International Conference

APPLIED STATISTICS 2010

PROGRAM and ABSTRACTS

September 19 - 22, 2010

Ribno (Bled), Slovenia

International Conference

APPLIED STATISTICS 2010

PROGRAM and ABSTRACTS

September 19 – 22, 2010

Ribno (Bled), Slovenia

Organized by

Statistical Society of Slovenia

Supported by

Slovenian Research Agency (ARSS)

Statistical Office of the Republic of Slovenia

ALARIX

RESULT d.o.o.

VALICON / SPSS Slovenia

ELEARN Web Services Ltd

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

311(082.034.2)

INTERNATIONAL Conference Applied Statistics (2010; Ribno)
Program and abstracts [Elektronski vir] / International
Conference Applied Statistics 2010, September 19–22, 2010, Ribno
(Bled), Slovenia ; organized by Statistical Society of Slovenia :
edited by Lara Lusa and Janez Stare. El. knjiga. - Ljubljana :
Statistical Society of Slovenia, 2010

Način dostopa (URL): http://conferences.nib.si/AS2010/AS2010-Abstra
cts.pdf

ISBN 978-961-92487-5-1

1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo
Slovenije
252476416

Scientific Program Committee

Janez Stare (Chair), Slovenia
Vladimir Batagelj, Slovenia
Maurizio Brizzi, Italy
Anuška Ferligoj, Slovenia
Dario Gregori, Italy
Dagmar Krebs, Germany
Lara Lusa, Slovenia
Mihael Perman, Slovenia
Jože Rovan, Slovenia
Willem E. Saris, The Netherlands
Vasja Vehovar, Slovenia
Tomaž Banovec, Slovenia
Jaak Billiet, Belgium
Brendan Bunting, Northern Ireland
Herwig Friedl, Austria
Katarina Košmelj, Slovenia
Irena Križman, Slovenia
Stanisław Mejza, Poland
John O'Quigley, France
Tamas Rudas, Hungary
Albert Satorra, Spain
Hans Waege, Belgium

Organizing Committee

Andrej Blejec (Chair)
Bogdan Grmek
Lara Lusa
Irena Vipavc Brvar

Published by: Statistical Society of Slovenia
Vožarski pot 12
1000 Ljubljana, Slovenia
Edited by: Lara Lusa and Janez Stare
Printed by: Statistical Office of the Republic of Slovenia, Ljubljana

PROGRAM

		Hall 1	Hall 2
Sunday	10.30 – 11.00	Registration	
	11.00 – 11.10	Opening of the Conference	
	11.10 – 12.00	Invited Lecture	
	12.00 – 12.20	Break	
	12.20 – 13.40	Social Science Methodology	Data Mining
	13.40 – 15.00	Lunch	
	15.00 – 16.20	Design of Experiments	Education
	19.00	Reception	
Monday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Biostatistics and Bioinformatics I	Statistical Applications I
	11.40 – 12.00	Break	
	12.00 – 13.20	Biostatistics and Bioinformatics II	Network Analysis
	13.20 – 14.30	Lunch	
	14.30	Excursion	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Modeling and Simulation I	Sampling Techniques and Data Collection
	11.40 – 12.00	Break	
	12.00 – 13.20	Econometrics	Mathematical Statistics I
	13.20 – 15.00	Lunch	
	15.00 – 16.40	Modeling and Simulation II	Mathematical Statistics II
Wednesday	9.10 – 10.30		Statistical Applications II
	10.30 – 10.50	Break	
	10.50 – 11.50	Modeling and Simulation III	
	12.00 – 12.20	Closing of the conference	
	12.30 – 14.00	Lunch	
	14.00 – 18.00	Workshop	

10.30–11.00 **Registration**

11.00–11.10 **Opening of the Conference**

11.10–12.00 **Invited Lecture** (Hall 1) *Chair: Vasja Vehovar*

1. **Direct and Indirect Causal Effects: A Helpful Distinction?**
   Donald Rubin

12.00–12.20 **Break**

12.20–13.40 **Social Science Methodology** (Hall 1) *Chair: Donald Rubin*

1. **Integration of Item Response Theory and Structural Equation Modeling with Dichotomous Data. An Application of Measurement Invariance in Political Preferences**
   Lluis Coromina

2. **Functional Independence Measure from the Classical Test Theory Perspective with an Eye on the Item Response Theory Approach**
   Gaj Vidmar and Helena Burger

3. **Comparing Student Grading Distributions**
   Matevž Bren and Darko Zupanc

4. **Integrated Modelling of European Migration**
   James Raymer, Jonathan J. Forster, Peter W.S. Smith, Jakub Bijak, Arkadiusz Wiśniowski and Guy J. Abel

12.20–13.40 **Data Mining** (Hall 2) *Chair: Aleš Žiberna*

1. **Comparison of Rough Sets and Discriminant Analysis in the Study of Intangible Assets**
   Agnes Maciocha

2. **Yellow Pages Analysis by Using Web Mining**
   Umman Tugba Simsek Gursoy and Serap Sahin

3. **Association Rules Analysis for E-Commerce Data: An Application in Turkey**
   Serap Sahin and Umman Tugba Simsek Gursoy

13.40–15.00 **Lunch**

15.00–16.20 **Design of Experiments** (Hall 1) *Chair: Janez Stare*

1. **A Construction Method of Incomplete Split-Block Designs Supplemented by Control Treatments**
   Shinji Kuriki, Stanisław Mejza and Iwona Mejza

15.00–16.00 **Education** (Hall 2) *Chair: Andrej Blejec*

1. **Trends of Inequal Gender Proportions in Most Demanding Educational Tracks**
   Gašper Cankar

19.00 **Reception**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Janez Stare*

1. **Models, Assumptions and Confidence Limits**
   John Copas

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics I** (Hall 1) *Chair: John Copas*

1. **Evaluation of Multiple Allele Probabilities at a Single Locus for Ungenotyped Members of Complex Pedigrees**
   Gregor Gorjanc

2. **Marginal Structural Models: Treatment in Reach Study**
   Nursel Koyuncu and Emmanuel Lesaffre

3. **Anisotropic and Inhomogeneous Hidden Markov Models for the Analysis of Water Quality Spatio-Temporal Data on a Cylindrical Lattice**
   Luigi Spezia

10.20–11.40 **Statistical Applications I** (Hall 2) *Chair: Germa Coenders*

1. **Determining the Risk of Contracting Cancer Using Logistic Regression**
   Berna Yazıcı, Betül Kan, Gonca Mert and Aynur Küçükçongar

2. **Inflation and Market Power in Agriculture: A Case Study in the Banana Industry in Israel**
   Tal Shahor

11.40–12.00 **Break**

12.00–13.20 **Biostatistics and Bioinformatics II** (Hall 1) *Chair: Gaj Vidmar*

1. **Threshold Calibration for High-Dimensional Classification with Class Imbalanced Data**
   Lara Lusa and Rok Blagus

2. **A Nonparametric Generalized Additive Model for a Dichotomous Dependent Variable on Forest Fire Data**
   Ahmet Sezer, Betül Kan and Berna Yazıcı

12.00–13.00 **Network Analysis** (Hall 2) *Chair: Nataša Kejžar*

1. **Using Social Network to Predict Behavior of Active Members of Web Forums**
   Aleš Žiberna and Vasja Vehovar

2. **Single and Multiple Name Generators for Measuring Social Support Networks**
   Maja Mrzel and Valentina Hlebec

3. **On Fide Network(s) of Chess Players**
   Kristijan Breznik and Vladimir Batagelj

4. **Structures of Collaboration in Slovenian Science System**
   Luka Kronegger, Patrick Doreian, Anuška Ferligoj and Franc Mali

13.20–14.30 **Lunch**

14.30 **Excursion**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Lara Lusa*

1. **Three-sided Hypothesis Testing: Simultaneous Testing of Superiority, Equivalence and Inferiority**
   Jelle Goeman

10.00–10.20 **Break**

10.20–11.40 **Modeling and Simulation I** (Hall 1) *Chair: Jelle Goeman*

1. **Hunting for Significance with the False Discovery Rate**
   Sonja Zehetmayer, Martin Posch and Peter Bauer

2. **A Permutation Test for Random Effect Models**
   Dario Basso, Livio Finos and Gianmarco Altoè

3. **Rank Tests in Measurement Errors Models**
   Radim Navratil

4. **Permutation Test for Partial Regression Coefficient on First-Order Autocorrelation**
   Pradthana Minsan and Pachitjanut Siripanich

10.20–11.40 **Sampling Techniques and Data Collection** (Hall 2) *Chair: Lluis Coromina*

1. **Algorithms for the Stratification of Skewed Populations**
   Jane M. Horgan and Sebnem Er

2. **Adaptive Cluster Sampling Based on Ranked Sets**
   Girish Chandra and Hukum Chandra

3. **Bayesian Bootstrap Approximation to Sampling Distributions**
   Naoto Niki and Yoko Ono

4. **Statistical Analyses of the Experimental Results on the Bond Behaviour of Reinforcing Steel and Concrete**
   M. Sami Donduren, M. Tolga Cogurcu, Mustafa Altin and Mehmet Kamanli

11.40–12.00 **Break**

12.00–13.20 **Econometrics** (Hall 1) *Chair: Aleša L. Dolinar*

1. **There is Only One Statistical SoftwaRe**
   Giovanni Millo

2. **Non-linearity, Complexity and Limited Measurement in the Relationship Between Domain and Overall Life Satisfaction**
   Monica Gonzalez, Germa Coenders, Marc Saez and Ferran Casas

3. **Plurality of Methods for the Categorical Variables: Empirics of Microfinance Impacts on Happiness in Thailand and Brazil**
   Thanawit Bunsit

12.00–13.20 **Mathematical Statistics I** (Hall 2) *Chair: Mihael Perman*

1. **Asymptotic Behavior of the Kernel Regression Estimator Under Different Rates of Censoring and Dependence**
   Zohra Guessoum and Elias Ould Said

2. **Minimal Sufficiency in Rare Populations**
   Mohammad Moradi, Jennifer Brown and Miriam Hodge

3. **Asymptotic Behavior of a Smooth Conditional Quantile Kernel Estimator for Censored and Dependent Data**
   Ourida Sadki and Elias Ould Said

13.20–15.00 **Lunch**

15.00–16.40 **Modeling and Simulation II** (Hall 1) *Chair: Germa Coenders*

1. **Loglinear Models for Contingency Table**
   Justyna J. Brzezińska

2. **A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Contingency Tables Via Simulation**
   Atilla Göktaş and Öznur İşçi

3. **A Comparison of the Most Commonly Used Measures of Association for Doubly Non-Squared Ordered Contingency Tables Via Simulation**
   Öznur İşçi and Atilla Göktaş

4. **On Robust Ridge Regression and Robust Liu Estimator**
   Betül Kan, Berna Yazıcı and Özlem Alpu

15.00–16.20 **Mathematical Statistics II** (Hall 2) *Chair: Damjan Škulj*

1. **A Class of Asymptotically Normal Degenerate Quasi U-Statistics**
   Aluisio Pinheiro, Pranab Kumar Sen and Hildete Prisco Pinheiro

2. **Ordering Life Distributions Through Interval Entropy Function**
   Fakhroddin Misagh, Gholam Hossein Yari and Rahman Farnoosh

9.10–10.30 **Statistical Applications II** (Hall 2) *Chair: Simona Korenjak-Černe*

1. **Hierarchical Clustering of Population Pyramids Presented as Histogram Symbolic Data**
   Nataša Kejžar, Simona Korenjak-Černe and Vladimir Batagelj

2. **The Survey Study about the Lifestyle of Young Generations in Thailand**
   Vacharaporn Suriya-bhivadh

3. **Numerical Realization Nonlinear Regressions Dependences**
   Victor Lyumkis

4. **Applications of Heavy Tailed Distributions for Modeling Human Behavior and Activities in Cyber Space**
   Mohammad Ali Baradaran Ghahfarokhi, Parvin Baradaran Ghahfarokhi and Rahim Bahramian Dehkordi

10.30–10.50 **Break**

10.50–11.50 **Modeling and Simulation III** (Hall 1) *Chair: Giovanni Millo*

1. **Control Charts For Weibull, Gamma and Lognormal Distributions**
   Derya Çalışkan and Canan Hamurkaroglu

2. **Statistical Modulation of a Human Health Problem in Albania**
   Luela Prifti, Etleva Beliu and Shpetim Shehu

3. **Non-linear Dimensionality Reduction for Functional Computer Code Modeling**
   Benjamin Auder

4. **A Bayesian Approach to Inferring the Contribution of Unobserved Ground Conditions to Observed Scores in Sports: The Example of Cricket**
   Scott R. Brooker and Seamus Hogan

12.00–12.20 **Closing of the conference** (Hall 1)

12.30–14.00 **Lunch**

14.00–18.00 **Workshop** (Hall 1)

1. **Practical Applications of Permutation Tests**
   Phillip Good

ABSTRACTS

Invited Lecture

Direct and Indirect Causal Effects: A Helpful Distinction?

