Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

António Pacheco · Rui Santos
Maria do Rosário Oliveira
Carlos Daniel Paulino *Editors*

New Advances in Statistical Modeling and Applications

Springer

Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

For further volumes:
http://www.springer.com/series/10104

Series Editors

Spanish Society of Statistics and Operations Research (SEIO)

Société Française de Statistique (SFdS)

Società Italiana di Statistica (SIS)

Sociedade Portuguesa de Estatística (SPE)

António Pacheco • Rui Santos •
Maria do Rosário Oliveira •
Carlos Daniel Paulino
Editors

New Advances in Statistical Modeling and Applications

Springer

Editors
António Pacheco
CEMAT and Departamento de Matemática
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal

Rui Santos
CEAUL and School of Technology
and Management
Polytechnic Institute of Leiria
Leiria, Portugal

Maria do Rosário Oliveira
CEMAT and Departamento de Matemática
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal

Carlos Daniel Paulino
CEAUL and Departamento de Matemática
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal

ISSN 2194-7767 ISSN 2194-7775 (electronic)
ISBN 978-3-319-05322-6 ISBN 978-3-319-05323-3 (eBook)
DOI 10.1007/978-3-319-05323-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014938629

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Dear reader, On behalf of the four Scientific Statistical Societies – the SEIO, Sociedad de Estadística e Investigación Operativa (Spanish Statistical Society and Operation Research); SFdS, Société Française de Statistique (French Statistical Society); SIS, Società Italiana di Statistica (Italian Statistical Society); and the SPE, Sociedade Portuguesa de Estatística (Portuguese Statistical Society) – we would like to inform you that this is a new book series of Springer entitled "Studies in Theoretical and Applied Statistics," with two lines of books published in the series: "Advanced Studies" and "Selected Papers of the Statistical Societies."

The first line of books offers constant up-to-date information on the most recent developments and methods in the fields of theoretical statistics, applied statistics, and demography. Books in this series are solicited in constant cooperation between the statistical societies and need to show a high-level authorship formed by a team preferably from different groups so as to integrate different research perspectives.

The second line of books presents a fully peer-reviewed selection of papers on specific relevant topics organized by the editors, also on the occasion of conferences, to show their research directions and developments in important topics, quickly and informally, but with a high level of quality. The explicit aim is to summarize and communicate current knowledge in an accessible way. This line of books will not include conference proceedings and will strive to become a premier communication medium in the scientific statistical community by receiving an Impact Factor, as have other book series such as "Lecture Notes in Mathematics."

The volumes of selected papers from the statistical societies will cover a broad range of theoretical, methodological as well as application-oriented articles, surveys and discussions. A major goal is to show the intensive interplay between various, seemingly unrelated domains and to foster the cooperation between scientists in different fields by offering well-founded and innovative solutions to urgent practice-related problems.

On behalf of the founding statistical societies I wish to thank Springer, Heidelberg and in particular Dr. Martina Bihn for the help and constant cooperation in the organization of this new and innovative book series.

Rome, Italy — Maurizio Vichi

Preface

The material of this volume was inspired by selected papers presented at SPE 2011, the XIX Annual Congress of the Portuguese Statistical Society. The annual congress of SPE is the most important statistics meeting taking place every year in Portugal, constituting a primary forum for dissemination of Statistics in Portugal and a privileged channel for scientific exchange between members of SPE and other statistical societies.

SPE 2011 was organized by Instituto Superior Técnico—Technical University of Lisbon and the School of Technology and Management—Polytechnic Institute of Leiria, by invitation from the Directive Board of the Portuguese Statistical Society (SPE). It took place from September 28 to October 1, at Hotel Miramar Sul, in the beautiful picturesque Portuguese sea town of Nazaré.

SPE 2011 continued paving the success of previous SPE congresses having an attendance in excess of 200 participants and included 140 communications from authors from 11 countries (Argentina, Austria, Brazil, England, Germany, the Netherlands, Portugal, Scotland, Spain, Switzerland, and the USA), aside from a 1-day mini-course on Longitudinal Data Analysis, given by M. Salomé Cabral (University of Lisbon) and M. Helena Gonçalves (University of Algarve).

