International Conference

APPLIED STATISTICS
2009

PROGRAM and ABSTRACTS

September 20 - 23, 2009

Ribno (Bled), Slovenia

International Conference

APPLIED STATISTICS 2009

PROGRAM and ABSTRACTS

September 20 – 23, 2009

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 (2009; Ribno)
Program and abstracts [Elektronki vir]/International Conference Applied Statistics 2009, September 20–23, 2009, Ribno (Bled), Slovenia ;
[organized by Statistical Society of Slovenia ; edited by Lara Lusa and Janez Stare].
- Ljubljana : Statistical Society of Slovenia, 2009
Način dostopa (URL): http://conferences.nib.si/AS2009/AS2009-Abstracts.pdf

ISBN 978-961-92487-3-7 1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
247358720

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	Biostatistics and Bioinformatics I	Canonical Correlation Analysis
	13.40 – 15.00	Lunch	
	15.00 – 16.20	Biostatistics and Bioinformatics II	Network Analysis
	16.20 – 16.40	Break	
	16.40 – 17.40	Biostatistics and Bionformatics III	Sampling Techniques and Data Collection
	19.00	Reception	
Monday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Bayesian Statistics	Statistical Applications - Economics I
	11.40 – 12.00	Break	
	12.00 – 13.00	Social Science Methodology	Education
	13.00 – 14.30	Lunch	
	14.30	Excursion	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.20	Mathematical Statistics	Bibliometrics
	11.20 – 11.40	Break	
	11.40 – 13.00	Econometrics I	Design of Experiments
	13.00 – 14.30	Lunch	
	14.30 – 15.30	Econometrics II	Data Mining
Wednesday	9.30 – 10.30	Statistical Applications - Economics II	Modeling and Simulation
	10.30 – 10.50	Break	
	10.50 – 12.10	Measurement	
	12.10 – 12.30	Closing of the conference	
	12.30 – 14.00	Lunch	
	14.00 – 17.30		Workshop

10.30–11.00 **Registration**

11.00–11.10 **Opening of the Conference** (Hall 1)

11.10–12.00 **Invited Lecture** (Hall 1) *Chair: Janez Stare*

1. **Every Missing not at Random Model for Incomplete Data Has Got a Missing at Random Counterpart with Equal Fit** *Geert Molenberghs, Michael G. Kenward, Geert Verbeke, Caroline Beunckens and Cristina Sotto*

12.00–12.20 **Break**

12.20–13.40 **Biostatistics and Bioinformatics I** (Hall 1) *Chair: Geert Molenberghs*

1. **Pseudo-Observations in Survival Analysis** *Maja Pohar Perme and Per Kragh Andersen*
2. **An Overwiev of Closed Form Methods for the Analysis of Clustered Data** *Yasemin Genç, Derya Öztuna, Selcen Yüksel and Can Ateş*
3. **Statistical Tools for Basic Research in Agro-Homeopathy** *Maurizio Brizzi and Lucietta Betti*
4. **Composite Interval Mapping and Skew-Normal Distribution** *Elisabete B. Fernandes, António Pacheco and Carlos Penha-Gonçalves*

12.20–13.20 **Canonical Correlation Analysis** (Hall 2) *Chair: Anuška Ferligoj*

1. **Effects of Data Categorization on Canonical Correlation Analysis: A Simulation Study** *Matjaž Širca, Špela Jezernik and Boris Cergol*
2. **The Impact of Missing Values on the Stability of Canonical Correlation Analysis Results: Simulations Using ESS Data** *Rok Platinovšek, Rok Okorn and Nejc Berzelak*
3. **Canonical Correlation Analysis: Influence of Different Methods on Teatments of Missing Values on Simulated Data** *Nina Klenovšek, Aleš Toman, Kocar Sebastjan and Romana Štokelj*

