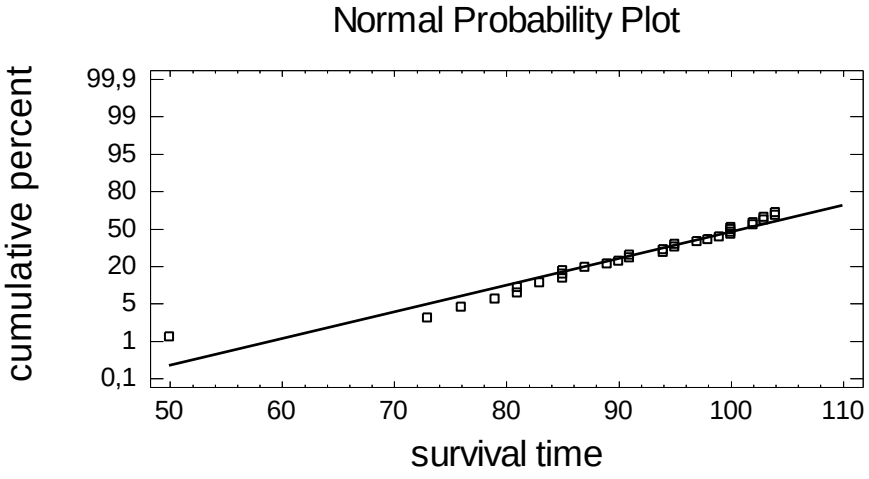

Fig. 3. Normal probability plot of survival time data.

Question 2: Estimate approximately the distribution parameters using the plot.

As the normal distribution is appropriate, they have to estimate the mean and standard deviation using the location option of the graph. Fig. 3. Show the estimation of the mean using the 50% cumulative percentage.

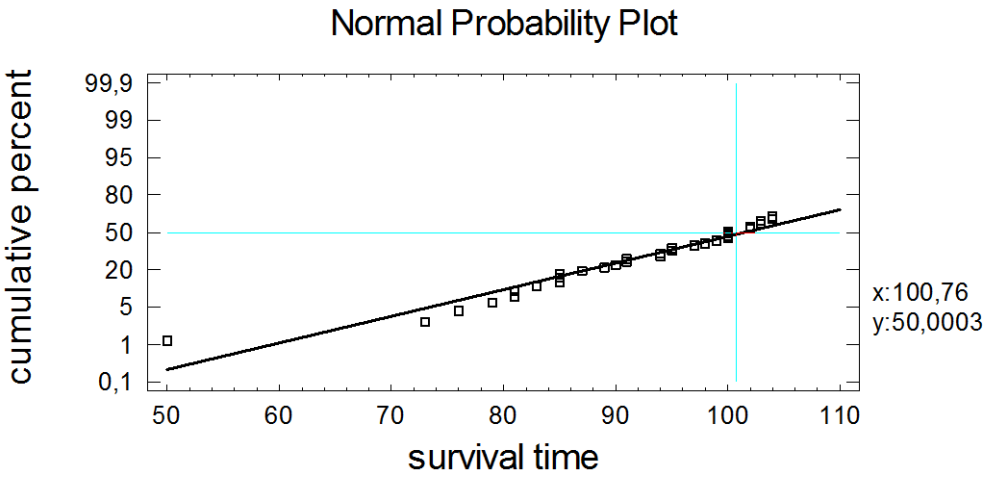

Fig. 4. Estimation of the distribution mean.

Question 3. What percentage of mice will have a survival time above 80 days? Estimate this percentage using the graph.

In Fig. 5 The plot is used to estimate the percentage (11,65%) below 80 days. The complementary (88,35%) is the answer to question 3.