For the pleasant and stimulating scientific and social environment enjoyed by participants in the event, we must thank in a very special way the members of the Organising Committee (Alexandra Seco, António Pacheco, Helena Ribeiro, M. Rosário de Oliveira, Miguel Felgueiras, and Rui Santos) and the Scientific Committee (António Pacheco, António St. Aubyn, Carlos A. Braumann, Carlos Tenreiro, and M. Ivette Gomes).

Last but not least, we must also thank the following four distinguished invited plenary speakers who have honoured us by contributing the first four papers of the volume: Fernando Rosado (University of Lisbon), Graciela Boente (Universidad de Buenos Aires), João A. Branco (Technical University of Lisbon), and M. Ivette Gomes (University of Lisbon).

The publication of this volume, which is the last stone in the SPE 2011 building, aims to disseminate some of the most important contributions presented at SPE 2011 to the international scientific community. The papers included in the volume mix in a nice way new developments in the theory and applications of Probability

and Statistics. There is a total of 27 papers which, for the convenience of readers, were arranged into the following four parts:

- Statistical Science
- Probability and Stochastic Processes
- Extremes
- Statistical Applications

The editors would like to thank all authors who submitted papers to the volume, the anonymous referees for their insightful criticism and excellent reviewing work that contributed to improve the scientific quality and presentation of the accepted papers, and the current Directive Board of SPE, especially Vice President Pedro Oliveira, for assistance during the preparation of this volume. The included papers were accepted for publication after a careful international review process that involved a minimum of 2 referees per paper and a total of more than 70 referees from 10 countries (Argentina, Belgium, France, Germany, Italy, Portugal, Russia, Spain, Switzerland, and the USA).

The editors are very pleased that their work comes to an end at the International Year of Statistics (Statistics 2013), which is a moment of worldwide celebration and recognition of the contributions of statistical science to the humanity. In addition, for them as well as for all participants in SPE 2011, it is a fond remembrance the fact that SPE 2011 paid tribute to the following former presidents of SPE who had retired in the previous year:

- M. Ivette Gomes (1990–1994)
- João A. Branco (1994–2000)
- Fernando Rosado (2000–2006)

SPE wanted, with the tribute, to thank these former popular presidents of SPE for their invaluable work for the progress of the Portuguese Statistical Society and its national and international recognition, as well as for the development of statistics in Portugal, and pay homage also to their strong personal qualities. In this respect, we and SPE would like to provide our most sincere thanks to Isabel Fraga Alves, Manuela Souto de Miranda, and M. Manuela Neves for having promptly and very kindly accepted to be first instance spokespersons in the tribute sessions of M. Ivette Gomes, João A. Branco, and Fernando Rosado, respectively. It was also very moving to the editors and the Organising Committee of SPE 2011 the fact that this event took place close to the end of the mandate as president of SPE of

- Carlos A. Braumann (2007–2011)

whose support, as well as that of the Directive Board of SPE, was invaluable and decisive for the great success of SPE 2011.

Lisbon, Portugal — António Pacheco, M. Rosário de Oliveira, Carlos Daniel Paulino

Leiria, Portugal — Rui Santos

July 2013

Contents

Contributors

Anabela Afonso Department of Mathematics and Research Center of Mathematics and Applications (CIMA-UE), University of Évora, Évora, Portugal

Airlane P. Alencar IME, University of São Paulo, São Paulo, Brazil

Valeska Andreozzi Centro de Estatística e Aplicações da Universidade de Lisboa, FCUL, Lisboa, Portugal

Nelson Antunes FCT of University of Algarve and CEMAT, Faro, Portugal

Paulo Araújo Santos Departamento de Informática e Métodos Quantitativos, Escola Superior de Gestão e Tecnologia, Instituto Politécnico de Santarém, Santarém, Portugal