13.40–15.00 **Lunch**

15.00–16.20 **Biostatistics and Bioinformatics II** (Hall 1) *Chair: Maurizio Brizzi*

1. **Changes in Forest Species Composition in the Last 50 Years on the Snežnik Area in Slovenia Studied by Compositional Data Approach** *Damijana Kastelec, Milan Kobal and Klemen Eler*
2. **Subgroup Discovery in Data Sets with Multi-Dimensional Responses: A Method and a Case Study in Traumatology** *Lan Umek, Blaž Zupan, Marko Toplak, Annie Morin, Jean-Hugues Chauchat, Gregor Makovec and Dragica Smrke*

3. **Class Imbalance Problem for Multivariate Classifiers Using High-Dimensional Data** *Rok Blagus and Lara Lusa*

15.00–16.20 **Network Analysis** (Hall 2) *Chair: Anuška Ferligoj*

1. **Online Dictionary of the Social Sciences as a Network** *Anja Žnidaršič*
2. **Dynamics of Scientific Co-authorship Networks of Slovenian Researchers** *Luka Kronegger*
3. **Clustering of Discrete Distributions: New R Package and Application on US Patent Data** *Nataša Kejžar, Simona Korenjak-Černe and Vladimir Batagelj*
4. **"Please Name the First Two People You Would Ask for Help": The Effect of the Limitation of the Number of Alters on Network Composition** *Tina Kogovšek and Valentina Hlebec*

16.20–16.40 **Break**

16.40–18.00 **Biostatistics and Bionformatics III** (Hall 1) *Chair: Gaj Vidmar*

1. **Susceptibility to Environment of Barley DH Lines Examined in a Series of Trials** *Tadeusz Adamski , Zygmunt Kaczmarek, Stanisław Mejza and Maria Surma*
2. **A Multivariate Approach to the Evaluation of Oilseed Rape Doubled Haploids on the Basis of a Series of Line x Tester Experiments** *Zygmunt Kaczmarek, Stanisław Mejza, Elzbieta Adamska, Teresa Cegielska-Taras and Laura Szała*
3. **Comparison of Ozone-Caused Tobacco Leaf Injury Degree Between Rural and Urban Exposure Sites** *Anna Budka, Klaudia Borowiak, Dariusz Kayzer and Janina Zbierska*
4. **Analysis of Hybrids Obtained by Diallel Crossing Scheme Based on Categorical Trait** *Anita Biszof and Stanisław Mejza*

16.40–17.40 **Sampling Techniques and Data Collection** (Hall 2) *Chair: Aleš Žiberna*

1. **Optimal Compression of Statistical Data by Minimization of Information Cost Function** *Igor Grabec*
2. **Performing Descriptive Statistical Analysis on Large Univariate Numerical Datasets Using Spreadsheets** *Ilija S. Hristoski*

19.00 **Reception**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Andrej Blejec*

1. **Creating Structured and Flexible Models: Some Open Problems** *Andrew Gelman*

10.00–10.20 **Break**

10.20–11.40 **Bayesian Statistics** (Hall 1) *Chair: Andrew Gelman*

1. **Building a Cross-Classified Multilevel Structure for Credit Scorecard: A Bayesian Monte Carlo Markov Chain Approach** *Alesia S. Khudnitskaya*
2. **Bayesian Model Averaging for Hierarchical Log-Linear Models and an Application to the Dumping Severity Data** *Haydar Demirhan and Canan Hamurkaroglu*
3. **Endogeneity of Store Attributes in Heterogeneous Store-Level Sales Response Models** *Harald Hruschka*

10.20–11.40 **Statistical Applications - Economics I** (Hall 2) *Chair: Janez Stare*