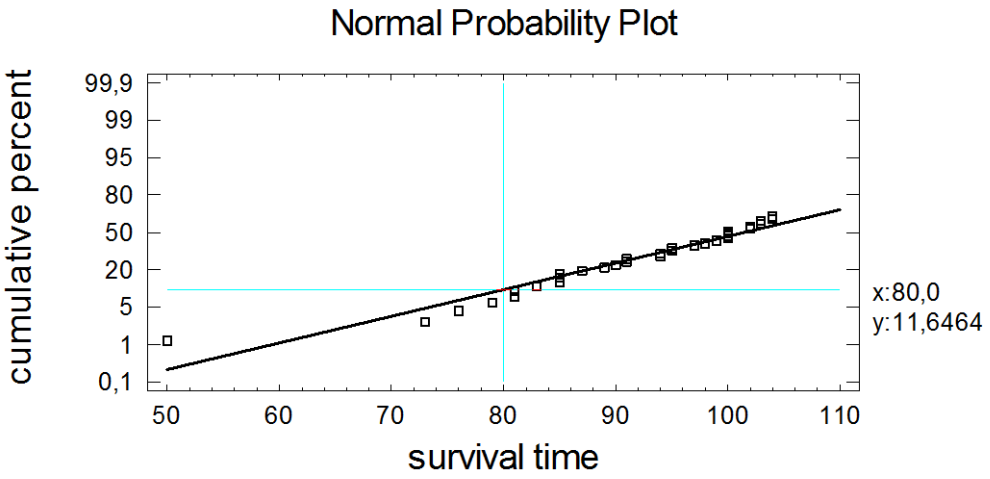

Fig. 5. Estimate of a percentage using the plot.

2.2.3 Inference on the mean and standard deviation of sulphur dioxide in a forest

The objective of this exercise is that students apply the inference methods (hypothesis tests and confidence intervals) for the mean and standard deviation of a normal population. They use sample data of sulphur dioxide concentrations in a forest area [10].

Question 1: Analyze if the data follow the normal distribution (Fig. 5)

Question 2: The average sulphur dioxide concentration is 20 mgr/m^3 in areas that are not affected by acid rain. Is the study area affected by this problem?

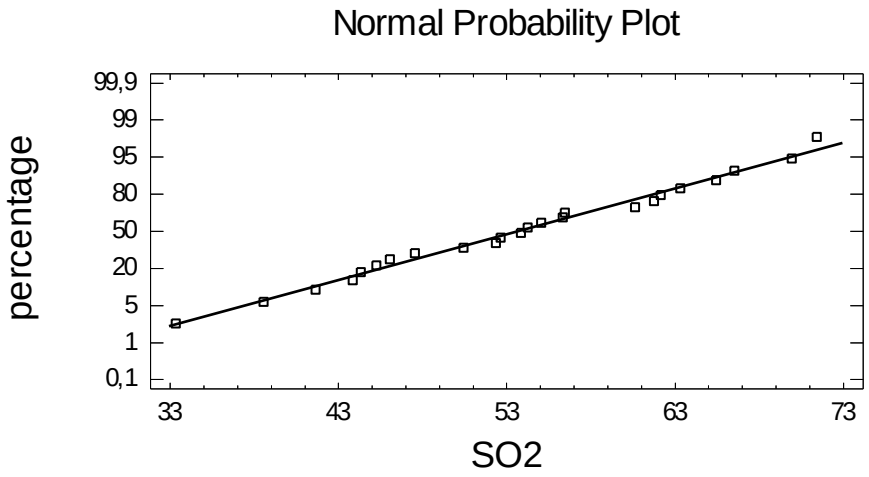

Fig. 5. Normal probability plot of the data sample.

The answer this question with the hypothesis test for the mean with the null hypothesis m=20. The software results are:

Hypothesis Tests for SO2

Sample mean = 53,9375
Sample median = 54,1

t-test

Null hypothesis: mean = 20,0
Alternative: not equal

Computed t statistic = 16,5173
P-Value = 0,0

Reject the null hypothesis for alpha = 0,05.

Question 3. Compute the confidence intervals for the mean and standard deviation.