Juan Lucas Bali Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

Graciela Boente Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

João A. Branco Department of Mathematics and CEMAT, Instituto Superior Técnico, TULisbon, Portugal

Carlos A. Braumann Department of Mathematics, Centro de Investigação em Matemática e Aplicações, Universidade de Évora, Évora, Portugal

Maria de Fátima Brilhante CEAUL and Universidade dos Açores, DM, Ponta Delgada, Portugal

Frederico Caeiro Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa and CMA, Caparica, Portugal

Clara Carlos Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal, Lavradio, Portugal

Luísa Carvalho University of Évora, Évora, Portugal

Clara Cordeiro University of Algarve and CEAUL, Faro, Portugal

José Carlos Dias BRU-UNIDE and ISCTE-IUL Business School, Lisboa, Portugal

Sara Simões Dias Departamento Universitário de Saúde Pública, FCM-UNL, Lisboa, Portugal

Luís Ferreira dos Santos Serviço de Cardiologia, Tondela-Viseu Hospital Center, Viseu, Portugal

Marta Ferreira University of Minho, DMA/CMAT, Braga, Portugal

Patrícia A. Filipe Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Universidade de Évora, Évora, Portugal

Maria Isabel Fraga Alves Faculdade de Ciências, Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, Lisboa, Portugal

M. Ivette Gomes Universidade de Lisboa, CEAUL and DEIO, FCUL, Lisboa, Portugal

Délia Gouveia CEAUL, CIMO/IPB and University of Madeira, Funchal, Portugal

Carla Henriques CMUC and Escola Sup. Tecnologia e Gestão, Inst. Polit. de Viseu, Viseu , Portugal

Lígia Henriques-Rodrigues CEAUL and Instituto Politécnico de Tomar, Tomar, Portugal

Paulo Infante Department of Mathematics and Research Center of Mathematics and Applications (CIMA-UE), University of Évora, Évora, Portugal

Gonçalo Jacinto CIMA-UE and ECT/DMAT of University of Évora, Évora, Portugal

Sandra Lagarto Colégio Luís Verney, CIMA-University of Évora, Évora, Portugal

Manuela Larguinho Department of Mathematics, ISCAC, Bencanta, Coimbra, Portugal

Luiz Guerreiro Lopes CIMO/IPB, ICAAM/UE and University of Madeira, Funchal, Portugal

Susete Marques Instituto Superior de Agronomia (UTL) and CEF, Tapada da Ajuda, Lisboa, Portugal

João Paulo Martins School of Technology and Management, Polytechnic Institute of Leiria, CEAUL-Center of Statistics and Applications of University of Lisbon, Lisbon, Portugal

Maria Oliveira Martins Unidade de Parasitologia e Microbiologia Médicas, IHMT-UNL, Lisboa, Portugal

Ana Cristina Matos Escola Sup. Tecnologia e Gestão, Inst. Polit. de Viseu, Viseu, Portugal

Sandra Mendonça CEAUL and University of Madeira, Funchal, Portugal

Manuel Cabral Morais CEMAT and Mathematics Department, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal

Isabel Natário Faculdade de Ciências e Tecnologia (UNL) and CEAUL, Quinta da Torre, Caparica, Portugal

M. Manuela Neves ISA, Technical University of Lisbon and CEAUL, Tapada da Ajuda, Lisboa, Portugal

Rui Nunes Research Centre for Spatial and Organizational Dynamics (CIEO), University of Algarve, Faro, Portugal

M. Manuela Oliveira Universidade de Évora and CIMA, Évora, Portugal

Patrícia Oom do Valle Research Centre for Spatial and Organizational Dynamics (CIEO), University of Algarve, Faro, Portugal

António Pacheco CEMAT and Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

Claudia Pereira MMEAD/ECT of University of Évora, Évora, Portugal

Dinis Pestana Faculdade de Ciências (DEIO), Universidade de Lisboa and CEAUL, Lisboa, Portugal

Alexandra Pinto Faculty of Medicine of Lisbon, Laboratory of Biomathematics, Lisboa, Portugal

