UNIVERSITAT POMPEU FABRA

idEC INSTITUT D'EDUCACIÓ CONTÍNUA

Causal Inference, Missing Data and Measurement Error

Short courses » 10th Applied Statistics Week

www.idec.upf.edu/asw

Barcelona, 28 June - 1 July 2004



Course 1. Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies - **Donald B. Rubin**

Course 2. Missing Data in Large-scale Surveys - **Nicholas T. Longford**

Course 3. Causal Relationships with Measurement Error in the Data - **Willem E. Saris**

Aims

The 10th **APPLIED STATISTICS WEEK** is being organized by the Universitat Pompeu Fabra (UPF) from 28 June to 1 July 2004, in Barcelona.

The **APPLIED STATISTICS WEEK** aims to provide a set of intensive short courses on a particular statistical theme of an applied nature. The courses are presented by acknowledged leading researchers of international stature, who are also known to have excellent teaching skills.

Past themes of the **APPLIED STATISTICS WEEK** have been: "Statistics in the Health Sciences" (1995), "Statistics in Classification and Pattern Recognition" (1996), "Design and Analysis of Survey Data" (1997), "Statistics in Environmental Science" (1998), "Statistics in Marketing Research" (1999), "Statistics in Society" (2000), "Statistics in Finance" (2001), "Statistical Genetics" (2002) and "Statistical Modeling for Behavioral and Genetic Studies" (2003). This year's Applied Statistics Week, the 10th, continues this tradition dealing with new topics that are of crucial importance in applied research. The goal of many research studies is to uncover causal relationships among the observed variables, because causation represents the ultimate understanding of the studied phenomena. The design of such studies, and the quality of the data they yield, usually do not satisfy the paradigms of experimentation - randomisation, completeness of the data and absence of measurement error. Such deficiency presents a formidable challenge to applied statistics.

This year's Applied Statistics Week comprises three courses that deal with the state of the art in the theory and practice related to these problematic and in some contexts contentious issues.

The first course by Professor Donald B. Rubin (Harvard University) will set up the language and the basic statistical tools for dealing with a concept that so frequently has been evaded by the statistics profession: causality. In recent years, however, there has been an intense debate on this essential concept that has gone beyond the boundaries of statistics (this debate has taken place in Economics, Social Sciences, Epidemiology, Computer Sciences, etc.). In this discussion, Donald Rubin has been the bright star that has set the statistical standards on this fundamental issue. Donald Rubin's course is designed for a broad range of applied researchers, such as applied statisticians, economists, social scientists, medical researchers, epidemiologists, computer scientists, philosophers of science, and the like.

Missing data is a ubiquitous problem in large-scale surveys, so any analysis that does not address it is problematic. Most established approaches either reduce the collected data by discarding the records that do not fit the straitjacket of the planned analysis or impute a value for each missing item, so that the data completed in this manner would fit into this straitjacket. These approaches are deficient because they either discard some collected data or pretend that more data is available for the analysis than has been collected by the survey. Nicholas T. Longford is a prominent researcher in several areas of applied statistics, including the treatment of missing data. He has recently presented a course on this topic at the Royal Statistical Society as part of his Campion Fellowship award for, among his other achievements, developing modern methods for handling missing data in large-scale surveys. Nick Longford features in his course a special lecture by the most innovative researcher in this field, Professor Donald B. Rubin.

Measurement error impregnates all the data that is collected in applied statistics. The effect of measurement error can be very severe when assessing causal relations among variables. Models that account for measurement error have been under development in social sciences research, especially with the introduction of LISREL by Karl G. Jöreskog in the early seventies. Structural equation models have since then been established as the basic toolkit to deal with measurement error and latent variables in social sciences. Since the late seventies, Professor Willem E. Saris has been at the forefront in using structural equation models in the practice of social science research, with special attention to problems of measurement. The course by Willem E. Saris will address the issue of how measurement error can distort standard analysis of causality, and will show how these problems can be taken care of by structural equations. Examples of application in different fields will be provided. A special lecture by Albert Satorra, on practical aspects of estimation and testing in SEM, will be given.

We feel that UPF is uniquely placed to foster such intensive courses of high quality in Spain and, indeed, in Europe as a whole. The University and the city of Barcelona enjoy an optimal location for gathering scholars from different parts of Spain and Europe, a city known for its strong work and innovation ethic as well as its weath of cultural and leisure activities. The courses are concentrated into one week to facilitate the enrollment of working professionals and the academic community. We also promote lively interactions between participants and instructors that assist in achieving the goal of improving the application of statistical concepts and methods to problems in our society.

