

VISUALIZATION FOR TEACHING ALL STEPS OF DATA BASED SCIENTIFIC RESEARCH

Clovis A. Peres

Departamento de Medicina Preventiva

Universidade Federal de São Paulo – São Paulo - Brasil

I - INTRODUCTION

The world has entered the era of globalization, but it's not yet safe to talk about statistical education at an international level. Countries have such diverse statistical cultures that improvements in statistical education cannot be overcome by a single strategy. This fact was pointed out by Vere-Jones in his presentation at the ICOTS 5, in Singapore in 1998. Even though statistical education problems are common to different countries, the solutions are almost always developed locally.

At the University of São Paulo, Brazil, the undergraduate and masters courses in statistics have been established from the very beginning by a highly trained faculty staff. After a several years we realized that although students had a good statistical curriculum, they were not able to do well in their jobs as professional statisticians. Their statistical knowledge was compartmentalized and they did not know how to synthesize it in order to solve practical problems. This difficulty was minimized by incorporating into the undergraduate curriculum a two semester course in statistical consulting. This experience was reported in an invited paper, under the title *Building Bridges Between the Academic and the Real World of Statistics*, at the ICOTS I, Sheffield, England, 1982. In 1985, a revised and more detailed paper was published in the *Journal of Educational Statistics* (Peres, 1985).

Discussions at subsequent ICOTS and other scientific meetings have shown that statistical education for statisticians at all levels seems to be very well established in most centers throughout world. There are excellent statistical educators as well as books for statistical topics at all levels.

Nevertheless, statistical education for non-statisticians, mainly from biological and sociological areas, still needs to be improved. Local experience in this area can contribute to this objective.

I worked for 30 years in the Department of Statistics at the University of São Paulo where I taught numerous statistics courses for statisticians and non-statisticians at different levels. I was also in charge of the Statistics Laboratory, thus having had the opportunity to help researchers in different specialties. This was when I realized that there must be a more appropriate way to teach statistics. The student who will become a statistician at least has some skills in mathematics which enables him to understand statistical concepts through mathematical justification. Intuitive arguments are seldom used. In his program he learns different topics, such as, probability, mathematical statistics, stochastic process, design of experiments, sampling techniques, inferential analysis, multivariate analysis, and so on. The connections or relationship between these subjects comes naturally when the student becomes a professional statistician. Students from other areas who will be statistics users, are often adverse to mathematics and lack an understanding of statistical methods through mathematical argument. However, their skills lie in the practice they have in solving real problems which enables them to understand underlying statistical concepts through intuitive argument. Furthermore, students engaged in master or advanced programs are willing to learn statistics for two main reasons. First, in order to be able

to understand scientific papers in which statistical methodology is presented. Second, in order to design and analyze their own scientific research. Statistical educators should capitalize on these necessities to plan a course that will fulfill the students' needs.

In this paper, we present a teaching model in statistics for post-graduate students engaged in masters programs in various medical specialties. The students come from all parts of Brazil as well as from other South American countries and they have little or no statistical background whatsoever. It is a 60-hour (15 weeks) course divided into: 30 hours of lectures and 30 hours of statistical analysis of research projects using STATA. It is designed for twenty students, in order to fit in the computer laboratory. The concepts that underline medical statistics are explained using intuitive appeal and in accordance with the circularity of the scientific method (Kempthorne, 1967). Heavy mathematical justification is avoided. I use to say "It's a course without formulas." However, some formulas are necessary to introduce concepts, but emphasis is given to interpretation. Since the students will be using a computer program to analyze their data, they will only have to interpret the output.

II - STRUCTURE OF THE COURSE

Before explaining the structure of the course, it is important to recall two facts from my professional career, that led me to design a special medical statistics course. When I taught basic statistics for biologists, back in 1968, I always used to begin by telling my students about the circularity of the scientific method, in order to explain the steps involved in scientific research. Unfortunately, after the first class I would then set aside the cycle and give the course in the traditional way.

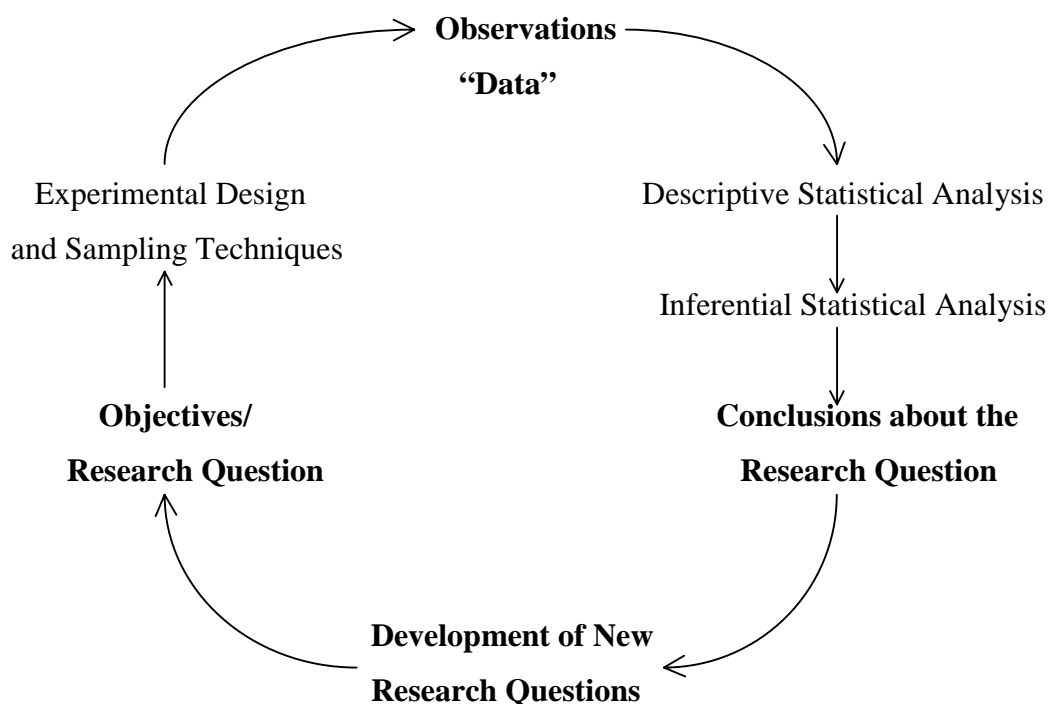
As a PhD student, in Wisconsin in the 70s, I learned from Professors George Box and William Hunter that "Scientific research is a process of guided learning. The object of statistical method is to make that process as efficient as possible," later published in the book *Statistics for Experimenters*, Box, Hunter, Hunter (1978).