Donald Rubin

Harvard University, Cambridge, United States of America; rubin@stat.harvard.edu
http://videolectures.net/as2010_rubin_dic/

Although the terminology of direct and indirect causal effects is relatively common, I believe that it is generally scientifically unhelpful without further explication. This assessment is based on repeated experience with the confusion it creates in important examples in the social and biomedical sciences. After reviewing my perspective on causal inference based on the concepts of potential outcomes and assignment mechanisms, this presentation will discuss two distinct ways to formalize the issues that arise in circumstances where this terminology is used.
The first is based on the concept of principal stratification, and the second is based on the concept of a compound assignment mechanism, which is just a special case of general assignment mechanisms. Simple artificial examples will be used to illustrate the differences between the two conceptualizations, and to show that it is easy to become confused when using the direct and indirect jargon to describe causal effects.
Even the great R.A. Fisher was a victim of this confusion, as will be documented.

Social Science Methodology

Integration of Item Response Theory and Structural Equation Modeling with Dichotomous Data. An Application of Measurement Invariance in Political Preferences

Lluis Coromina

University of Girona, Girona, Spain; lluis.coromina@udg.edu

Item Response Theory (IRT) models focus on binary or ordinal manifest variables, these models are probabilistic with nonlinear relations between hypothetical constructs and their observed variables. IRT focus on location and discrimination of each item. It is mainly used in education and psychology since late 1960s.
Structural Equation Modeling (SEM) has its measurement component called Confirmatory Factor Analysis (CFA). It specifies linear relationships on factor loading of each observed variable on the latent construct of interest. IRT and CFA have similarities in its measurement part when dichotomous or ordinal data is used. Due to its equivalence in the measurement part, these models can be used together in a complete model. The paper explains the similarities and differences between both methods and it shows how they can be combined when dichotomous data is used.
Besides the integration of IRT and SEM in a model, the paper also explores the measurement invariance for such integrated model. It is done using multiple group analysis; in our case we study invariance according gender, education level and age group variables. The political application uses dichotomous data from Spanish citizens on the preference of the level of supranational decision making for eight policies (immigration, defence, environment, fighting against organized crime, interest rates, aid to developing countries, agriculture and welfare).

Functional Independence Measure from the Classical Test Theory Perspective with an Eye on the Item Response Theory Approach

Gaj Vidmar[1] *and Helena Burger*[2]

University Rehabiltation Institute, Republic of Slovenia, Ljubljana, Slovenia
[1]gaj.vidmar@ir-rs.si
[2]helena.burger@ir-rs.si

Functional Independence Measure (FIM) is arguably the main outcome measure in rehabilitation medicine and an important casemix tool. It was devised in 1984 in the USA to be used as universal assessment tool in the Uniform Data System for medical rehabilitation. Its use has since been reported in over one thousand published articles in patients with various pathologies and for various purposes. At the University Rehabilitation Institute in Ljubljana, which is the only rehabilitation hospital in the country and thus provides comprehensive rehabilitation for the whole territory of Slovenia (admitting over 1300 cases per year), compulsory FIM assessment at admission and discharge has been performed (and integrated into the hospital information system) since 2004.
FIM consists of the motor and the cognitive subscale with 13 and 5 items, respectively, all rated on a 7-point scale. After numerous publications proved its acceptable reliability and validity within the classical test theory perspective, it has been more recently almost equally widely criticised from the item response theory perspective, especially from the Rasch model perspective.
Based on the data from the 2004-2006 period, we previously reported on comprehensive FIM analyses at our Institute (including patient demographics, duration of rehabilitation, other classifications, admission and discharge diagnosis, complications and treatment cessations, and type of admission, treatment and discharge) with emphasis on the general aspects of data visualisation for decision support and healthcare quality monitoring, as well as on modelling FIM patient progress. This paper incorporates data from the 2007-2009 period and focuses more deeply on psychometric issues. Internal validity is tested (in terms of unidimensionality, using Parallel Analysis and the Minimum Average Partial Test, and internal consistency) and the relationship between independence level of single items and the corresponding total subscale scores is assessed using ordinal logistic regression modelling.

Comparing Student Grading Distributions

Matevž Bren[1] *and Darko Zupanc*[2]

[1]University of Maribor, Ljubljana, Slovenia; matevz.bren@fvv.uni-mb.si
[2]National Examination Center, Ljubljana, Slovenia; darko.zupanc@guest.arnes.si

In comparing student achievements in one subject or in one class/school with the other/s there are two research questions to be considered. First is: do the two grade distributions differ and the second: to what extend is this true.
In our contribution we will present ordinal statistics to answer these ordinal questions. The first question can be answered by the standard MWW test and the second with the application of ordinal dominance graph (ODG) and directed dissimilarity $d = P(X > Y) - P(X < Y)$. Properties of the directed dissimilarity d will also be discussed: antisymmetric, vanishes on the diagonal, $|d|$ is dissimilarity.
In demonstration we will apply real data on student grading in upper secondary education in Slovenia and we will set criteria whether difference between grading distributions is small, medium or large.

Integrated Modelling of European Migration

James Raymer[1], *Jonathan J. Forster*[2], *Peter W.S. Smith*[3], *Jakub Bijak*[4], *Arkadiusz Wiśniowski*[5] *and Guy J. Abel*[6]

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom
[1]raymer@soton.ac.uk
[2]J.J.Forster@soton.ac.uk
[3]pws@soton.ac.uk
[4]J.Bijak@soton.ac.uk
[5]A.Wisniowski@soton.ac.uk
[6]G.J.Abel@soton.ac.uk

In order to fully understand the causes and consequences of international population movements in Europe, researchers and policy makers need to overcome the limitations of the various data sources, including inconsistencies in availability, definitions, data-collection techniques and quality. In this presentation, we propose a Bayesian model for harmonising and correcting the inadequacies in the available data and for estimating the completely missing flows for 31 EU and EFTA countries and providing meaningful measures of uncertainty of the estimates. We have built a hierarchical Poisson-log-normal model that assumes the flows are measured with error, reflecting discrepancies between the data-collection systems of countries in relation to the United Nations recommendation for the measurement of migration flows. These measurement aspect of the model focuses on duration of stay, coverage of subpopulations and accuracy of the data collection system. We also have built in an explanatory model to borrow strength from covariates, patterns over time, and to predict unreported flows. The model is analysed by means of an MCMC technique implemented both in WinBUGS and R. The results represent a synthetic data base of international migration with measures of uncertainty for both the flows and model parameters. Having such a data base allows us to better understand the underlying mechanisms and reasons for recent migration trends.

Data Mining

Comparison of Rough Sets and Discriminant Analysis in the Study of Intangible Assets

Agnes Maciocha

Dublin Institute of Technology, Dublin, Ireland; amaciocha@dit.ie

This paper presents a comparison of a data mining method with a traditional statistical method. Each technique was applied to a sample of both Virtual and Traditional Organisations focusing on their Intangible Assets. According to literature the virtual organization has emerged to create flexibility and efficiency, i.e. better exploitation of resources and development of capabilities within groups of organizations. To that end, it was decided to analyse the differences between the two types of companies from the Intangible Assets perspective. The objective was to search for patterns within the Intangible Assets data that could distinguish between these two types of companies. For data mining, Rough Sets Analysis was used, and for the traditional method discriminant analysis was applied. The purpose of this comparative analysis was to determine which method is more accurate in classifying data of this nature. The literature suggests that Rough Sets are more accurate in predicting group categorization. The results will be discussed and the strengths and weaknesses of each method will be investigated.

Yellow Pages Analysis by Using Web Mining

Umman Tugba Simsek Gursoy[1] *and Serap Sahin*[2]

Department of Quantitative Methods, Istanbul University, Istanbul, Turkey
[1]tugbasim@istanbul.edu.tr
[2]srp_shn@yahoo.com

In recent years, in business, marketing, medical and automotive sectors, rapidly growing data causes problems, such as data collection and data storage. The need for new techniques has occurred because of increased data, to find solutions to the problems. Data mining have been used for this purpose.
By using Data mining techniques, end-users can perform queries, generate reports and apply analysis. These techniques are developed for end-users in order to take quick and efficient decisions.
In this project, "Yellow Pages" data, are analyzed by using Web Mining that has become a popular subject in recent years. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services.
The most popular analysis on the Web site is to identify the most visited pages. With this project, "Yellow Pages" data is going to be analyzed by using web mining to identify which pages the customers visit frequently.

Association Rules Analysis for E-Commerce Data: An Application in Turkey

Serap Sahin[1] *and Umman Tugba Simsek Gursoy*[2]

Department of Quantitative Methods, Istanbul University, Istanbul, Turkey
[1]srp_shn@yahoo.com
[2]tugbasim@istanbul.edu.tr

Association Rules Analysis is one of the important techniques of data mining. Association rule mining finds interesting associations and correlation relationships among large set of data items. This powerful exploratory technique has a wide range of applications in many areas of business practice and also research, such as analysis of consumer preferences or human resource management.
The data are valuable resources for e-commerce companies which operate many functions in the Internet environment such as product promotion, taking orders, sales and marketing, customer relationship management.
The biggest computer retailing firms' data are analyzed to identify which items are bought together by the customers. Association Rules Analysis is the rule of analyzing which items frequently occur together in the same transaction. Clementine program, which is the data mining model of SPSS, is used to analyze data.

Design of Experiments

A Construction Method of Incomplete Split-Block Designs Supplemented by Control Treatments

Shinji Kuriki[1], *Stanisław Mejza*[2] *and Iwona Mejza*[3]

[1]Department of Mathematical Sciences, Osaka Prefecture University, Sakai, Japan; kuriki@ms.osakafu-u.ac.jp
[2]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; smejza@up.poznan.pl
[3]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; imejza@up.poznan.pl

We give a construction method of incomplete split-block designs supplemented by control treatments, using a modified Kronecker product of two square lattice designs for test row and column treatments. We consider a mixed linear model for the observations with a three step randomization, i.e., the randomization of blocks, the randomization of the rows (or columns) within each block and the randomization of the columns (or rows) within each block. We characterize such incomplete split-block designs with respect to the general balance property and we give the stratum efficiency factors for the incomplete split-block designs.