Supporting institutions

The ten editions of the Applied Statistics Week have been organized in cooperation with the Institut d'Estadística de Catalunya (IDESCAT) and, since the year 2003, sponsored by the Fundación BBVA.

Programme director

Albert Satorra. Department of Economics and Business, Universitat Pompeu Fabra, Barcelona.

Target audience

Researchers in all the fields involved with statistical applications with the problems of causality, missing data and errors in variables. This comprises applied researchers in economics, finance, marketing and business, social science, medicine, epidemiology, computer science, and others.

Place

Classes will be held at IDEC (Continuing Education Institute, Universitat Pompeu Fabra) – Balmes, 132, 08008 Barcelona.

Language

All the courses will be taught in English.

Limited number of seats in each course

To make possible a wide interaction between participants and instructors, coffee breaks and the one-and-a-half hour lunch break is scheduled as an integral part of the course. The size of the class for each course is limited to 32. Applications will be processed as a first come first served basis.

Course 1. Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies

Donald B. Rubin. Harvard University.

This course will present the Rubin Causal Model perspective on understanding and teaching statistical inference for causal effects through potential outcomes. There are three parts to the course. The first part establishes the foundation, which is based on the concepts of units, treatments, and potential outcomes, as well as replication, the stability assumption, and the assignment mechanism. The second part presents inference based solely on the assignment mechanism; this perspective generalizes Fisher's (1925) and Neyman's (1923) randomization-based methods for finding significance levels, unbiased estimates, and confidence intervals, and emphasizes the central role of the propensity score (Rosenbaum and Rubin, 1983). The third part presents inference based on predictive models for the distribution of the missing potential outcomes, formally, Bayesian posterior predictive inference, due to Rubin (1978); the material presented, however, is not in terms of formal models but far more intuitive multiple imputation methods for the missing potential outcomes. In practice, the predictive approach is ideal for creating statistical procedures, whereas the assignment-based approach of Fisher is ideal for traditional confirmatory inference, and the assignment-based approach of Neyman is ideal for evaluating procedures. For best practice, being facile with all three approaches is important. There is essentially no prerequisite knowledge for this course, as the material is based on an introductory course taught at Harvard University and designed for students with very little quantitative background. Examples are presented from a variety of fields, including medicine, education, and economics.

An example of causal inference occurs in the situation where a sample is split into two groups. The first group is given a medicine (treatment W=1), and the other group is the control group (W=0). Assume that the variable Y denotes the score on a medical test that is observed in both groups. The aim of causal inference is to estimate the effect of treatment (W=1) versus control (W=0) on the outcome Y. This effect is defined as the comparison of (i) the response in the population if everyone were exposed to the treatment, W=1 (that is, the collection of values of Y under W=1, Y(1)) versus (ii) the response in the population if everyone were exposed to control, W=0 (that is, the collection of values of Y under W=0, Y(0)).

Even if this example does not connect to your current research, the seminar should still be interesting because it is also an introduction to statistics where causality is explicitly the object of inference. Many statisticians end their statistical inference with statements only about association, even when it can not be denied that there is some implicit causality implied. The seminar shows how to generate explicit causal statements when conducting statistical inference.

Program

Day I:

Part I: Framework for causal inference - Rubin's Causal Model.

Part II: Methods of causal inference based on the assignment mechanism.

Day II:

Part III: Causal inference based on predicting missing potential outcomes.

Part IV: Complications including noncompliance, censoring due to death, and other principal stratification examples.

The instructor

Donald B. Rubin is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University, where he has served as Chairman for many years. He has over 300 publications (including several books) on a variety of topics, including computational methods, causal inference in experiments and observational studies, survey methods, techniques for handling missing data, Bayesian methods, multiple imputation, matched sampling, and applications in many areas of social and biomedical science. Professor Rubin is a Fellow of the American Statistical Association, the Institute for Mathematical Statistics, the International Statistical Institute, the Woodrow Wilson Society, the John Simon Guggenheim Society, the New York Academy of Sciences, the American Association for the Advancement of Sciences, and the American Academy of Arts and Sciences. He is also the recipient of three of the most prestigious awards available to statisticians: The Samuel S. Wilks Medal of the American Statistical Association, the Parzen Prize for Statistical Innovation and the R.A. Fisher Lectureship of the Committee of Presidents of Statistical Societies.

Date: Monday, 28 June 2004, from 9.30 to 13.30, and from 15.00 to 18.00 hours, and Tuesday, 29 June 2004, from 9.30 to 13.30, and from 15.00 to 18.00 hours.