Based on these two experiences I came to the conclusion that statistical methods could only contribute to the efficiency of scientific research if used adequately at each step of the scientific method in the order established by its circularity. The situation can be represented diagrammatically by **Figure 1**.

Figure 1 can be thought of as a **map** that allows the student to visualize how the three different areas of statistics interact, in a sequence, from the first to the last step of a research project.

Through the map it is easy to show the students how all the steps are interconnected in such a way that good performance in one will facilitate that in the next. Thus, if the scientific research question is well defined, the appropriate design will be easily chosen. Also, if the appropriate data collection method is specified in the design, the spread sheet and the data set will be well organized. Consequently, if the data set is very well prepared, the appropriate descriptive analysis will be performed and the scientific research question may or not be confirmed. Lastly, if the students have the scientific research question formulated in a statistical way, they will be able to choose the appropriate technique to analyze the data.

Figure 1. *Scientific Research and Statistical Cycle*



To conduct a course following this philosophy, real research projects are selected in such a way, that the statistical analysis, in the last step of the map, covers statistical topics of progressively greater complexity.

Since this course is for non-statisticians, who probably don't have a deep statistical knowledge, it has been divided into three main parts. In Part I the students are instructed in the basic statistical concepts and definitions and in Part II they are given information about the first statistical step of the map. It is only in Part III that the map will actually be used. The course is given as follows:

PART I- DEFINITIONS AND BASIC CONCEPTS

- 1.1. **Definition of Statistics**
- 1.2. **Types of variables (data)**
- 1.3. **Measurements of individual variability:** Standard Deviation, Range
- 1.4. **Frequency Distribution**
- 1.5. **Properties of the Normal Distribution**

PART II- SCIENTIFIC RESEARCH CYCLE

- 2.1. **Scientific Research and Statistical Cycle**
- 2.2. **Sampling techniques**
- 2.3. **Measurement of the variability of one sample to another in terms of summarized measures:** Standard error, Central Limit Theory
- 2.4. **Experimental Design**

PART III - RESEARCH PROJECTS

- 3.1. **Statistical analysis of the relation between two categorical dichotomous variables for non related samples:**
- 3.2. **Statistical analysis of the relation between two categorical dichotomous variables for related samples:** McNemar's Test and Kappa Coefficient
- 3.3. **Statistical analysis of the relation between two categorical nominal and ordinal polychotomous variables:** Pearson's chi-square test.
- 3.4. **Statistical analysis of the relation between one continuous response variable and nominal variables:** analysis of variance for several factors for related and non-related samples.
- 3.5. **Statistical analysis of the relation between one continuous response variable and several numerical variables:** Multiple regression analysis.
- 3.6. **Statistical analysis of the relation between one categorical dichotomous response variable and several numerical and several categorical variables:** Logistic Regression Analysis.

BIBLIOGRAPHY

ALTMAN, D.G. (1997) **Practical Statistics for Medical Research**. Chapman and Hall

BOX, G.E.P., HUNTER, W.G. and HUNTER, J.S. (1978) **Statistics for Experimenters**. New York: John Wiley and Sons.

KEMPTHORN, O. (1967) **The Design and Analysis of Experiments**. New York: John Wiley and Sons.

PERES, C.A., MORETTIN, P.A. (1982) *Building Bridges Between the Academic and the Real World of Statistics*. Proceedings of the First International Conference on Teaching Statistics, Sheffield, England, 1982.

PERES, C.A., MORETTIN, P.A., NARULA, S.C. (1985) *Educating and Training Undergraduate Applied Statisticians*. Journal of Educational Statistics, 10(3): 283 – 292.

VERE-JONES, D. (1998) *Background Influences on the Development of Statistical Education*, Proceedings of the Fifth International Conference on Teaching of Statistics, Singapore, June 1998.

SUMMARY

We present a model for teaching statistics to post-graduate students enrolled in masters programs in diverse medical specialties. The concepts that underline statistical techniques are explained using intuitive appeal according to the order established by the circularity of the scientific method (Kempthorne, 1967). The students use a computer program to analyze the data and the emphasis is on interpreting the output.

RESUMÉ

Nous présentons un modèle pour l'enseignement des statistiques adressé aux étudiants de Maîtrise des Sciences Médicales. Les concepts qui sont à la base de cet enseignement font appel à l'intuition et à l'ordre établi par la circularité des méthodes scientifiques (Kempthorne, 1967). Pour l'analyse et l'interprétation des données les étudiants utiliseront l'ordinateur.