Education

Trends of Inequal Gender Proportions in Most Demanding Educational Tracks

Gašper Cankar

National Examinations Centre, Ljubljana, Slovenia; gasper.cankar@guest.arnes.si

Programme for international student assessment (PISA) provides wealth of high quality data for informed decisions in education. Based on results from PISA 2006 author demonstrates trend of increased feminization in most demanding educational tracks of upper secondary education in Slovenia and also in some other Central European countries like Croatia, Hungary, Austria, Czech Republic and Slovak Republic. In case of Slovenia and Croatia the trend is so explicit that science results for boys and girls by educational track in those two countries on PISA 2006 produce Simpson's paradox which will be demonstrated and explained. Since those trends reflect broad characteristics of school systems and societies in each country they are likely to remain constant in years since 2006. Implications of such trends will be disscussed and explored.

Invited Lecture

Models, Assumptions and Confidence Limits

John Copas

University of Warwick, Warwick, United Kingdom; jbc@stats.warwick.ac.uk
http://videolectures.net/as2010_copas_mac/

Confidence intervals reflect our uncertainty about a parameter of interest, and models reflect our assumptions about the context of the data. Some of these assumptions may be justified by background knowledge, but others will be rather arbitrary. Statistics text books advise that before assuming a model we should check that it gives a good fit to the data (by using goodness-of-fit tests or graphical diagnostics). But does a well-fitting model necessarily mean a good confidence interval? Looking at the robustness of confidence limits to model choice suggests some rather basic questions about our use of models and assumptions in statistics.

Biostatistics and Bioinformatics I

Evaluation of Multiple Allele Probabilities at a Single Locus for Ungenotyped Members of Complex Pedigrees

Gregor Gorjanc

University of Ljubljana, Biotechnical Faculty, Animal Science Department, Domžale, Slovenia;
gregor.gorjanc@bf.uni-lj.si

Development of molecular genetics led to the abundant source of genotype information. However, there are often some individuals that are not (yet) genotyped or their genotype is in conflict with the genotype of relatives. In such cases the evaluation of genotype probabilities based on the genotypes of relatives can be performed using the so called peeling methods. However, standard peeling methods can only be used with relatively simple pedigrees. For complex pedigrees the approximate methods can be used. Here such a method is presented that employs a linear mixed model which considers multiple allele loci, genotype and/or pedigree errors, unknown allele frequency in a founding population, and pedigrees of practically unlimited size. An application of this method is shown for the PrP gene in sheep in Slovenia.

Marginal Structural Models: Treatment in Reach Study

Nursel Koyuncu[1] *and Emmanuel Lesaffre*[2]

[1]Hacettepe University, Ankara, Turkey; nkoyuncu@hacettepe.edu.tr
[2]Erasmus University Medical Center, Rotterdam, Netherlands;

An important aim of clinical studies is to how treatment influences the course of disease of a patient. Estimating the casual effect of a time-varying covariate is difficult or impossible. The control of time-varying covariates is a fundamental problem in analyzing data and interpreting results. Standard approaches for adjustment of confounding are biased. Marginal structural models (MSMs) are casual models designed to adjust for time-dependent confounding in observational studies of time-varying treatments.
Reach is a study conducted by Erasmus MC Medical Centre in Netherlands for rheumatoid arthritis (RA), which is a chronic debilitating disease of the joints. Treatment in RA improved strongly over the last 25 years with the occurence of Metotrexate (MTX) and, in the late nineties, of the biologicals. In this study we use the marginal structural models to estimate the effect of treatment on patients suffering from rheumatoid arthritis.

Anisotropic and Inhomogeneous Hidden Markov Models for the Analysis of Water Quality Spatio-Temporal Data on a Cylindrical Lattice

Luigi Spezia

Biomathematics & Statistics Scotland, Aberdeen, United Kingdom; luigi@bioss.ac.uk

Motivated by a real data problem, an anisotropic and inhomogeneous spatio-temporal Hidden Markov model (HMM) with an unknown number of states is made up on a cylindrical lattice. A Bayesian inference procedure, based on a reversible jump Markov chain Monte Carlo algorithm, is proposed to estimate both the dimension and the unknown parameters of the model. The real data problem is the modelling in time and in space of the concentrations of three dissolved inorganic nitrogens recorded monthly by the Scottish Environmental Protection Agency in the 56 major Scottish rivers. The 56 gauging stations can be linked to create a circle and so the spatio-temporal data set can be displayed on a cylinder. The states of the hidden Markov process allows the classification of the observations in a small set of groups. The different states can represent increasing levels of pollution. In the Bayesian model presented here, the hidden Markov process is an anisotropic and inhomogeneous Potts model. The Potts model is widely used in statistical mechanics to model the spins of elementary particles that are placed on a lattice. Here the hidden Potts model is assumed to be anisotropic (i.e., variant under rotations) and inhomogeneous (i.e., variant under translations). Anisotropy is due to the presence of two different parameters describing the link between neighbouring pixels: one for the temporal relation and the other for the spatial relation. Inhomogeneity is established by assuming that the spatial relation is a function of the physical distance between two neighbouring sites.

Statistical Applications I

Determining the Risk of Contracting Cancer Using Logistic Regression

Berna Yazıcı[1], *Betül Kan*[2], *Gonca Mert*[3] *and Aynur Küçükçongar*[4]

[1]Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey; bbaloglu@anadolu.edu.tr
[2]Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey; bkan@anadolu.edu.tr
[3]Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey; goncamert_ist@hotmail.com
[4]Gazi University School Of Medicine Department Of Pediatrics. Department Of Child Health And Diseases, Ankara, Turkey; aynurcon@yahoo.com

A research is conducted in order to determine the risk of contracting cancer of the children. The sample is defined as the children who have been treated in Gazi University Hospital, Department of Children Health and Diseases. The factors that can be effective as a risk of contracting cancer among the children have been listed and regarding that, a questionnaire is formed. The survey is conducted to 616 children, 233 girls and 383 boys. The survey lasted 1.5 years. The dependent variable for binary logistic model was defined as contracting cancer or not. As a result of logistic regression, the significant variables in the model are mother's age, mother's education, breast feeding duration, and if there is a cancer occurrence in the family or not. The logistic regression model for the risk of contracting cancer among the children is constructed, and interpreted. Also the cell probabilities and prediction values are given in a resulting table.

Inflation and Market Power in Agriculture: A Case Study in the Banana Industry in Israel

Tal Shahor

Department of Economics, Max Stern Academic College of Emek Yezreel, Yaad, Israel;
tals@yvc.ac.il

In a competitive market, the marginal cost is equal to production price and the allocation of resources is efficient. On the other hand, in a market where firms have market power, production price is higher than marginal cost and there is inefficient resource allocation. The purpose of this study is to test the hypothesis that inflation influences market power. If inflation increases market power, then the importance of fighting inflation increases. The measure of market power used here is the markup. This is the ratio between the price and the marginal cost. Under competitive conditions and risk neutrality, the markup equals one and increases with the market power of the firms. It is my intention to discuss the above issue within the context of the banana industry in Israel. I chose this industry because the banana industry in Israel is centralized thus facilitating a relatively orderly data collection, and the building up of a panel database for about 50% of banana growers in Israel. The results of this study lead to the conclusion that, in the banana industry, there is no relationship between inflation and market power.

Biostatistics and Bioinformatics II

Threshold Calibration for High-Dimensional Classification with Class Imbalanced Data

Lara Lusa[1] *and Rok Blagus*[2]

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, Ljubljana, Slovenia
[1]lara.lusa@mf.uni-lj.si
[2]rok.blagus@mf.uni-lj.si

Frequently the classifiers are developed using class-imbalanced data, i.e., data sets where the number of subjects in each class is not equal. Standard classification methods used on class-imbalanced data produce classifiers that do not accurately predict the smaller class. We previously showed that additional challenges exist when the data are both class-imbalanced and high dimensional, i.e., when the number of samples is smaller than the number of measured variables.
Most classification methods base the classification rule on a numerical variable that is produced by the classification algorithm, for example on the probability for a sample to belong to a class (as for penalized logistic regression), on the proportion of samples among the nearest neighbors that belong to a class (for k-NN), on the proportion of bootstrap trees that classified the sample in a given class (for random forests). If the value of this numerical variable is above a pre-specified threshold, then the new sample is classified in a given class.
We evaluated if we could improve the performance on imbalanced data of some classifiers by estimating the threshold value upon which their classification rule is based. We addressed the issue on how choose the threshold value. We estimated the threshold (on a training set) maximizing the Youden's index (sensitivity + specificity-1), the positive or the negative predictive value, or their sum. The results obtained on independent test sets were evaluated both in terms of class-specific predictive accuracies and of class-specific predictive values, and we compared the empirically determined thresholds with the thresholds commonly used in practice.
In this talk we will show the simulation-based results obtained using penalized logistic regression models.

A Nonparametric Generalized Additive Model for a Dichotomous Dependent Variable on Forest Fire Data

Ahmet Sezer[1], *Betül Kan*[2] *and Berna Yazıcı*[3]

Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey
[1]a.sezer@anadolu.edu.tr
[2]bkan@anadolu.edu.tr
[3]bbaloglu@anadolu.edu.tr

Parametric regression methods enjoy simplicity, but they suffer from both inflexibility in modeling complicated relationships between the response and explanatory variables and assumptions about the residuals. Splines provide the advantage that they do not require assumptions about the form of the estimation equation.
This paper applies generalized additive models using penalized regression splines to investigate relationship between forest fires and important explanatory variables. It is well known that forest fires are one of the major environmental concerns in all over the world. They not only create economical and ecological damage, but also threaten human life.
This study develops a particular generalized additive model for probabilistic risk assessment, especially to estimate probabilities for occurrence of fires given the explanatory variables. Our intension here is to consider a model which best explains the structure of the dataset in term of GCV.

Network Analysis

Using Social Network to Predict Behavior of Active Members of Web Forums

Aleš Žiberna[1] *and Vasja Vehovar*[2]

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
[1]ales.ziberna@fdv.uni-lj.si
[2]vasja.vehovar@fdv.uni-lj.si

In this paper we use social networks of active members in online communities (i.e. authors) to predict their future behavior. We limit our discussion to web forums, the most generic type of on-line communities. With web forums (or discussion boards) we understand here websites where users discuss various topics by posting messages in existing topics/threads or by opening new ones.
We first address the problem of measuring ties between participants (members) where there is no direct information (i.e. quotation, direct reply indication) about the tie between two members. Without such direct information it is not entirely unambiguous whether a massage responds to a certain massage or is unrelated and is simply posted in the same thread after that massage. Based on the solution to this problem we also construct corresponding social network. However, while such networks are suitable for description of members' behavior, they are not very suitable for the prediction of their future behavior. Here we suggest a very specific method for creating networks of members that is specially designed for prediction. The main characteristic of our approach is that we take into account all available data while putting more weights on the more recent data (messages). In addition to networks of members, we also use the network (two-mode networks) of threads (topics) to generate predictions.
The network data and other information related to the posting habits of the members are used to develop an algorithm that predicts future behavior of the members. Ultimate goal of our research is to generate explicit recommendations for members of online communities about which members and topics/threads to follow and then use these predictions/recommendations as the basis for more user friendly forum interface.