Efigénio Rebelo Research Centre for Spatial and Organizational Dynamics (CIEO), University of Algarve, Faro, Portugal

M. Luísa Rocha Universidade dos Açores (DEG) and CEEAplA, Ponta Delgada, Portugal

Carlos J. Roquete Instituto de Ciências Agrárias e Ambientais Mediterrânicas, Universidade de Évora, Évora, Portugal

Fernando Rosado CEAUL-Center of Statistics and Applications, University of Lisbon, Lisbon, Portugal

Thelma Sáfadi DEX, Federal University of Lavras, Lavras, Brazil

Tiago Salvador Instituto Superior Técnico, Technical University of Lisbon, Lisboa, Portugal

Rui Santos CEAUL and School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal

Ricardo Sousa Higher School of Health Technology of Lisbon, Polytechnic Institute of Lisbon, CEAUL-Center of Statistics and Applications of University of Lisbon, Lisbon, Portugal

Part I

Statistical Science

The Non-mathematical Side of Statistics

João A. Branco

Abstract

It is well recognized and accepted that mathematics is vital to the development and the application of statistical ideas. However, statistical reasoning and proper statistical work are grounded on types of knowledge other than mathematics. To help to understand the nature of statistics and what its goals are some major aspects that make statistics different from mathematics are recalled. Then non-mathematical features are considered and it is observed how these are diverse and really indispensable to the functioning of statistics. Illustrations of various non-mathematical facets are brought about after digging into statistical analyses attempting to end the Mendel–Fisher controversy on Mendel's data from breeding experiments with garden peas. Any serious statistical study has to take into account the mathematical and the non-mathematical sources of knowledge, the two sides that form the pillars of statistics. A biased attention to one side or the other not only impoverishes the study but also brings negative consequences to other aspects of the statistical activity, such as the teaching of statistics.

1 Statistics and Mathematics

Although there is a general consensus among statisticians that mathematics is essential to the development and practice of statistics there is also disagreement and confusion about the amount and the level of sophistication of mathematics used in connection to statistical work. The role of mathematics has been viewed differently throughout the times within the statistical community.

When statistics was at its beginnings, the need for some mathematics was felt, surely because a theoretical basis for statistics was missing. The precise nature of

J.A. Branco (✉)
Department of Mathematics and CEMAT, Instituto Superior Técnico, TULisbon, Portugal
e-mail: jbranco@math.ist.utl.pt

A. Pacheco et al. (eds.), *New Advances in Statistical Modeling and Applications*,
Studies in Theoretical and Applied Statistics, DOI 10.1007/978-3-319-05323-3_1,

the mathematical statistics that arose in the early twentieth century was naturally associated with little data available and the lack of computing power. In William Newmarch Presidential address to the Statistical Society of London on "Progress and Present Conditions of Statistical Inquiry" [22] we can appreciate this concern. Newmarch examines, together with seventeen other fields of statistical interest, the topic "Investigations of the mathematics and logic of Statistical Evidence" saying that it

> . . . relates to the mathematics and logic of Statistics, and therefore, as many will think, to the most fundamental enquire with which we can be occupied . . . This abstract portion of the enquiries we cultivate is still, however, in the first stages of growth. (p. 373)

Ronald Fisher's celebrated book *Statistical Methods for Research Workers* [5] begins with a lapidary first sentence that had tremendous impact on the future development of statistics:

> The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. (p. 1)

This view may be quite natural knowing that Fisher was involved in deep mathematical thinking to establish the foundations of statistics [4]. This potential definition of statistics could have had the same importance as any other, but coming from such an outstanding scientist it had decisive influence in valuing, possibly too highly, the role of mathematics and of mathematicians in the progress of statistics. Too many mathematical abstractions invade the realm of statistics. Mathematical Statistics was born and grew so strongly that it was identified, in some quarters, with Statistics itself. Even today statistical courses and statistical research continue to take place under the umbrella organization of departments of mathematics and the teaching of statistics at school is conducted mainly by teachers of mathematics.