Confidence Intervals for SO2

95,0% confidence interval for mean: 53,9375 +/- 4,25041 [49,6871;58,1879]
95,0% confidence interval for standard deviation: [7,82326;14,1199]

3 ASSESSMENT RESULTS

The computer laboratory sessions are assessed with the reports that students present. Their average mark account for 35% of the total final mark in the subject. An analysis of variance of the marks obtained during the last academic year, shows that there are significant differences between students and activities (Table 2). Fig. 6 represents the least significant differences plot for students. Two students have had a significantly lower average of the computer classes assessments marks. The comparison of assessment activities marks can be seen in Fig. 7. Activities 2 (univariate descriptive analysis), 4 (discrete probability distributions) and 10 (introduction to advanced data analysis) have worse results. The concepts that students have to apply in these activities are more difficulties especially in activity 10. The correlation between the average mark in the computer laboratory activities and the final exam mark is significant and equal to 0,64. The computer laboratory classes allow improving students' ability with technologies, and placing emphasis to the interpretation of analysis results. They comment at the end of tje course, that compter class activities have helped them more to understand the statistical concepts than the in class activities done in the lectures.

Table 2. Analysis of Variance of the marks by student and activity.

Analysis of Variance for Mark

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
Student	369,063	30	12,3021	2,53	0,0000
Activity	136,272	9	15,1413	3,12	0,0014
RESIDUAL	1312,27	270	4,86024		
TOTAL	1817,6	309			