1. **Advertising Expenditure Components in Time; A Case Study of 17 European Countries in the Period 1994-2007** *Katarina Košmelj and Vesna Žabkar*
2. **Investigating the Relationship Between Electricity Consumption, Export and GDP in Turkey via the Multivariate Time Series Analysis** *Derya Ersel, Yasemin Kayhan Atilgan and Süleyman Günay*
3. **Nonlinearities, Weak Integration and the Globalization of Stock Markets** *Rui C. Menezes*
4. **Parametric Estimation of the Customer Churn Risk** *Sofia L. Portela and Rui C. Menezes*

11.40–12.00 **Break**

12.00–13.00 **Social Science Methodology** (Hall 1) *Chair: Tina Kogovšek*

1. **Ordered Dissimilaritiy of Students Gradings Distributions – Ordinal Statistics to Answer Ordinal Questions** *Matevž Bren and Darko Zupanc*
2. **The Cultural Capital of Immigrant Families and the Impact on Student Performance** *Alina S. Botezat*
3. **Changes in the Marital Structure of the Population of Vojvodina in Respect to their National-Ethnic and Confessional Affiliation** *Katarina J. Čobanović and Valentina T. Sokolovska*

12.00–13.00 **Education** (Hall 2) *Chair: Patrick Wessa*

1. **Difficulties in Teaching Statistics in Slovenian Secondary Schools** *Andreja Drobnič Vidic and Simona Pustavrh*
2. **Mathematics in Slovene General Matura – Discrepancy of Grades at Basic and Higher Level of Achievement** *Alenka Hauptman*
3. **Does Examinee Choice in Educational Testing Work? One Example from Slovenian General Matura Examinations** *Gašper Cankar*

13.00–14.30 **Lunch**

14.30 **Excursion**

9.10–10.00 **Invited Lecture** (Hall 1) *Chair: Matevž Bren*

1. **Statistics of Compositional Data** *Gerald van den Boogaart*

10.00–10.20 **Break**

10.20–11.20 **Mathematical Statistics** (Hall 1) *Chair: Gerard van den Boogaart*

1. **Fuzzy Hypothesis Testing in Linear Regression** *Duygu İçen and Süleyman Günay*
2. **Incongruence of Model Fit Indices and Other Evidence of Model Quality in SEM** *Roman Konarski*

10.20–11.20 **Bibliometrics** (Hall 2) *Chair: Janez Stare*

1. **The Use of Statistical Methods in Library and Information Science Literature** *Güleda Düzyol and Sevil Bacanli*
2. **Peer-Reviews and Bibliometrical Methods: Two Sides of the Same Coin?** *Primož Južnič, Matjaž Žaucer, Miro Pušnik, Tilen Mandelj, Stojan Pečlin and Franci Demšar*
3. **Measuring the Citation Impact of Statistic Journals with Structural Equation Modelling Analysis** *Güleda Düzyol, Duygu İçen and Süleyman Günay*

11.20–11.40 **Break**

11.40–13.00 **Econometrics I** (Hall 1) *Chair: Stanslav Mejza*

1. **Testing Tobler's Law in Spatial Panels: A Test for Spatial Dependence Robust Against Common Factors** *Giovanni Millo*
2. **Modeling SBITOP Stock Index Time Series Through Decomposition** *Aleša Lotrič Dolinar*

11.40–13.00 **Design of Experiments** (Hall 2) *Chair: Katarina Košmelj*

1. **On the Efficiency of Some Incomplete Split-Plot $\times$ Split-Block Designs with Control Treatments** *Katarzyna Ambroży and Iwona Mejza*
2. **Graphic Analysis of Interaction in Full Factorial Designs: A Critical Study** *Dulce G. Pereira and Paulo Infante*
3. **Statistical Quality Control for Business Indicators of Healthcare Quality: General Considerations and a Specific Proposal** *Gaj Vidmar and Rok Blagus*

13.00–14.30 **Lunch**

14.30–15.30 **Data Mining** (Hall 1) *Chair: Lluis Coromina*

1. **Discriminant Analysis Versus Random Forests on Qualitative Data: Contingent Valuation Method Applied to the Seine Estuary Wetlands** *Salima Taibi and Dimitri Laroutis*