Course 2. Missing Data in Large-scale Surveys

Nicholas T. Longford. De Montfort University.

The course will review traditional methods for dealing with missing values in large-scale surveys, highlight their deficiencies and describe two methods for efficient estimation with incomplete data - the EM algorithm and multiple imputation. The EM algorithm is suitable for simpler problems. The course will focus on the method of multiple imputation, starting with the standard setting of (secondary) analysts wishing to make a range of inferences using software tools that cannot deal with incomplete data. The terminology associated with the process that brings about the missing values will be discussed and the importance of collecting information about it and modelling it emphasised. The theory underlying the

method will not be dealt with in detail, but its consequences will be elaborated. Multiple imputation will be motivated by several examples, including some from outside the narrow confines of missing data, such as measurement error and misclassification.

The course will focus on the method of multiple imputation and its applications in the analysis of large-scale surveys, especially in the context of government statistics. The method has a computationally intensive component, but after its conclusion a range of analyses can be conducted with the same software (programs or algorithms) that would be suitable if there were no missing data. The method is universal - applicable to any such analysis, irrespective of its complexity.

The principal references for the method are Rubin, 1987 (Multiple Imputation for the Non-response in Surveys, Wiley and Sons) and Rubin, 1996 (Multiple imputation after 18+ years, JASA 91, 473-528). Another important reference is Little and Rubin, 1987 (Statistical analysis with missing data, Wiley and Sons(2nd edition 2002)) and Dempster, Laird, and Rubin, 1977 (Maximum likelihood from incomplete data via EM algorithm, JRSS B 39, 1-38). The presenter's main references are Longford et al, 2000 (Handling missing data in diaries of alcohol consumption, JRSS A 163, 381-402) and Longford, 2004 (Missing data and small-area estimation in the UK Labour Force Survey, JRSS A 167, in print).

Target audience

The course is addressed to statistical analysts working with large-scale surveys. Understanding of ordinary (linear) regression and some of its extensions (generalised linear models or multilevel models) and proficiency in some statistical software (Splus, SAS, SPSS, or similar) will be assumed. Participants/delegates are encouraged to submit problems for the discussion (surgery), preferably in advance, to ntlagi@hotmail.com.

Program

1. Introduction. Motivating examples. Background.
   How to talk about missing data (terminology).
   Imputation and deletion methods. Deterministic and stochastic methods.
2. The replication-related standard error. Valid inference.
   The processes of sampling and missingness.
   Information external to the data.
   EM algorithm and incomplete information.
3. Multiple imputation - the setting and the basics. Proper imputation and the practice.
4. Examples.
5. Donald B. Rubin. Making the best of what you've got.
6. Discussion. Surgery.

The instructor

Nicholas T. Longford is a Senior Research Fellow at De Montfort University in Leicester, England. He was awarded the first Campion Fellowship in 2000 for a proposal to introduce and develop modern methods for handling missing data in large-scale surveys and for applying novel methods in small-area estimation in UK government statistics (monograph entitled New Analytical Equipment for the Survey Statistician, Springer-Verlag, to appear in 2004). His other professional interests include multilevel analysis (monograph entitled Random Coefficient Models, Oxford University Press, 1993), dealing with model uncertainty, incorporating costs and taking account of the consequences in statistical decisions, including in model selection. In 2003, he was an Academic Visitor at Massey University, New Zealand and in 2002 a Visiting Professor at the University of Khartoum in the Sudan. Since 2001, he has been a Consultant in the Methodology Group of the Office for National Statistics. He is an Associate Editor of the Journal of the Royal Statistical Society, Series A, and the British Journal of Mathematical and Statistical Psychology, and a former Editor of the Journal of Multivariate Analysis.

Date: Wednesday, 30 June 2004, from 9.30 to 13.30, and from 15.00 to 18.00 hours.

Course 3. Causal Relationships with Measurement Error in the Data

Willem E. Saris. University of Amsterdam.

Although the effect of measurement error on the estimators of relations has been known for a long time, and the procedures for correcting for measurement error are also well known (Jöreskog, 1969), this is often ignored in practice. A reason for this is that the models to correct for measurement error looked too complex for most applied researchers.

The course will shows that measurement errors in experimental and observational studies can be accounted for by Structural Equation Modelling (SEM), and that these procedures are efficient and simpler than is generally thought. It will be shown that measurement error erodes essential properties of standard estimators, and so correction for it is essential (Saris and Andrews, 1991). We will distinguish between latent and observed variables, the latter being subject to measurement error.