John Tukey was one of the first statisticians to perceive that this line of thought was leaving aside crucial aspects of the subject matter of statistics. He opens his revolutionary paper on "The Future of Data Analysis" [27], by showing his dissatisfaction with the inferential methodology as well as the historical development of mathematical statistics and announcing a new era for statistics:

> For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. . . . All in all, I have come to feel that my central interest is in data analysis, . . . (p. 1)

Tukey's insight of the nature of statistics appears even more profound if we take into consideration the fact that he was a former pure mathematician and that his paper was published in the *Annals of Mathematical Statistics*, a true sanctuary of mathematical statistics. His ideas took time to be assimilated by the community but they raise immediate discussions about the purpose of statistics and the role of mathematics in the development of statistics. One wonders if Tukey's paper had any influence in the decision by the Institute of Mathematical Statistics to split the *Annals of Mathematical Statistics*, just a few years later in 1972, into two different journals, erasing the words "Mathematical Statistics" from the title list of its journals. Many distinguished statisticians have made contributions to these

discussions (see, for example, [2, 3, 15, 17, 18, 29]). In 1998 the statistical journal *The Statistician* published four papers giving a critical appraisal on statistics and mathematics [1, 11, 25, 26] commented by 21 qualified discussants. The debate is very illuminating and reveals a general consensus among the four authors and the discussants, possibly extensible to a great majority of statisticians, that mathematics is a necessary tool to be used as much as statistics needs it, but no more than that in what concerns statistical practice. More recently another author [20] gives a retrospective of the influential role of Tukey's 1962 paper, connected with the issue of statistics and mathematics.

The idea of freeing statistics from the rigidity and limitations of classical inference and going in the direction of data analysis, as advanced by Tukey, was followed by others, as in the area of robust statistics. In robust statistics [10, 14] a unique true model is not accepted. Instead robust statistics considers a family of approximate models, a supermodel, and it tries to find estimates that are as good as possible for all models of the family. The search for procedures that are optimal under a given model gives way to the search for procedures that are stable under a supermodel.

The advent of new technologies has opened the door to the production of a multitude of huge data sets of all kinds and in all fields of activity. It is evident that classical statistical methods often relying on assumptions of independence, normality, homogeneity and linearity, are not ready for a direct analysis of unstructured and non-homogenous data sets with a very large number of observations and variables and where the number of dimensions sometimes exceeds the number of observations. In these circumstances what can be done to analyse this kind of data? According to the spirit emanating from Tukey's concept of data analysis one should do everything sensible, using all means, arguments and tools, including all standard statistical methods and computing facilities, to get to the point, to answer as better as possible the question that we think only the data can help to answer properly. There was a time when statisticians were very interested in the study of methods to analyse small data sets and the discovery of asymptotic behaviours was a temptation that many were willing to try. Today we look around and see floods of data coming from every field such as astronomy, meteorology, industry, economy, finance, genomics and many others. Much of the immense work needed to analyse this sea of data is being done by professionals other than statisticians, working in areas traditionally not considered within the statistical arena (data mining, neural networks, machine learning, data visualization, pattern recognition and image analysis) but having some overlap with statistics. Since most of the tools of data analysis are of statistical nature it comes as a surprise when we see that statisticians are somehow reluctant to get involved with the analysis of large data sets [16]. That position can have negative consequences for the future of statistics. At a time when new statistical methods are needed, to face the complexity of modern data, statisticians should be committed to develop necessary theoretical studies to guarantee the progress of statistics. But to search for the convenient methods that statistics is needing, statisticians should first understand what are the problems and difficulties one encounters when analysing large data sets, and that can only be achieved with a steady involvement in the analysis of this type of data.

In this quick journey on a long road we have seen statistics and mathematics always together but apart, showing different phases of engagement because they are actually distinct undertakings. Next, in Sect. 2, differences between statistics and mathematics are highlighted. In Sect. 3, we discuss an example to illustrate that statistics appreciates its mathematics companionship but needs other companions to arrive at its purposes. Final remarks are presented in Sect. 4.