2. **Using Decision Trees for Classification in Actuarial Analysis** *Damla Barlas and Omer Esensoy*
3. **Analysis of the eDonkey Data – In Search of Pedophilia** *Aleš Žiberna, Vasja Vehovar and Matej Kovačič*
4. **Supervised Learning for Automatic Target Recognition** *Gerard Brunet*

14.30–15.30 **Econometrics II** (Hall 2) *Chair: Aleša Lotrič Dolinar*

1. **Management Accounting in Austrian Family Enterprises - An Empirical Research** *Christine Duller*
2. **The Attitude of the Government to the Arab Minority in Israel: A Study of Government Fiscal Allotments to Local Authorities** *Tal Shahor*

9.30–10.30 **Statistical Applications - Economics II** (Hall 1) *Chair: Jože Rovan*

1. **Path Analysis and Cumulative Measures Applied to the Spanish Customer Satisfaction Indices: A Marketing Application for Automobile Industry** *López Caro Cristina, Mariel Chladkova Petr and Fernandez Aguirre Karmele*
2. **Forecasting of Incurred But Not Reported Reserves with Generalized Linear Models** *Tuğba Tunç and Gençtürk Yasemin*

9.30–10.30 **Modeling and Simulation** (Hall 2) *Chair: Vasja Vehovar*

1. **Poisson Mixture Regression Model: Application to Financial Data** *Fátima Gonçalves and Susana Faria*
2. **A Group Sequential Form for the Partially Grouped Log Rank Test and A Simulation Study** *Yaprak Parlak Demirhan and Haydar Demirhan*

10.30–10.50 **Break**

10.50–12.10 **Measurement** (Hall 1) *Chair: Damijana Kastelec*

1. **Measurement and Expressions of Uncertainty of Material Characteristics** *Athanasios Papargyris and Dimitrios Papargyris*
2. **Measurement of Supranational Policy Level of Decision Making. A Practical Method for Helping Policy Makers** *Lluis Coromina and Willem E. Saris*
3. **Pitfalls and Remedies in Testing the Calibration Quality of Rating Systems** *Wolfgang Aussenegg, Florian Resch and Gerhard Winkler*

12.10–12.30 **Closing of the conference** (Hall 1)

12.30–14.00 **Lunch**

14.00–17.30 **Workshop** (Hall 2)

1. **Statistics Education and Educational Research Based on Reproducible Computing** *Patrick Wessa, Bart Baesens, Stephan Poelmans and Ed van Stee*

ABSTRACTS

Invited Lecture

Every Missing not at Random Model for Incomplete Data Has Got a Missing at Random Counterpart with Equal Fit

Geert Molenberghs[1,2], *Michael G. Kenward*[3], *Geert Verbeke*[2,1], *Caroline Beunckens*[1] *and Cristina Sotto*[1]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium;
[2]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Katholieke Universiteit Leuven, Belgium;
[3]Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK
geert.molenberghs@uhasselt.be

Over the last decade, a variety of models to analyze incomplete multivariate and longitudinal data have been proposed, many of which allowing for the missingness to be not at random (MNAR), in the sense that the unobserved measurements influence the process governing missingness, in addition to influences coming from observed measurements and/or covariates. The fundamental problems implied by such models, to which we refer as sensitivity to unverifiable modeling assumptions, has, in turn, sparked off various strands of research in what is now termed sensitivity analysis. The nature of sensitivity originates from the fact that an MNAR model is not fully verifiable from the data, rendering the empirical distinction between MNAR and random missingness (MAR), where only covariates and observed outcomes influence missingness, hard or even impossible, unless one is prepared to accept the posited MNAR model in an unquestioning way. We show that the empirical distinction between MAR and MNAR is not possible, in the sense that each MNAR model fit to a set of observed data can be reproduced exactly by an MAR counterpart. Of course, such a pair of models will produce different predictions of the unobserved outcomes, given the observed ones. This is true for any model, whether formulated in a selection model (SeM), pattern-mixture model (PMM), or shared-parameter model (SPM) format. Specific attention will also be given to the SPM case, since we are able to provide a formal definition of MAR in this case. Theoretical considerations are supplemented with illustrations based on a clinical trial in onychomycosis and on the Slovenian Public Opinion survey. The implications for sensitivity analysis are discussed. Missing data can be seen as latent variables. Such a view allows extension of our results to other forms of coarsening, such as grouping and censoring. In addition, the technology applies to random effects models, where a parametric form for the random effects can be replaced by certain other parametric (and non-parametric) form, without distorting the model's fit, latent classes, latent variables, etc.