Correct model specification ensures that estimators have the claimed properties. Model checking is therefore a major issue (Saris, Satorra and Sörbom 1987). Procedures for model evaluation, goodness-of-fit testing and power analysis will be given special attention. These procedures will be illustrated by examples from psychology, medical, economics and social science research.

Target audience

The course is targeted at all researchers who are involved in empirical research and have basic knowledge of correlation and regression analysis.

Program

Part I:

1. two approaches to estimating causal models
2. the effect of systematic and random measurement errors
3. detecting measurement errors
4. estimating structural models with latent variables: ML, GLS and robust estimation

Part II:

1. testing structural models: Likelihood ratio test, goodness of fit and power considerations.
2. prediction of quality of questions (SQP).
3. correction for measurement error: examples.

The instructor

Willem E. Saris is Professor in Methodology of Social Science Research at the University of Amsterdam, and Visiting Professor at ESADE, the Business School of the Universitat Ramon Llull. He is member of the Coordination Team of the European Social Survey, held in 22 countries and sponsored by the EU Commission and the European Science Foundation. His special expertise is detection and correction for measurement error in social science research. He has been working on measurement problems in social science research during his whole academic career. He has published many methodological papers showing how structural equation models can be used to correct for measurement error in the data, and an elementary introduction to structural equation models (Saris and Stronkhorst, 1984). He has also written books on the improvement of measurement by use of continuous scales, detection and correction for individual response functions, and computer assisted data collection (Saris, 1991). Currently he is working with his research team on assessing the effect of the different designs of survey questions on the quality of the data collected. On this issue he has developed a program (SQP) that assesses the quality of questions before they are used in practice. He has shown that in these issues of measurement error, structural equation models plays a central role.

Date: Thursday, 1 July 2004, from 9.30 to 13.30, and from 15.00 to 18.00 hours.

Course fees

The course fees include course notes, a Diploma of Continuing Education Institute, meals and refreshments during the courses.

- Course 1: 320 € if paid before the 28th of May 2004, otherwise 470 €.
- Course 2: 210 € if paid before the 28th of May 2004, otherwise 290 €.
- Course 3: 210 € if paid before the 28th of May 2004, otherwise 290 €.

For those who register for 2 courses after the 28th of May 2004, the fee is reduced by 10%.
For those who register for 3 courses after the 28th of May 2004, the fee is reduced by 20%.

SPECIAL FEES for doctoral students and participants in a previous APPLIED STATISTICS WEEK:

- Course 1: 230 €.
- Course 2: 140 €.
- Course 3: 140 €.

Members of the IDEC's Alumni Association will enjoy a 20% discount.

Non accumulative discounts.

Registration

The registration form must be sent by the 28th of May to take advantage of the reduced fees. Payment can be made by bank transfer (including a copy of the bank transfer), by credit card or by bank cheque to the Institut d'Educació Contínua. In the case of foreign payments, the cheque should be in convertible euros.

Information and enrolment

IDEC (Continuing Education Institute)
Balmes, 132 · 08008 Barcelona - Spain
Telephone: (+34) 93 542 18 50 · Fax: (+34) 93 542 18 08
http://www.idec.upf.edu · e-mail: info@idec.upf.edu

Further information about the Applied Statistics Week can be found at:
http://www.idec.upf.edu/asw

APPLIED STATISTICS WEEK

Name:

Surname:

Home Address:

Telephone:

Fax:

ID Number/Passport (compulsory):

E-mail:

Organization:

Position:

Work Address:

Telephone:

Fax:

E-mail:

REGISTRATION (Mark your choice with an X)

[ ] Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies

[ ] Missing Data in Large-scale Surveys

[ ] Causal Relationships with Measurement Error in the Data

PAYMENT (Mark your choice with an X)

[ ] Bank cheque to Institut d'Educació Contínua

[ ] Bank transfer to

0182-5734-33-0010003294 Banco Bilbao Vizcaya Argentaria

(a copy of the bank transfer should be sent by fax)

[ ] Credit card payment

Visa card number:

Expiry date:

Cardholder's printed name:

Date:

Signature:

Send to: IDEC · Balmes 132 · 08008 Barcelona, Spain · Fax Number: (+34) 93 542 18 08 · e-mail: info@idec.upf.edu

Sponsored by:

Fundación BBVA

In cooperation with:

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Information
www.idec.upf.edu

Balmes, 132
08008 Barcelona
Telephone (+34) 93 542 18 50
Fax (+34) 93 542 18 08
info@idec.upf.edu