Single and Multiple Name Generators for Measuring Social Support Networks

Maja Mrzel[1] *and Valentina Hlebec*[2]

Faculty of Social Sciences, Ljubljana, Slovenia
[1]maja.mrzel@gmail.com
[2]valentina.hlebec@fdv.uni-lj.si

Survey data collection techniques of egocentric social networks can be divided into single and multiple name generators. In a single name generator we use just one question and in an multiple name generator we use several questions to generate individual's social network. Studies have shown that different egocentric network is measured, if we use single or multiple name generator. (Hlebec and Kogovsek, 2006) In our paper we wish to determine, what is the difference between obtained information of individual's social network gathered with these two techniques. The data used in this study was collected in 2002 (multiple name generator) and 2007 (single name generator). The 2002 sample is a representative sample of adult population in Slovenia (n = 5013) and the 2007 sample is a quota sample defined by gender and three age groups (n = 558). In both techniques the exchange approach was used, where individual names persons he/she collaborates with in various exchanges. This approach is particularly suitable for measuring social support. Given the differences in surveys questionnaire in 2002 and 2007, we compared results of emotional support, instrumental support and socializing, and the size and composition of the whole network of a social support.

On Fide Network(s) of Chess Players

Kristijan Breznik[1] *and Vladimir Batagelj*[2]

[1]Graduate student of Statistics, University of Ljubljana , Ljubljana, Slovenia;
kristijan.breznik@mfdps.si
[2]Faculty of Mathematics and Physics, Ljubljana, Slovenia;
vladimir.batagelj@fmf.uni-lj.si

At the Fide (world chess federation) web site data on the results of games and tournaments are available; from January 2008 on the single game level. From these data some (temporal) networks can be obtained. Additional data about chess players (rating, age, gender, country, title, ...) are also available.
Collecting the data we run in to some problems: there exist different players with the same name; the same player is entered into the Fide base under different names (different writting, typos); some players passed away during the time of analysis and they are no longer in Fide base of players; for some unknown reason Fide also does not publish Elo ratings for players from several countries; etc. We discuss some approaches how to deal with these problems and produce consistent data sets.
Some chess players presume that the best players of the world are almost exclusively playing between themselves, avoiding to play against low rated opponents in order to keep their high Elo chess rating - they mainly play in closed, also called berger, tournaments. Another interesting question is how much the result of the game depends on the color of pieces. It is obviously harder to win a chess game with black pieces, but in the Elo system this is not considered in evaluation of the result. In the paper we deal with these and some other similar questions on the basis of data from Fide base using network analysis.
The programs for collecting the data from the Fide web site and producing networks were written in R. For analysis of networks we used Pajek.

Structures of Collaboration in Slovenian Science System

Luka Kronegger[1], *Patrick Doreian*[2], *Anuška Ferligoj*[3] *and Franc Mali*[4]

[1]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
luka.kronegger@fdv.uni-lj.si
[2]Department of Sociology, University of Pittsburgh, Pittsburgh, United States of America;
[3]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
[4]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;

Co-authorship as form of scientific collaboration presents the major interaction mechanism between actors at the micro-level of individual scientists. Wide range of mechanisms fostering collaboration produce different local patterns within general network which can be described from the perspective of research groups, research topics, intensiveness of collaboration or comparison of entire research disciplines. In our research we observed and compared collaborative structures in complete longitudinal co-authorship networks for four research disciplines. Dataset was split into 4 five-year intervals and clustered with blockmodeling technique. The main question of the research was to what extent the obtained multi-core periphery structure is determined by organizational structure of the institutions, special topics in the scientific field and other factors that foster scientific collaboration as research projects.

Invited Lecture

Three-sided Hypothesis Testing: Simultaneous Testing of Superiority, Equivalence and Inferiority

Jelle Goeman

Leiden University Medical Center, Leiden, Netherlands; j.j.goeman@lumc.nl
http://videolectures.net/as2010_goeman_tsh/

We propose three-sided testing, a testing framework for simultaneous testing of inferiority, equivalence and superiority in clinical trials, controlling for multiple testing using the partitioning principle. Like the usual two-sided testing approach, this approach is completely symmetric in the two treatments compared. Still, because the hypotheses of inferiority and superiority are tested with one-sided tests, the proposed approach has more power than the two-sided approach to infer non-inferiority or non-superiority. Applied to the classical point null hypothesis of equivalence, the three sided testing approach shows that it is sometimes possible to make an inference on the sign of the parameter of interest, even when the null hypothesis itself could not be rejected. Relationships with confidence intervals are explored, and the effectiveness of the three-sided testing approach is demonstrated in a number of recent clinical trials.

Modeling and Simulation I

Hunting for Significance with the False Discovery Rate

Sonja Zehetmayer[1], *Martin Posch*[2] *and Peter Bauer*[3]

Medical University of Vienna, Vienna, Austria
[1]sonja.zehetmayer@meduniwien.ac.at
[2]martin.posch@meduniwien.ac.at
[3]peter.bauer@meduniwien.ac.at

When testing a single hypothesis, it is common knowledge that increasing the sample size after non-significant results and repeating the hypothesis test several times at unadjusted critical levels inflates the overall Type I error rate severely. Surprisingly, if a large number of hypotheses are tested controlling the False Discovery Rate (a frequently used error criterion for large scale multiple testing problems), such 'hunting for significance' has asymptotically no impact on the error rate. More specifically, if the sample size is increased for all hypotheses simultaneously and only the test at the final interim analysis determines which hypotheses are rejected, a data dependent increase of sample size does not affect the False Discovery Rate. This holds asymptotically (for an increasing number of hypotheses) for all scenarios but the global null hypothesis where all hypotheses are true. To control the False Discovery Rate also under the global null hypothesis, we consider stopping rules where stopping before a predefined maximum sample size is reached is possible only if sufficiently many null hypotheses can be rejected.

A Permutation Test for Random Effect Models

Dario Basso[1], *Livio Finos*[2] *and Gianmarco Altoè*[3]

[1]Department of Managemnt and Ingegniring, University of Padua, Padova, Italy;
dario@stat.unipd.it
[2]Department of Statistical Sciences, University of Padua, Padova, Italy;
livio@stat.unipd.it
[3]Department of General Psychology and Department of General Psychology, University of Padua, Padova, Italy; gianmarco.altoe@unipd.it

The mixed effects models have been widely explored and extended in many directions in the last five decades. Most of the efforts, have been aimed to the definition of new models and on methods to estimate their parameters. The estimation methods usually rely on (some form of) ML requiring then strong distributional assumptions of -both - random effects and error terms. The demand of knowledge of the underlying model is just partially compensated by the fact that only heavy departure from the assumptions will cause serious biases to the inference (e.g. hypotheses testing). Even when this is not the case, the power of the inference is potentially compromised. In this work we present a permutation test for mixed effects model. We restrict our discussion to hierarchical models with clustered observations (e.g. subjects with many trials each) and the mean of each cluster being a random process. The method tests the hypothesis of independence of mean of the cluster with categorical classifications (i.e. factors, anova-like setting) and/or with continuous predictors. The distributional assumptions are weakened and only existence of the second moment of the error and random terms is required. We show that exact solutions are available for designs with equal observations on each cluster. We also propose an approximate solution for the general case of unbalanced designs. The quality of the type I error control and the power seems to be promising from simulations and from theoretical argumentations. Moreover the method is very general and its extension to a multivariate response framework is immediate. Some applications to the psychological field are shown and discussed. The R codes are provided.

Rank Tests in Measurement Errors Models

Radim Navratil

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic; navratil@karlin.mff.cuni.cz

In many practical applications, when the values of the random variable of our interest are obtained by measurement, it can happen that we do not get the accurate value of the random variable, but we get the value affected by measurement error. Application of parametric methods in this case is not very convenient, because we do not know the exact distribution of the errors and their estimation can make the situation more complicated. It shows that just the rank tests are suitable for this problem. We will consider the linear regression model and testing of linear hypothesis and as a special case of this model two-sample tests (e.g. comparison of male and female saleries). Further we show using of one-sample tests of symmetry (e.g. testing the hypothesis that the new treatment is better than the current, or that older twin has different properties than younger, or wheather the left eye can see sharper than the right one). At first we remind basic idea of construction the rank tests for this models and mention some properties required for testing the hypotheses. The contribution further states when and under which assumptions we can use the classical rank tests, defined for models without measurement errors, in models with measurement errors and how it affects the power of this tests (the power will decrease). This theoretical results are illustrated on examples or simulation studies.

Permutation Test for Partial Regression Coefficient on First-Order Autocorrelation

Pradthana Minsan[1] *and Pachitjanut Siripanich*[2]

School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand
[1]pradthana.m@gmail.com
[2]oui52@as.nida.ac.th

In the regression model, we assume that the error terms are independent identically distributed. Furthermore, in case of testing hypothesis on the regression coefficient, we assume that the error terms are normally distributed. However, in time series regression, it is quite likely that the errors are serially correlated. The major consequences of the autocorrelation problem are the Ordinary Least Squares (OLS) estimators still remain unbiased, consistent, but they no longer have the minimum variance. Therefore, the usual t and F tests are not generally reliable. An alternative to this traditional is to use a permutation test. Permutation test was applied for linear model in many fields but rarely in time series regression. In this article we proposed permutation test for testing partial regression coefficient for time series with first-order autocorrelation(AR1) where the distribution of error terms is not necessary normal. The Prais-Winsten transformation was applied so that errors are i.i.d. and then exchangeability are also obtained. In addition, we proposed a permutation method that can be directly conducted to the test without fitting back to the model which is not the same as the others well-known permutations test, e.g., Freedman and Lane (1983) and Manly (1997). The asymptotic distribution of the proposed test is also considered. The type I error rate and power of the proposed method and the others permutation methods are studied using numerical simulation.

Sampling Techniques and Data Collection

Algorithms for the Stratification of Skewed Populations

Jane M. Horgan[1] *and Sebnem Er*[2]

[1]Dublin City University, Dublin, Ireland; jane.horgan@dcu.ie
[2]Istanbul University, Istanbul, Turkey; sebnemer@istanbul.edu.tr

Over the past number of years, many computational procedures have been developed to obtain optimum stratification boundaries, those that give the minimum variance (for example, Keskinturk & Er (2007), Kozak (2004), Lavallee & Hidiroglou (1988)). In this paper, we examine these methods in terms of efficiency, ease-of-use and success in arriving at the optimum boundaries, and compare them to the ridiculously simple geometric method of Gunning and Horgan (2004).

Adaptive Cluster Sampling Based on Ranked Sets

Girish Chandra[1] *and Hukum Chandra*[2]

[1]Tropical Forest Research Institute, Jabalpur, India; gchandra23@yahoo.com
[2]Indian Agricultural Statistics Research Institute, Delhi, India; hchandra@iasri.res.in

In many surveys, characteristic of interest is sparsely distributed but highly aggregated, in such situations the adaptive cluster sampling is very useful. Examples of such populations can be found in fisheries, mineral investigations (unevenly distributed ore concentrations), animal and plant populations (rare and endangered species), pollution concentrations and hot spot investigations, and epidemiology of rare diseases. This paper deal with the problem in which the value of the characteristic under study on the sampled places is low or negligible but the neighbourhoods of these places may have a few scattered pockets of the same. We proposed an adaptive cluster sampling theory based on ranked sets. Different estimators of the population mean are considered and the proposed design is demonstrated with the help of two simple examples of small and large populations.