Biostatistics and Bioinformatics I

Pseudo-Observations in Survival Analysis

Maja Pohar Perme[1] *and Per Kragh Andersen*[2]

[1]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia; maja.pohar@mf.uni-lj.si
[2]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark; P.K.Andersen@biostat.ku.dk

Survival analysis has developed as an independent area within statistics where methods dealing with incomplete data (right-censoring, left-truncation) are discussed. Without incomplete data, the survival time T would be observed for all individuals and standard methods for quantitative data or methods for binary outcomes could be applied. Pseudo-observations present a general approach to analyzing survival data as they are defined for each individual at each follow-up time. We review the idea of pseudo-observations and illustrate their application to various regression models as well as graphical goodness of fit testing. The methods are illustrated using a data set from bone marrow transplantation.

An Overwiev of Closed Form Methods for the Analysis of Clustered Data

Yasemin Genç[1], *Derya Öztuna*[2], *Selcen Yüksel*[3] *and Can Ateş*[4]

Department of Biostatistics, Faculty of Medicine, Ankara University, Ankara, Turkey
[1]genc@medicine.ankara.edu.tr
[2]dgokmen2001@yahoo.com
[3]selcenpehlivan@yahoo.com
[4]can.ates@gmail.com

The most common problem encountered in many medical studies is the incorrect statistical analysis of multiple observations taken from the same subject. While observations from different subjects can be considered statistically independent, different observations from the same subject are correlated. This data structure is called as "clustered data" in statistical literature. According to simulation studies, when the within-correlation coefficient is positive, ignoring the correlation results in inflation to Type I error rate. Clustered data occur frequently in several fields of studies. For example, in periodontal studies, observations can be taken from multiple sites (gums, teeth or tooth surfaces) on each subject. In ophthalmologic studies, the subject again is the cluster, but two eyes are measured, for example; the presence of uveitis can be exposed by considering either or both eyes of a patient with Behçet's disease. A special class of the studies is community intervention trials in which medical practices, factories, or entire cities are taken as cluster. This study reviews closed-form methods for the analysis of clustred data. In addition, user-friendly program "ClusteredData" written in C# language on Microsoft Visual Studio.NET 2005 platform, will be introduced. This program compares two dependent/independent proportions or two means from independent samples, calculates confidence intervals for sensitivity/specificity and area under receiver operating characteristics (ROC) curve and compares two areas under ROC curves.

Statistical Tools for Basic Research in Agro-Homeopathy

Maurizio Brizzi[1] *and Lucietta Betti*[2]

[1]Dipartimento di Scienze statistiche "Paolo Fortunati", University of Bologna, Bologna, Italy; maurizio.brizzi@unibo.it
[2]Dipartimento di Scienze e tecnologie agro-ambientali, University of Bologna, Bologna, Italy; lucietta.betti@unibo.it