2 Statistics Is Not Mathematics

No one is interested in discussing whether economics or physics is not mathematics, but statisticians are usually attracted and concerned with the subject of the title of this section. The reason may be found in the recurrent historical misunderstanding between statistics and mathematics, as perceived by the contents of the previous section and of the references found there. To distinguish statistics from mathematics one could start from their definitions but we would find that there is not a unique definition for any of the subjects. To avoid a long discussion and some philosophical considerations we consider only a few characteristics typical of statistics that will serve to highlight the differences of the two subjects. These characteristics have been referred to by many statisticians at large, in particular by those interested in the teaching of statistics, and are:

1. *Origin*: Mathematics and Statistics are both very old. One might say that mathematics was born when primitive men first started to count objects and living things. The origin of statistics is associated with the moment when men first felt the need to register the results of counting, with the interest to remember the past and try to foresee the future. However, statistics, as an academic discipline, is much younger than mathematics, only a little over a century old, while one can speak of hundreds and hundreds of years for the various branches that form the present undergraduate students' mathematical curriculum. Statistics grew outside mathematics prompted primarily by questions and problems arising in all aspects of human activity and all branches of science. The first statisticians were truly experimental scientists [7]. Experimental scientists needing to analyse complex data had—as they have now and will always have—an important catalyst role in broadening the field of statistics and the development of new statistical methods.
2. *Variability*: In a world without variability, or variation, there would be no statistics. But variability abounds in our world and the uncertainty it generates is everywhere. The role of statistics is to understand variability by caring about identifying sources, causes, and types of variation, and by measuring, reducing and modelling variation with the purpose of control, prediction or simple explanation. Statistical variation does not matter much to mathematics, a relevant fact that has to be used to distinguish the two disciplines.
3. *No unique solutions*: Statistics results depend on many factors: data under analysis, model choice and model assumptions, approach to statistical inference, method employed and the personal views of the statistician who does the analysis. Instead of well-identified solutions as is usual in mathematics, various

solutions of a nondeterministic nature, leading to different interpretations and decisions is a common scenario in statistics.

4. *Inductive reasoning*: Two types of reasoning that work in opposite directions can be found in Mathematics and Statistics. Deductive reasoning is used in mathematics: from a general principle known to be true we deduce that a special case is true. Statistical inference uses inductive reasoning: the evidence we find in a representative sample from a population is generalized to the whole population.
5. *Scientific method*: In [19], Mackay and Oldford view the statistical method as a series of five stages: Problem, Plan, Data, Analysis and Conclusion (PPDAC). By comparison with the scientific method for the empirical sciences they conclude that the statistical method is not the same as the scientific method. But, although statistics is a unique way of thinking, it follows the general principles of scientific method. It is embedded in almost every kind of scientific investigations adding rigor to the scientific process and scientific results. To the contrary, mathematics, as generally accepted, does not follow the scientific method.
6. *Context*: Data needs statistics to be analysed and statistics needs data to work. With no data one does not need any statistics. But data are numbers in a context, and that is why context is essential in statistical work. Context is crucial even before we have data because knowing context one can decide how data may be collected to better conduct the statistical analysis. The conclusions of a statistical study have to recall the context to answer properly the questions formulated in the beginning of the study. The case of mathematics is different. The work of mathematics is mainly abstract, it deals with numbers without a context. While context is the soil that makes statistics grow well it may be a drawback that disrupts the natural development of mathematics. That is why context may be sometimes undesirable for mathematicians.

Other aspects typical of statistics but not of mathematics, and certainly not the only ones, are the terminology and the language, the measurement and design issues associated with the collection of proper statistical data, the interpretation of statistical results and the communication of statistical ideas and statistical results to a large and diverse audience.

The idea to set apart statistics from mathematics is not intended to say that mathematics is not important to statistics but to justify that statistical knowledge and statistical reasoning, specific as they are, must be envisioned and cared about as a unique scientific process that must be let to develop freely without any constraints from other fields, in particular from mathematics with which it has a strong connection.