Bayesian Bootstrap Approximation to Sampling Distributions

Naoto Niki[1] *and Yoko Ono*[2]

[1]Tokyo University of Science, Tokyo, Japan; niki@ms.kagu.tus.ac.jp
[2]Niigata University of International and Information Studies, Niigata, Japan;
onoyk@nuis.ac.jp

Bayesian bootstrapping (BBS, in short) in a broad sense, including nonparametric bootstrapping as a special case, is a versatile method for estimating the sampling distribution of an estimator $T(F_n)$ of a population parameter T(F). Here, F_n is the empirical distributions based on observations of size n, F is the (unknown) population and T is a (known) functional with 'gentle characters' typically written as a C-infinity function of moments defined in a neighborhood of the population moments.
Assume that there exists an (unknown) functional T_F such that $T(F) = T_F(U)$ for a (well known) distribution U, e.g., the uniform or normal distributions, then approximation to U by BBS from U_n, i.e., a random sample of size n from U, should be our main concern, since $T(F_n)$ and $T_F(U_n)$ have the same sampling distribution.
Algebraic and numerical works from this viewpoint have revealed that BBS with the Dirichlet prior $Dir[n; c, ..., c]$ for c around 1/2 gives robust approximation to the target distribution. Numerical comparisons are also made with several resampling methods.

Statistical Analyses of the Experimental Results on the Bond Behaviour of Reinforcing Steel and Concrete

M. Sami Donduren[1], *M. Tolga Cogurcu*[2], *Mustafa Altin*[3] *and Mehmet Kamanli*[4]

Selcuk University, Konya, Turkey
[1]sdonduren@selcuk.edu.tr
[2]mtolgac@selcuk.edu.tr
[3]maltin@selcuk.edu.tr
[4]mkamanli@selcuk.edu.tr

In this study, bond behaviour between the concrete and the iron is investigated experimentally with prepared 16 different properties, 8 experiment samples are produced with links and 8 of experiment samples are produced without links. Iron diameter used in links is 8 and diameters of the iron which goes through the midpoint of the samples are 12 (BÇI and BÇIII). Heights and the width of samples are chosen as 20 cm. When choosing sample longitudes, from convection at TS500, touching heights are computed as 30, 50, 60 and 70 cm. Sample concretes are concrete C16 and C25. 3 samples are taken from each one of the concretes prepared and 28 day average cylinder pressure strengths are found as 219.50 for C16 and 293.75 for C25. Experiment samples are prepared in four groups. By keeping the concrete quality the same, the ratio between the maximum loads of the samples which have BÇI and BÇIII irons is found as 1100/5500. By keeping irons the same, the ratio between the maximum loads of the samples which are prepared with C16 and C25 concretes is found as 1100/2500 with respect to the samples which are prepared without links, the samples prepared with links are found to have 33% more average strenghts. Bond results are anlyzed with variance analysis on SPSS program and with security edge controls.

Econometrics

There is Only One Statistical SoftwaRe

Giovanni Millo

Generali R&D, Trieste, Italy; giovanni_millo@generali.com

I show how a variety of tasks in econometric computing, usually performed through spreadsheet manipulation and use of one or more software packages, can be accomplished effectively through R, substituting many different tools and interfaces with just one, free, open and platform-independent software environment ; how in this environment the need for manual intervention in summarizing or documenting the results of research is virtually eliminated through comprehensive automated interfacing; and how new possibilities in econometric programming arise from the distinctive features of the R system. I briefly introduce the R Project and give examples of R's scalability, illustrating the different available levels of usage, from useR of precompiled procedures to programmeR of new methods, to show that R can be far less difficult to use than its reputation goes. Not being tied to any particular GUI and/or editor, R can be used together with a general-purpose editor of your choice. Graphical user interfaces are available as well, so that R can pretty much be anything for everyone. I discuss reproducibility of results as compared with point-and-click statistical software, giving a quick overview of the logging facilities and interfacing possibilities of R geared towards the production of statistical reports and scientific papers. I describe some implementation examples where programming challenges are overcome by taking advantage of object-orientation features, of the status of functions as first-class objects, of the special 'dots' argument, of implicit recursion and of explicit parsing and evaluation of dynamic text. The concept of abstraction is discussed and applied to tasks and data types. Lastly, I give some examples of useful integration with other statistical disciplines through which R overcomes the limitations of typical 'single-purpose' applications, allowing researchers to use one interface and one syntax instead of having to switch between many tools.

Non-linearity, Complexity and Limited Measurement in the Relationship Between Domain and Overall Life Satisfaction

Monica Gonzalez[1], *Germa Coenders*[2], *Marc Saez*[3] *and Ferran Casas*[4]

Quality of Life Research Institute. University of Girona, Girona, Spain
[1]monica.gonzalez@udg.edu
[2]germa.coenders@udg.edu
[3]marc.saez@udg.edu
[4]ferran.casas@udg.edu

In this article we defend that the adoption of a non-linear approach, theoretically framed on complexity theories, can make some contribution to the models which explain the levels of satisfaction with life as a whole through the combination of the levels of satisfaction in different life domains.
Three models have been tested: 1) A linear model; 2) Rojas' (2006) constant elasticity of substitution model; 3) a model with all possible quadratic and interaction product terms (Gonzalez et al. 2006, 2008).
The dependent variable is censored above, which can lead to bias and to spurious non-linear relationships. Therefore, data have been analysed twice taking into account or not limited measurement of satisfaction with life as a whole: by least squares (OLS for the linear and quadratic and interaction term models and non-linear least squares for the constant elasticity of substitution model) and as a censored-above regression model, also known as Tobit model (by maximum likelihood in all models).
Results show that: a) any of the two non-linear models fits better than the linear one; b) any of the models failing to take into account limited measurement fits worse; c) the non-linear model with quadratic terms and interaction effects fits best; d) the rank order of the goodness of fit of the models, and thus model choice, depends on whether limited measurement is accounted for; e) the significance and even sign of the non-linear effects depends on whether limited measurement is accounted for.

Plurality of Methods for the Categorical Variables: Empirics of Microfinance Impacts on Happiness in Thailand and Brazil

Thanawit Bunsit

University of Bath, Bath, United Kingdom; tb238@bath.ac.uk

This paper aims to explore and compare different types of econometric models for capturing impacts of microfinance on the level of self-reported happiness and subjective wellbeing indicators such as life domain satisfaction, positive and negative affects. Using a comparison of microfinance case studies from Thailand (The village fund programme and the savings group fund) and Brazil (Social bank programme), this research examines the impact of small loans for the poor on their happiness and subjective wellbeing indicators. The data is from household surveys and in-depth interviews, collected by the author in 2008-2009 in rural areas of both countries.

When measuring the impact of microfinance on happiness and other related subjective indicators, the issues of discrete choice problems involving binary, ordered or multinomial discrete alternatives cannot be avoided. In this study, for example, self-reported happiness can be measured in a categorical variable (1 = not too happy, 2 = fairly happy, and 3 = very happy). A large body of previous literature acknowledges the inappropriateness of using linear regression when the dependent variable is categorical. By using different types of model, including ordinary least square, ordered probit and ordered logit models, multinomial logit model, binary choice model, and Heckman selection model, results have been generated showing differences and similarities among those methods which can assist in selecting the optimal model.

Mathematical Statistics I

Asymptotic Behavior of the Kernel Regression Estimator Under Different Rates of Censoring and Dependence

Zohra Guessoum[1] *and Elias Ould Said*[2]

[1]University of Scientific and Technologie Houari Boumediene, Algiers, Algeria;
z0guessoum@hotmail.com
[2]Université du littoral Côte d'opale, Lille, France; ouldsaid@lmpa.univ-littoral.fr

Consider a real random variable (rv) Y and a sequence of strictly stationary rv's $(Y_i)_{i\geq 1}$ with common unknown absolutely continuous disribution function (df) F and let $(C_i)_{i\geq 1}$, be a sequence of censoring rv's with common unknown df G. In contrast to statistics for complete data studies, censored model involves pairs $(T_i, \delta_i)_{i=1,\ldots,n}$ where only $T_i = Y_i \wedge C_i$ and $\delta_i = \mathbb{I}_{\{Y_i \leq C_i\}}$ are observed. Let X be an $\mathbb{R}^d$-valued random vector. Let $(X_i)_{i\geq 1}$ be a sequence of copies of the random vector X and denote by $X_{i,1}, \cdots, X_{i,d}$ the coordinates of X_i. The study we perform below is then on the set of observations $(T_i, \delta_i, X_i)_{i\geq 1}$. In regression analysis one expectes to identify, if any, the relationship between the Y_i's and X_i's. This means looking for a function $m^*(X)$ describing this relationship that realizes the minimum of the mean squared error criterion. It is well known that this minimum is achieved by the regression function $m(x)$ defined on $\mathbb{R}^d$ by

$$m(x) = \mathbb{E}\left(Y \,|X = x\right).$$

A feasible estimator of $m(x)$ is given by: $m_n(x) = \frac{r_{1,n}(x)}{\ell_n(x)}$ where

$$r_{1,n}(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} \frac{\delta_i T_i}{\bar{G}_n(T_i)} K_d\left(\frac{x - X_i}{h_n}\right) \tag{1}$$

$l_n(x)$ is the kernel estimate of the density function $l(x)$ of the covariate X and G_n is the Kaplan-Meier (1958) estimator (KME). Under some assumptions me have the rate of uniform convergence

$$O\left(\sqrt{\frac{\log n}{nh_n^d}} + \sqrt{h_n^{d(\nu-2)} \log n}\right) + O\left(h_n\right)$$

Simulations of the convergence are given for increazing size of sample with censoring's and dependent's rates.

Minimal Sufficiency in Rare Populations

Mohammad Moradi[1], *Jennifer Brown*[2] *and Miriam Hodge*[3]

[1]Razi University, Kermanshah, Iran; moradi_m@razi.ac.ir
[2]University of Canterbury, Canterbury, New Zealand;
jennifer.brown@canterbury.ac.nz
[3]University of Canterbury, Canterbury, New Zealand; miriam.hodge@canterbury.ac.nz

It is well understood that for conventional survey designs the set of unordered distinct units in a sample is minimally sufficient statistic. This means that when performing inference with the sample, the sample design is irrelevant; only the value of the sampled units are required. In addition, having less information than this will not allow us to perform inference. We will show that this does not hold for finite rare populations. For finite rare populations the set of unordered distinct rare units in a sample is a minimally sufficient statistic. The information collected from non-rare units, and even the number of non-rare units sampled does not affect the inference we can perform with the sample.