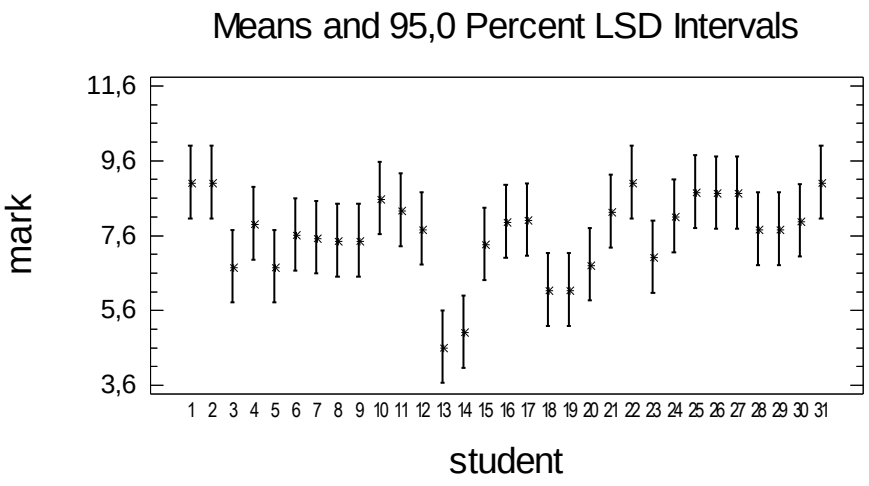

Fig 6. Comparison of average marks by student

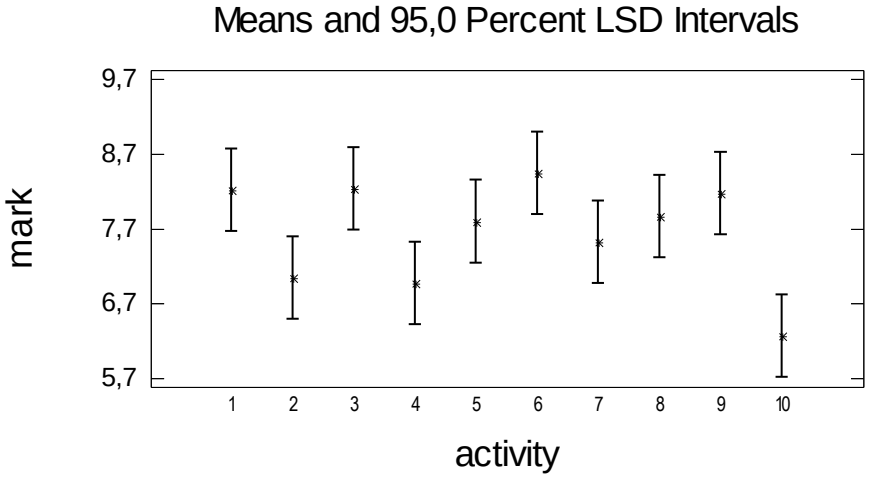

Fig. 7. Comparison of average marks by activity

4 CONCLUDING REMARKS

Teaching statistics at the university level is constantly changing due to the influence of modern technology. The use of statistical software for computations and visual representations, enable students' active knowledge constructions by "doing" and "seeing" statistics. In this paper a case study

of teaching statistical data analysis using software is described. The subject is in the environmental science degree at the Polytechnic University of Valencia (Spain). It is taught in the second semester of the fourth year of the specialization. A computer monitor projected on a screen is used to introduce in the classroom the basic concepts of data analysis.

During the semester there are ten computer laboratory classes in which students work in teams of two people and have to apply data analysis techniques using software. SPSS and Statgraphics are the programs available at the university campus for use in the classes. The students' activities in the computer lab consist in the analysis of environmental data of different fields (air and water pollution, meteorology, etc). The first three activities are related with descriptive analysis (unidimensional and bidimensional frequency tables, univariate descriptive techniques for quantitative data, and bivariate descriptive techniques for quantitative data). The assessment of these sessions has shown that student's have more difficulties and misunderstandings, when interpreting univariate descriptive parameters (location, dispersion and shape) and statistical descriptive plots (histograms, box-whisker plots). In the fourth session in the computer lab students have to apply discrete probability models (Binomial and Poisson), to answer different questions, among which their application to design a sampling procedure is included. The next session is related with continuous probability models (probability plots, normal, lognormal, uniform, exponential and extreme value distributions.

The evaluations results indicate that there are more difficulties in students' application of discrete models than in the use of continuous. Inference analysis is the target of the sessions six to eight. Session six requires the use of hypothesis test and confidence intervals for the mean and standard deviation of a normal population, and hypothesis test on the correlation coefficient. The next one is devoted to the comparisons of means and standard deviations of two normal populations. Session eight is used to introduce the analysis of variance. The two last course computer lab sessions introduce more advanced data analysis methods: the multiple regression model, and descriptive analysis of temporal and multivariate data. The comparison of students' marks in all the evaluations shows that they are significantly smaller in sessions two, four and ten. There are statistical differences between the activities and the students. They are heterogeneous in their backgrounds which is an important factor to influence the teaching-learning process of the course

REFERENCES

[1] Barnett, V. (2004). Environmental Statistics. Chichester, UK: Wiley.

[2] Guttorp, P. (2003). Environmental statistics- a personal view. International Statistical Review, 71(2), pp. 169-179.

[3] Piegorsch, W.W. & Edwards, D. (2002). What Shall We Teach in Environmental Statistics?. Environmental and Ecological Statistics, 9 (2), pp. 125-150.

[4] Moore, D.S. (1997). "New pedagogy and new content: the case of statistics. International Statistical Review, 65(2), pp. 123- 137.

[5] McGinnis, J.R., Wanatabe, T. & McDuffie, A.M. (2005). University mathematics and science faculty modeling their understanding of reform based instruction in a teacher preparation program: voices of faculty and teacher candidates. International Journal of Science and Mathematics Education, 3(3), pp. 407-428.

[6] Biehler, R. (1997). Software for Learning and for Doing Statistics. International Statistical Review, 65(2), pp. 167-189.

[7] Berthouex, P.M. & Brown, L.C. (2002). Statistics for Environmental Engineers, Second Edition. Boca Raton, FL: CRC Press.

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

[9] Piegorsch , W.W. & Bailer, A.J.(1997). Statistics for Environmental Biology and Toxicology. New York: Chapman&Hall, p. 527.

[10] Roberts, L. (1983). Is acid deposition killing west German forests?. Bioscience, 83(5), pp. 302-305.