Any inattention to this is likely to distort the natural progress of statistics. One area where this may happen is the teaching of statistics. If teachers and scholars fail to explain clearly the true nature of statistics and the specificity of statistical thinking, their students, detached from the reality of statistics, will tend to propagate a wrong message. And this state of affairs is not uncommon if we think that, on the one hand, the statistics taught at school level is often part of the mathematics curriculum and the teachers who are trained to teach mathematics have, in general, little contact with statistics and no experience whatsoever with the

practice of statistics. On the other hand, at the university level, introductory courses of statistics face the limitations of time allocated to these courses and with little time the syllabus concentrates on formal methods putting sometimes more weight on mathematical aspects than should be the case. Besides, statistics is a difficult subject to teach: students don't feel comfortable with uncertainty and probability, and how do we teach, in the first instance, the ideas of variability or data analysis? How can lecturers, in a limited amount of time, make their students understand that: (1) to have a good knowledge of the context is important, (2) good interpretation of statistical results requires ability, (3) making final decisions about a problem has to be based upon conclusions of statistical analyses, generally not unique and (4) they must exert good communication skills to dialogue with those who have posed the problem, know well the context and expect to follow the statistical arguments and results? Some aspects can only be learned by getting involved with problems of the real world, that is, by doing statistics.

Interest in the teaching of statistics is not new [13, 28] but it grew tremendously when it was felt that citizens living in a modern society should be statistically literate and statistics was then introduced in the school mathematics curriculum. Many obvious questions, that are not easy to answer, were then put forward: Who is going to teach statistics? Who can? What to teach? How to do it? and so on. The International Statistical Institute realizing the scale of the problem and its interest to the community created IASE (International Association of Statistical Education) to promote statistical education. IASE organizes conferences and other actions concerning the teaching and the learning of statistics. Today statistical education is a topic of research that attracts a large number of people who publish the results of their investigations in specialized journals. Ideas of changing curricula, styles and methods of teaching and learning are in the air [9, 12]. Although school elementary courses and university introductory courses are very distinct and run in completely different scenarios there are reasons to believe that the difficulties encountered in passing the statistical message in both cases share some form of influence of two general conditions: mathematical and non-mathematical aspects of statistics and the relative importance that is given to each one of them.

Next, we look at a statistical article [23] trying to identify and discuss various non-mathematical aspects of the analysis. Any other non-theoretical work could be used to illustrate the role of the non-mathematical aspects but this particular one has the advantage that the author of the present work is a joint author of that paper and then he can review and quote from it more freely.

3 The Non-mathematical Side of Statistics

The title of the paper mentioned at the end of the previous section, "A statistical model to explain the Mendel–Fisher controversy", is self-explanatory in what the authors want to do. The question is how they arrive at that model and what they are doing with it. Let us review the various phases of this work.

3.1 The Problem

Gregory Mendel, known as the founder of scientific genetics, published, as early as 1866, his two laws of heredity (the principle of segregation and the principle of independence) [21]. This amazing discovery was arrived at after continuous meticulous work, during more than 7 years, on controlled experiments by cross-breeding garden pea plants. Mendel cultivated and tested around 29,000 plants. Inspired by good judgement and based on empirical calculations (proper statistical methods did not exist at the time) on the registered data of the artificial fertilization Mendel worked out the laws of hereditary. But despite being an extraordinarily revolutionary achievement it was forgotten until 1900, during 35 years, when it was rediscovered by independent researchers.