Asymptotic Behavior of a Smooth Conditional Quantile Kernel Estimator for Censored and Dependent Data

Ourida Sadki[1] *and Elias Ould Said*[2]

[1]USTHB, Algiers, Algeria; osadki@usthb.dz
[2]Univ. Littoral, Calais, France; ouldsaid@lmpa.univ-littoral.fr

We study a smooth estimator of the conditional quantile function in the censorship model in the α-mixing case. Consider a sequence of strictly stationary rv's $T_1, T_2, \ldots$ with common unknown absolutely continuous df F. In many situations, we observe only censored lifetimes of items under study. That is, assuming that $C_1, C_2, \ldots, C_n$ are n censoring rv's with common unknown continuous df G, we observe only the n pairs $\{(Y_i, \delta_i)\ , i = 1, 2, \ldots, n\}$, where $Y_i = T_i \wedge C_i$ and $\delta_i = I_{\{T_i \leq C_i\}}$.
Let X be a real-valued rv and $F(.|.)$ be the conditional df of T given $X = x$.
We observe $\{(Y_i, \delta_i, X_i)\ , i = 1, 2, \ldots, n\}$. We suppose that $\{C_i, i \geq 1\}$ and $\{(X_i, T_i), i \geq 1, \}$ are independent.
We also suppose that $\{T_i, i \geq 1, \}$ and $\{C_i, i \geq 1, \}$ are two independent sequences of stationary α-mixing rv's.
For all fixed $p \in]0, 1[$, the pth conditional quantile of F given $X = x$ is defined by

$$\xi_p(x) = \inf\{t \in \Re;\ \ F(t|x) \geq p\}.$$

For our smoothed kernel estimator of the conditional quantile given by

$$\xi_{p,n}(x) = \inf\{t;\ \ F_n(t|x) \geq p\}$$

with

$$F_n(t|x) = \frac{\sum_{i=1}^{n} \frac{\delta_i}{\overline{G}_n(Y_i)} K\left(\frac{x-X_i}{h_n}\right) H\left(\frac{t-Y_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)},$$

where K is a probability density function, h_n is a sequence of positive real numbers which goes to zero as n goes to infinity, H a distribution function and $\overline{G}_n$ is the Kaplan-Meier estimator of G.
We show that this estimator converges uniformly almost surely over a compact set and give a rate of convergence of this estimator; and we establish his asymptotic normality. We also establish that, under some regularity conditions, the kernel estimator of the conditional quantile suitably normalized is asymptotically normally distributed. An application to prediction and confidence bands are given. Some simulations are drawn to lend further support to our theoretical results for finite samples sizes.

Modeling and Simulation II

Loglinear Models for Contingency Table

Justyna J. Brzezińska

Karol Adamiecki University of Economics in Katowice, Katowice, Poland;
justyna.brzezinska@ae.katowice.pl

Loglinear models have many applications in practice. They can be used for the analysis of counts but also in the context of models for two- and more-way contingency tables. In this paper loglinear models are used to model the association between categorical variables. Loglinear models (Poisson regression) will be presented to model counts in two-way contingency tables. The main purpose of this paper is to present different classes of three-way loglinear models. The first model is "complete independence", where there is no interactions and all two-way and three-way interactions are "0". The second model called "joint independence" allows only one-way interaction. The next model is "conditional independence" and this class of models contains two-way interaction. Finally, "homogeneous association" is a model with all two-way interactions. The likelihood ratio and chi-square statistics are measures of goodness of fit for categorical data in contingency table. There are also two criteria that enable to choose among reasonable models: AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria). Loglinear models are important in social and marketing science. With technological achievement and software available data analyses become easier and less complicated for researchers.

A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Contingency Tables Via Simulation

Atilla Göktaş[1] *and Öznur İşçi*[2]

Mugla University, Mugla, Turkey
[1]gatilla@mu.edu.tr
[2]oznur.isci@mu.edu.tr

Spearman and Pearson correlation coefficient, Gamma coefficient, Kendall's tau-b, Kendall's tau-c, and Somers' d are the most commonly used measures of association for doubly ordered contingency tables. So far there has been no study expressing a priority on those measures of association. The aim of this study is to compare those measures of association for several types and different sample sizes of generated squared doubly ordered contingency tables and determine which measures of association are more efficient. It is found that both the sample sizes and the dimension of the doubly ordered contingency tables play a significant role on the effect of those measures of association.

A Comparison of the Most Commonly Used Measures of Association for Doubly Non-Squared Ordered Contingency Tables Via Simulation

Öznur İşçi[1] *and Atilla Göktaş*[2]

Mugla University, Mugla, Turkey
[1]oznur.isci@mu.edu.tr
[2]gatilla@mu.edu.tr

Spearman and Pearson correlation coefficient, Gamma coefficient, Kendall's tau-b, Kendall's tau-c, and Somers' d are the most commonly used measures of association for doubly either squared or non-squared ordered contingency tables. So far there has been no study expressing a priority on those measures of association. The aim of this study is to compare those measures of association for several types and different sample sizes of generated non-squared doubly ordered contingency tables and determine which measures of association are more efficient. It is found that both the sample sizes and the dimension of the doubly non-squared ordered contingency tables play a significant role on the effect of those measures of association.

On Robust Ridge Regression and Robust Liu Estimator

Betül Kan[1], *Berna Yazıcı*[2] *and Özlem Alpu*[3]

[1]Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey; bkan@anadolu.edu.tr
[2]Department of Statistics, Science Faculty, Anadolu University, Eskişehir, Turkey; bbaloglu@anadolu.edu.tr
[3]Eskişehir Osmangazi University, Faculty of Arts and Sciences, Department of Statistics, Eskişehir, Turkey; oalpu@ogu.edu.tr

Multicollinearity is a common problem for multiple regression analysis and it causes coefficient estimates with big variances. There are many methods proposed by different researchers including biased estimation methods like 'Ridge regression' suggested by Hoerl et al. (1970), and 'Liu estimator' suggested by Liu K. (1993) to overcome the multicollinearity. On the other hand, the data set should not include the outliers in order to get representative results for dependent variable. Robust regression methods are commonly used to remove the effects of outliers. In this study robust ridge regression and robust liu estimator are taken into account for a real life problem with both multicollinear and outlier. The results are compared with each other in terms of their effectiveness and biasedness.

Mathematical Statistics II

A Class of Asymptotically Normal Degenerate Quasi U-Statistics

Aluisio Pinheiro[1], *Pranab Kumar Sen*[2] *and Hildete Prisco Pinheiro*[3]

[1]University of Campinas, Campinas, Brazil; pinheiro@ime.unicamp.br
[2]University of North Carolina at Chapel Hill, Chapel Hill, United States; pksen@bios.unc.edu
[3]University of Campinas, Campinas, Brazil; hildete@ime.unicamp.br

In complex diversity analysis, specially arising in genetics, genomics, ecology and other high-dimensional (and sometimes low sample size) data models, typically subgroup-decomposability (analogous to ANOVA decomposability) arises. In group-divergence of diversity measures in a high-dimension low sample size scenario, it is shown that Hamming distance-type statistics lead to a general class of quasi U-statistics having, under the hypothesis of homogeneity, a martingale (array) property, providing key to the study of general (nonstandard) asymptotics. The class of quasi U-statistics based on a general m-th degree kernel, stationary of order r, and having the novelty that it can be applied for any i.i.d. random vectors of arbitrary (and even increasing) dimension K, for which asymptotic normality is proven under mild regularity conditions. Neither the stochastic independence nor homogeneity of the marginal probability laws play a basic role.
The first author acknowledges the support of FAPESP 2009/14176-8 and CNPq.

Ordering Life Distributions Through Interval Entropy Function

Fakhroddin Misagh[1], *Gholam Hossein Yari*[2] *and Rahman Farnoosh*[3]

Islamic Azad University, Science and Research Branch, Tehran, Iran
[1]misagh@iaut.ac.ir
[2]yari@iust.ac.ir
[3]rfarnoosh@iust.ac.ir

Traditional measure of uncertainty is the differential entropy commonly referred to as Shannon information measure. In the literature of information theory, measures of uncertainty in past and residual lifetime distributions have been proposed. In reliability theory and survival analysis, the residual entropy measures the expected uncertainty contained in remaining lifetime of a component and the past entropy can be viewed as the entropy of the inactivity time of a component. Furthermore, the notions of interval entropy have explored the use of information measures for double truncated random variables. In this paper we introduce a new method of ordering among life distributions in terms of interval entropy; we study the Shannon information measure to ordering two sided truncated random variables. The new method is a generalization of dynamic methods of ordering. We discuss some applications of this method in reliability theory and survival analysis.

Statistical Applications II

Hierarchical Clustering of Population Pyramids Presented as Histogram Symbolic Data

Nataša Kejžar[1], *Simona Korenjak-Černe*[2] *and Vladimir Batagelj*[3]

[1]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
natasa.kejzar@fdv.uni-lj.si
[2]Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia;
simona.cerne@ef.uni-lj.si
[3]Faculty of Mathematics, University of Ljubljana, Ljubljana, Slovenia;
vladimir.batagelj@fmf.uni-lj.si

Population pyramid is a very popular presentation of the age-sex distribution of the human population of a particular region. Its shape is influenced not only by demographical indicators, but also by many other social and political characteristics, such as birth control policy, wars, life-style etc.
In the paper Clustering of population pyramids (Korenjak-Černe, Kejžar, Batagelj, Informatica, 2008) clusters of world countries with similar pyramidal shapes were obtained using Ward's hierarchical clustering. The corresponding clusters' shapes can offer additional insight about countries to field-related researchers.
In order to get clusters where the gender and size of population are also taken into account we present data as histogram symbolic data (Billard, Diday, 2006). For their analysis we adapt the generalized Ward's hierarchical clustering procedure (Batagelj, 1988).
The changes of the pyramids' shapes, and also changes of the countries inside main clusters will be examined for the years 1996, 2001, and 2006.

The Survey Study about the Lifestyle of Young Generations in Thailand

Vacharaporn Suriya-bhivadh

Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand; Vacharaporn@acc.chula.ac.th

The behaviors of young generations (18 to 29 years of age) in Thailand had changed and deviated enormously from the traditional way of the Thai culture and affected the society at large. To learn and understand their lifestyles might lessen the gap between the elders and youngsters, and thus would help lessening the social problems. This survey research aims to study three main aspects of young generations' lifestyle; attitudes towards education, attitudes towards the traditional Thai culture and their usage of high technology devices – computer and mobile phone.
By using the questionnaires and face to face interview, the data were collected from four groups of people – those who were 18 to 21 years, 22 to 25, 26 to 29 and those over 30. The last group gave the pictures of how the adults thought towards the young generations. A total of 1020 young people and 430 older people were included and the survey time spanned from June 2008 to April 2010. By using Descriptive Statistics and the statistical test of difference, it was found that fifty percent of younger groups valued the higher educations and more than half of each young sample group still treasured the old traditional Thai culture. Internet and mobile phone were now parts of their daily lives. However it was found that the three young groups were quite different in spending their money and time on the mobile phone and computer. The adults expected the youngsters to be more helpful to the elders and be more restricted to the traditional culture than they did at present. Should the statistical model about young generations' lifestyle be built, the explanation about the ways of young generations behaviors will be clearer and more interesting.

Numerical Realization Nonlinear Regressions Dependences

Victor Lyumkis

Transport and communication institute, Riga, Latvia; vlyumkis@yahoo.com

The problem of a finding of estimations of parameters on observations of a dependent variable y at known values x in nonlinear regressions models remains actual. The method of the least squares applied to the solving of such problems, remains convenient means, however in some cases there are arises the problems connected, for example, with bad stability of calculations. In the present work the accent on use of the modern integrated packages, in which quality of application of numerical methods rather effectively is done. Such package is free-of-charge package R, which author recommends, in particular, at realization logit and probit nonlinear regresses. The accent on construction of these regresses also for the so-called grouped data is done and corresponding estimations of the maximal likelihood are realized. Thus the grouped data are understood as such data when some observations of a dependent variable y take place at the same value of an explaining variable x. In work opportunities of realization of nuclear estimations regression dependences also are discussed. Now in package R convenient means of a finding of nuclear estimations with a simultaneous optimum choice of a step of a window h are offered. Various examples of a finding of parameters nonlinear regression models by classical methods and their comparison with nuclear models in package R are considered. In our opinion, introduction of package R in practice of work of engineers at forecasting various indicators would be rather reasonable.