Ronald Fisher, known as the founder of modern statistics [24], and a great geneticist, soon got interested in Mendel's work. In 1911 he made a first analysis of Mendel's results, and having found that they were exceptionally good questioned the authenticity of the data presented by Mendel. Twenty five years later, in 1936, Fisher came back to review the problem and performed a thorough and rigorous analysis of the same data and of all the Mendel experiments supposed to generate that data. He reinforced his previous opinion concluding that the data are simply "too good to be true" [6], what became a truly demolishing assessment for Mendel's image. Apparently Fisher's veiled accusation of forgery was ignored until the centennial celebration of Mendel's 1866 publication when it suddenly came to light and a stream of controversial opinions, about the relevant question, started to flood the publication spaces with tens of papers, including the recent book "Ending the Mendel–Fisher Controversy" [8] which really does not manage to accomplish what its title promises. Pires and Branco [23] present a short chronological account of the major facts of this controversy almost century-old controversy. They refer to the vast bibliography that has been produced, some of which is very illuminating for the sake of understanding the problem and the discussion of the analyses proposed by the various contributors.

The relevant question is: is Fisher right? That is, has Mendel's data been faked? Since Mendel's laws are right, we must start by asking if Fisher's analysis is correct, because if it is not then the reason for the accusation would be lost. A second question is: if Fisher is right can we think of other possible reasons why Mendel's data conforms so well to his model, instead of immediately accusing him of scientific misconduct?

3.2 Data in Context

To answer the first question Fisher's analysis must be reviewed. As mentioned in [23] only the part of Fisher' paper related to a chi-square analysis is considered here. It is in fact the extremely high p-value obtained by Fisher in that analysis that

served mostly to support his attack on Mendel and that has also been the bone of contention among the scientists involved in the debate.

To understand the data as prepared by Fisher in order to apply the chi-square goodness-of-fit test it is necessary to get into the genetic background, to follow the details of a large number of complex and delicate experiments, to be aware of the subtle problems of measurement of the experimental results and finally to understand Mendel's theory. Mendel's paper is simple, clear and certainly the best to help the reader in these matters, but there are other useful sources. Pires and Branco [23] give an organized summary of relevant aspects of the experimentation and give comments on the data that help to understand why and how the chi-square can be used. Knowing the context and understanding the data is essential also to follow the arguments advanced by many researchers to defend their proposals to solve the controversy.

Mendel concentrated on the transmission of seven binary traits or characteristics of garden pea plants (two traits observed in the seeds and five observed in the plants). One trait has two forms (phenotypes), A (named dominant) and a (named recessive), just like seed shape (round, A, or wrinkled, a) and flower colour (purple, A or white, a). He tried various types of cross fertilization and observing the traits of the offsprings and comparing the results with his expectations he consolidated his theory. Following [23] and a classification used by Fisher, the experiments can be classified into single trait experiments, bifactorial experiments and trifactorial experiments according to the number of traits considered in each crossing, one, two or three. Fisher included in his analysis more complex experiments classified into two new categories: gametic ratios and illustrations of plant variation experiments. In accordance with Mendel's theory, crossing a number of plants pure lines (those whose offsprings are always similar to their parents) and then crossing the resulting offsprings (no pure lines any more, called hybrids) then the offsprings of this last crossing will be of the two original phenotypes A and a in the proportion of 3:1. That is, in a population of n offsprings the number of phenotypes A (success), n_A, will be distributed as a binomial distribution, $n_A \sim Bin(n, p)$, where p is the probability of success ($p = 3/4$ in this case of the ratio 3:1), under the standard hypotheses: each observation is considered a Bernoulli and trials are independent. A more thorough and complete description of this interpretation, extended to all cases of cross breading included in the study, is in [23]. To test Mendel's theory we consider the number of successes, $n_A \sim Bin(n, p)$, and the hypothesis H_0: $p = p_0$, where p_0 is the true probability of success under Mendel's theory. The observed value of the test statistic to test H_0 against H_1: $p \neq p0$ is given by $\chi = (n_1 - np_0)/\sqrt{np_0(1-p_0)}$. Assuming n is large the p-value of the test is $P\left(\chi_1^2 > \chi^2\right)$.

3.3 Fisher's Chi-Square Analysis

Having assumed the binomial model (in some cases a multinomial model was assumed) and independence of experiments Fisher tested H_0 applying a chi-square goodness of fit test and then he summed up all the chi-square statistics and degrees