Applications of Heavy Tailed Distributions for Modeling Human Behavior and Activities in Cyber Space

Mohammad Ali Baradaran Ghahfarokhi[1], *Parvin Baradaran Ghahfarokhi*[2] *and Rahim Bahramian Dehkordi*[3]

[1]Ministry of Labor and Social Affairs - International Labor Organisation, Tehran, Iran; mali.baradaran@gmail.com
[2]Faculty of Entrepreneurship, Tehran University , Tehran , Iran; p.baradaran.g@gmail.com
[3]Faculty of Management - Tehran University, Tehran, Iran; R_BAHRAMIAN_D@yahoo.com

Nowadays, in this modern society, an important part of human activities and behavior leaves electronic traces in cyber space in form of server logs, e-mails, loan registers, credit card transactions, making telephone calls, online games, web browsing, computer operations, blogs , etc. This huge amount of generated data allows to observe human behavior and communication patterns which could be worthy in economical and business like advertisements, market researches and social and behavioral science. One important question is which model can clearly indicate the validity of distributions in mimicking the human dynamics in many real life systems and the other is how we can estimate the factors and coefficients of the model more precisely. For example when individuals execute tasks based on some perceived priority, the timing of the tasks will be heavy tailed.
In this paper, first we show the characteristics of human activities and dynamics in cyber space then we explain why heavy tailed distribution is a good candidate for modeling in this case with analyzing data. At the end, we compare the estimated parameters of heavy tailed distribution with other distributions and we demonstrate and interpret our results.

Modeling and Simulation III

Control Charts For Weibull, Gamma and Lognormal Distributions

Derya Çalışkan[1] *and Canan Hamurkaroglu*[2]

Hacettepe University, Ankara, Turkey
[1]deryacal@hacettepe.edu.tr
[2]caca@hacettepe.edu.tr

In this study the control limits of $\bar{X}$ and R control charts for skewed distributions are obtained by considering the classic, the weighted variance (WV), the weighted standard deviations (WSD) and skewness correction (SC) methods. These methods are compared by using Monte Carlo simulation. Type I risk probabilities of these control charts are compared with respect to different subgroup sizes for skewed distributions which are Weibull, gamma and lognormal. Simulation results show that Type I risk of SC method is less than that of other methods. When the distribution is approximately symmetric, then the type I risks of Shewhart, WV, WSD,and SC $\bar{X}$ charts are comparable, while the SC R chart has a noticeable smaller Type I risk.

Statistical Modulation of a Human Health Problem in Albania

Luela Prifti[1]*, Etleva Beliu*[2] *and Shpetim Shehu*[3]

Polytechnic University of Tirana, Tirana, Albania
[1]luela_p@yahoo.com
[2]etlevallagami@yahoo.com
[3]shpetimi49@yahoo.com

The air pollution from the industry activity is very dangerous for the human health. This paper aims to analyze the data collected in three sites: two polluted and the ones not, positioned in the south of Albania. Using ANOVA we analyze the influence of the site in the hematological and pneumological field. We build a multivariable regress model for the pneumology using smoke, time stay and the age as independent variables in this model. The covariance method used on the model shows that avoiding the smoke variable there is no difference between three sites in the pneumological field. The dependence of the smoke from the time stay is shown using the multi ANOVA method.

Non-linear Dimensionality Reduction for Functional Computer Code Modeling

Benjamin Auder

UPMC Paris 6, Paris, France; Benjamin.Auder@gmail.com

Estimating quantities from high-dimensional data sets is very challenging. One way of dealing with them is to find appropriate d-dimensional representations with small d. Our work is about dimensionality reduction of continuous functions from $[a, b]$ to R, sampled on a finite grid of D elements. At a larger scale, we develop a model for the computer code CATHARE (from French nuclear research center). This code, written Φ, simulates the thermohydraulic behavior of some parts of a nuclear reactor's vessel; it has p-dimensional inputs and continuous curves as outputs. A few hundreds samples $(x_i, y_i = \Phi(x_i))$ are available, where x_i are initial state parameters.
The model built is composed of three main parts

1. dimensionality reduction: each y_i is represented by $z_i \in R^d$, satisfying some topological constraints;
2. regression to learn the relation between inputs x_i and representations z_i;
3. mapping of vectorial representations $\hat{z}$ to the estimated corresponding curves $\hat{y}$. The two last steps allow to predict new outputs.

The main assumption for the first step is that output curves lie on a (smooth) manifold. Three dimensionality reduction methods are compared: Functional Principal Component Analysis, and two nonlinear approaches (Lin et al., Riemannian Manifold Learning for Nonlinear Dimensionality Reduction, in European Conference on Computer Vision, 2006, pp. 44-55; Zhan et al., Incremental Manifold Learning Algorithm Using PCA on Overlapping Local Neighborhoods for Dimensionality Reduction, in International Symposium on Advances in Computation and Intelligence, 2008, pp. 406-415).
Predicted curves are generally more trustful visually in the non-linear case, because small irregularities are preserved (although not optimally). The curves obtained using the PCA basis look somewhat too much smooth. The validation step shows that both methods are applicable, with future possible improvements of the non-linear ones.

A Bayesian Approach to Inferring the Contribution of Unobserved Ground Conditions to Observed Scores in Sports: The Example of Cricket

Scott R. Brooker[1] *and Seamus Hogan*[2]

Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand
[1]srbrooker@gmail.com
[2]seamus.hogan@canterbury.ac.nz

This paper is part of a wider research programme using a dynamic-programming approach to modelling the choices about the amount of risk to take by teams and players in International Cricket. An important confounding variable in this analysis is the ground conditions (size and shape of stadium, condition of playing surface and weather conditions) that affect the trade off between risk and return that teams and players face. This variable does not exist in our historical data set and would in any event be very difficult to accurately observe on the day of a match.
In this paper, we consider a way of estimating a distribution for the ground conditions using only the information contained in the scores and result of the match. In our approach we use the difference between the cumulative density function of scores and a probit estimate of the probability of each score being a winning score in order to infer the extent to which high scores on average reflect easy conditions rather than good performance. Using a Monte Carlo method we estimate the percentage of the variation in total scores that is due to the variation in conditions and we subsequently use Bayes' Law to estimate a distribution of conditions for each match. We develop our method using the example of cricket and we outline some potential applications of the method to other sporting contests.

Workshop

Practical Applications of Permutation Tests

Phillip Good

statcourse.com, United States of America; drgood@statcourse.com
http://videolectures.net/as2010_good_pap/

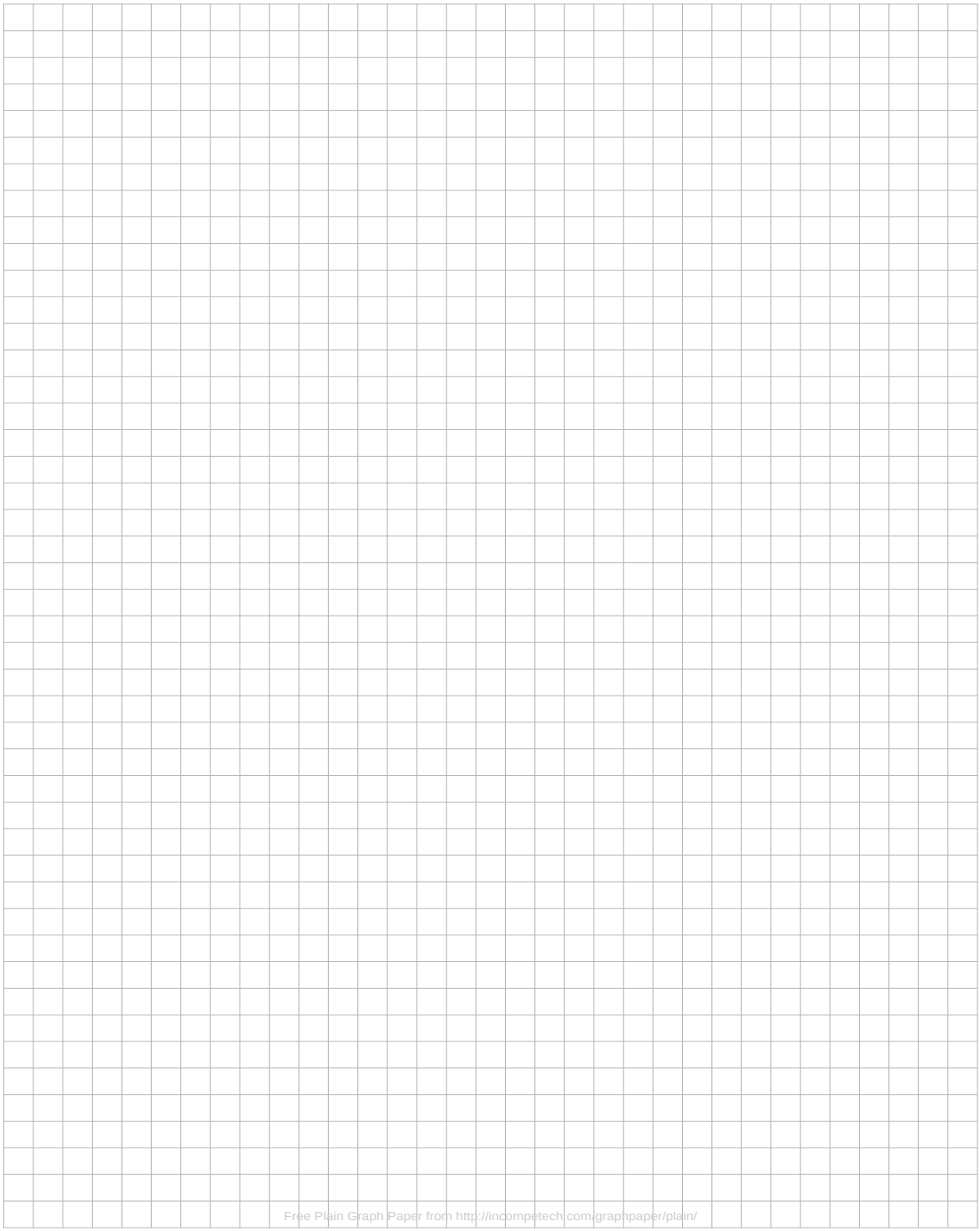
As research during the past sixty-five years has amply demonstrated, permutation tests should be used to do all of the following:

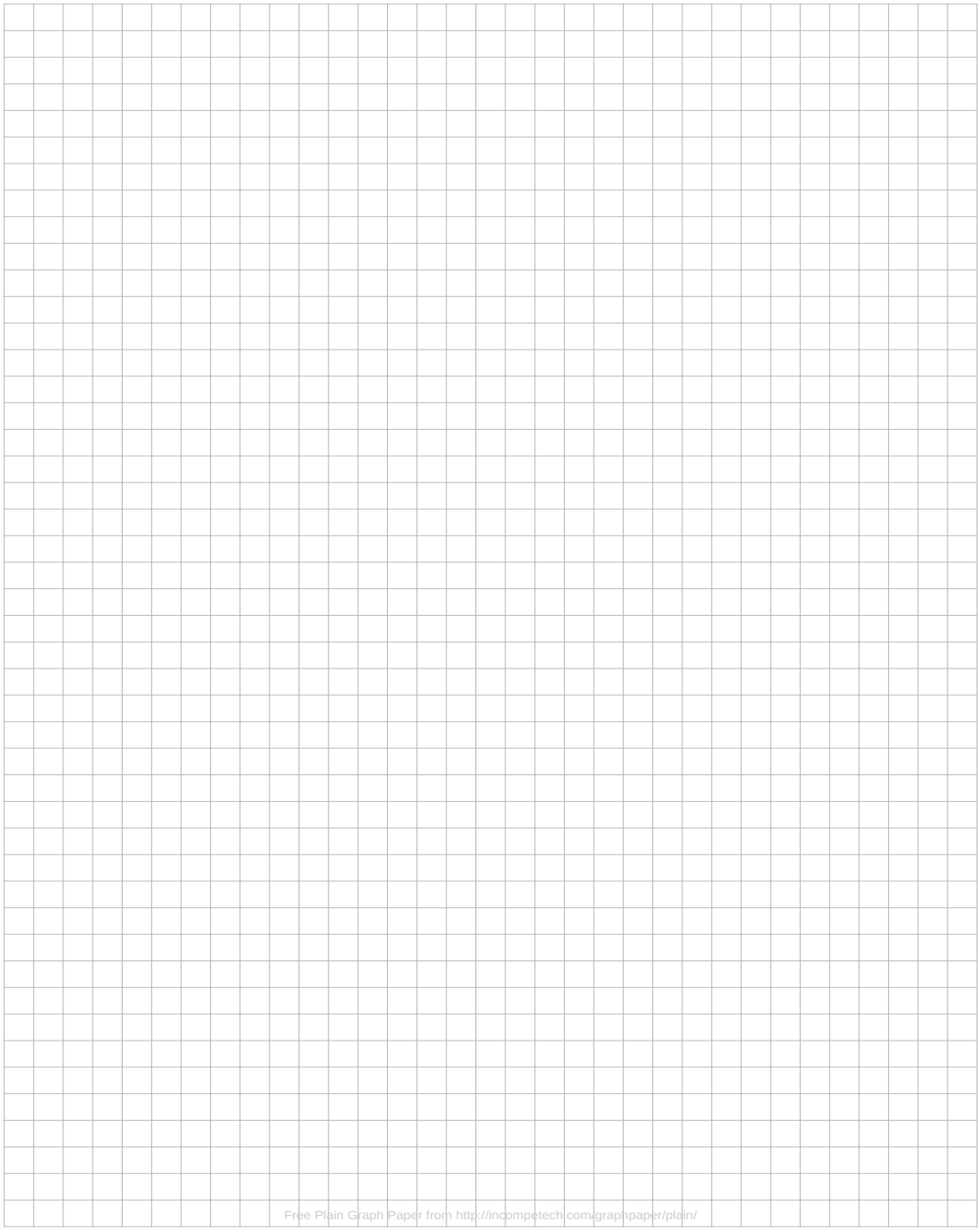
1. Make a multivariate comparison of population means.
2. Analyze the one-way layout.
3. Compare variances.
4. Analyze cross-over designs.
5. Analyze contingency tables.

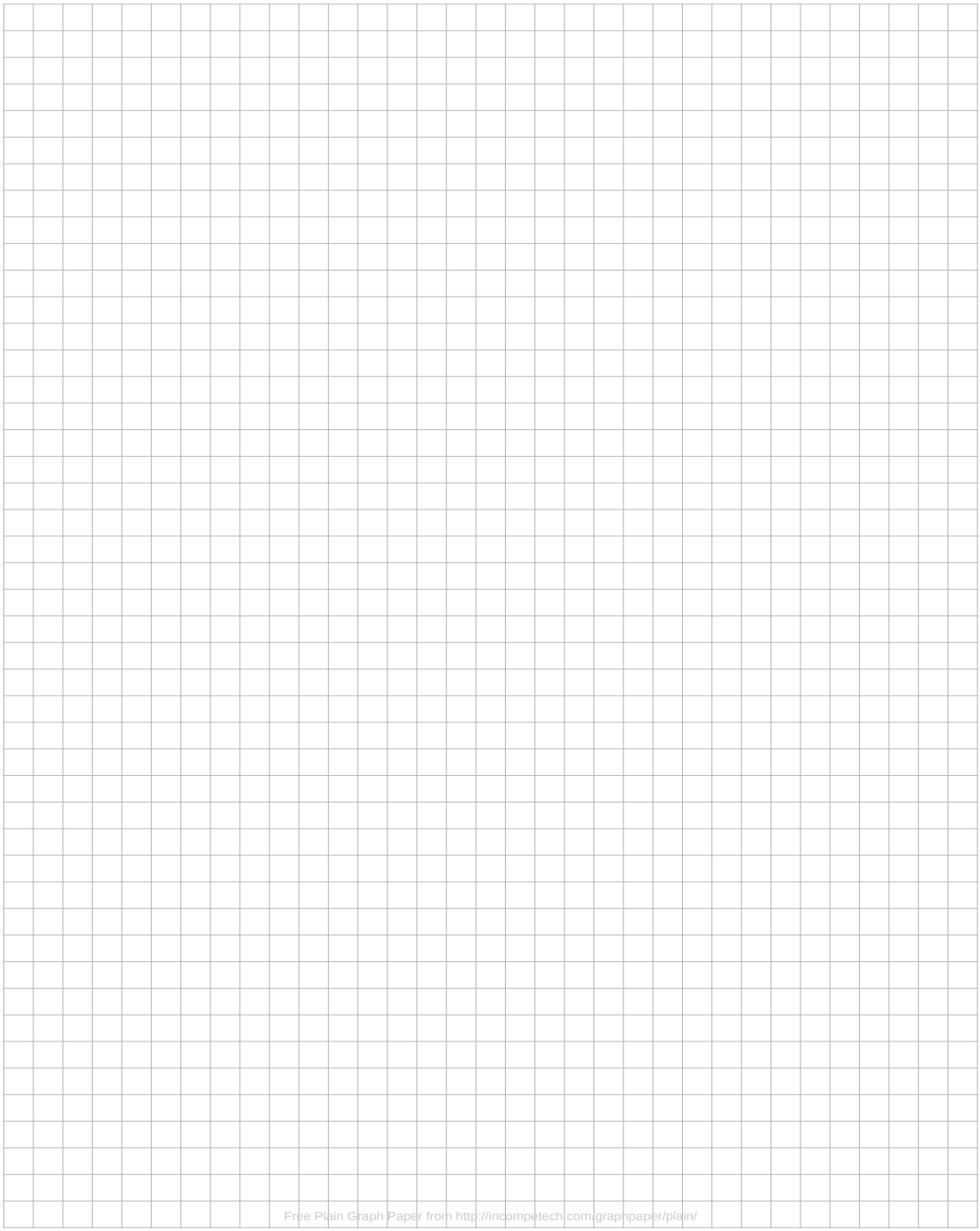
For permutation tests are exact, distribution free, and most powerful when the observations are exchangeable. We consider their application to such disparate fields as archeology, microarrays, and quality control.

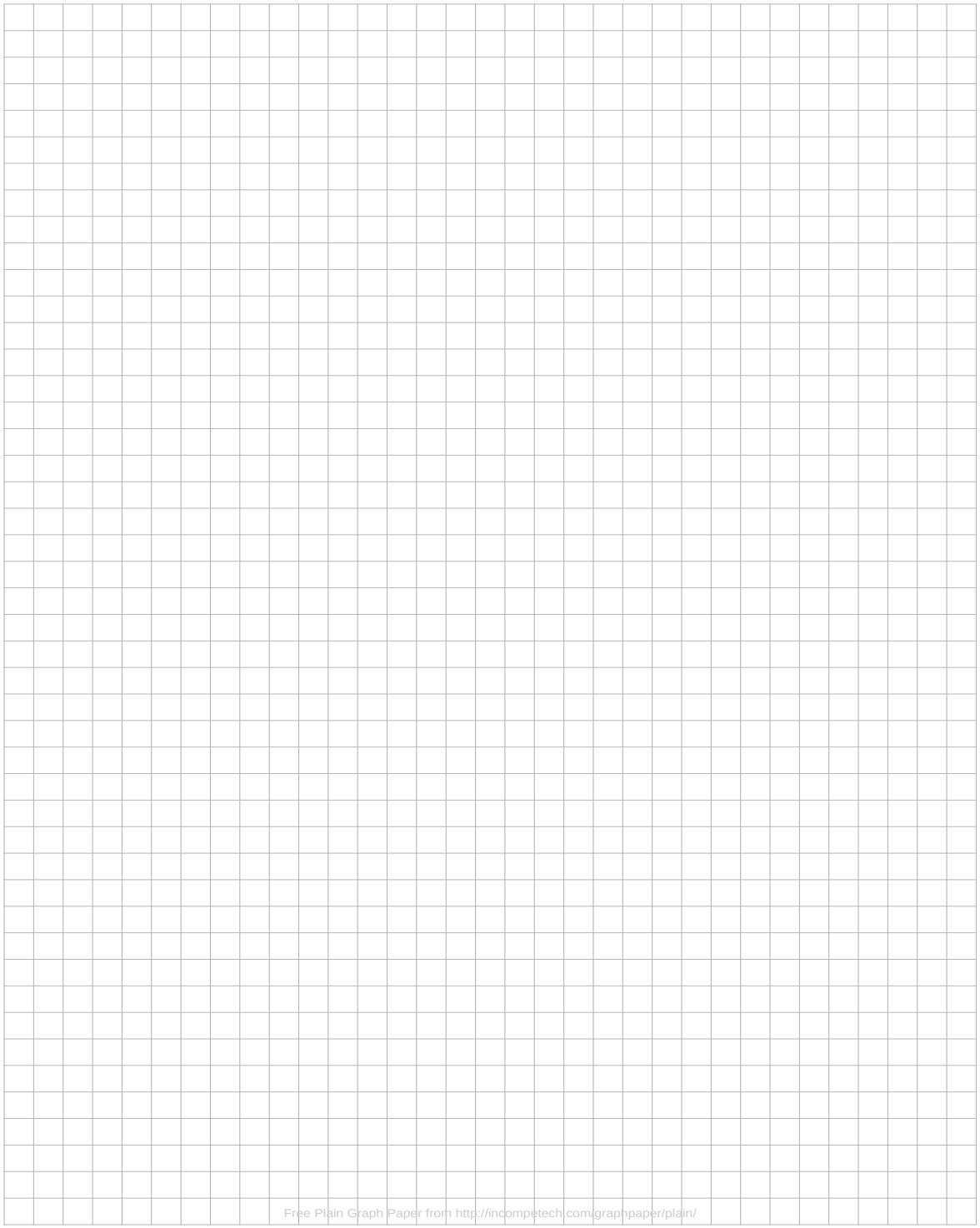
INDEX OF AUTHORS

Index of Authors


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/


Free Plain Graph Paper from http://incompetech.com/graphpaper/plain/

SUPPORTED BY

www.arrs.gov.si/en

www.valicon.si

www.alarix.si

www.result.si

www.sweetsurveys.com