The effectiveness of homeopathic treatments has been discussed for many years among physicians and researchers, being still an open point in scientific community. This field surely deserves a thorough statistical analysis in order to add empirical evidence to personal opinions of "experts", either in favour or against this therapies. One of the most repeated criticisms is the eventuality of a placebo effect; a possible way to avoid such effect is to apply experimental models where "patients" are not human beings, but plant species. A series of experiments was carried out by an Italian research group (including the Authors), during the last 15 years, in which large samples of wheat seeds and seedlings were previously stressed with material doses of Arsenic trioxide (As2O3) and then treated either with distilled water (control group) or with Arsenic trioxide (AP kdH), diluted and potentised at the k-th potency. The main working variables are germination (number of non-germinated seeds out of 33, located in a common Petri dish, after 96 hours) and seedling growth (stem length, after the same period of observation), in order to detect if there may be a stimulating effect in AP-treated samples. In particular, AP45dH treatment seems to show a repeatedly significant stimulating effect, both in germination and seedling growth, and even to induce a relevant decrease in variability of experimental results. The present study is a meta-analysis of these findings, with a special stress on the role played by statistical methods (Poisson model, Odds Ratio, Non-parametric tests and others).

Composite Interval Mapping and Skew-Normal Distribution

Elisabete B. Fernandes[1], *António Pacheco*[2] *and Carlos Penha-Gonçalves*[3]

[1]CEMAT and Instituto Gulbenkian de Ciência, Lisboa, Portugal; ebfernandes@fc.ul.pt
[2]CEMAT and Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal; apacheco@math.ist.utl.pt
[3]Instituto Gulbenkian de Ciência, Lisboa, Portugal; cpenha@igc.gulbenkian.pt

The composite interval mapping, CIM, combines interval mapping (IM) with multiple-marker regression analysis, which controls the effects of quantitative trait locus (QTL) in other intervals or chromosomes onto the QTL that is being tested and thus increases the precision of QTL detection. This approach makes use of the assumption that the quantitative phenotype follows a normal distribution. Many phenotypes of interest, however, follow a highly skewed distribution, and in these cases the false detection of a major locus effect may occur. An interesting alternative is to consider a skew-normal mixture model in CIM, and the resulting method is here denoted as skew-normal CIM. This method, which is similar to CIM, assumes that the quantitative phenotype follows a skew-normal distribution for each QTL genotype. The maximum likelihood estimates of parameters of the skew-normal distribution are obtained by the expectation-maximization (EM) algorithm. The proposed model is illustrated with real data from an intercross experiment that shows a significant departure from the normality assumption. The performance of the skew-normal IM is assessed via stochastic simulation. The results indicate that the skew-normal IM has higher power for QTL detection and better precision of QTL location as compared to CIM.

Canonical Correlation Analysis

Effects of Data Categorization on Canonical Correlation Analysis: A Simulation Study

Matjaž Širca[1], *Špela Jezernik*[2] *and Boris Cergol*[3]

Postgraduate students of Statistics, University of Ljubljana, Ljubljana, Slovenia
[1]Nova Ljubljanska Banka d. d., Postojna, Slovenia; matjaz.sirca@nlb.si
[2]XLAB d. o. o., Velenje, Slovenia; spela.jezernik@imfm.si
[3]Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia;
boris.cergol@imfm.si

In this article we investigate the effects of data categorization on results of canonical correlation analysis. Canonical correlation analysis (CCA) is a general multivariate statistical method for investigating the correlation between two sets of variables. We begin by estimating the correlation matrix of five variables taken from the European Social Survey (2006-2). We then use this matrix to simulate the data from a five-dimensional normal distribution. We generate a 100 samples, each consisting of a 1000 units. Simulated data is then categorized in two different ways, namely into categories of the same width (equidistant) and into categories of different width. In the second case we decide to create categories by using a quantile cut, which means that the resulting categories contain approximately the same number of units. Both types of categorization are investigated using the following numbers of categories: 11, 9, 7, 5 and 3. The comparison of results is based on means and standard deviations of several CCA parameters, calculated for each sample. In addition to all the different categorizations, we also consider the case of non-categorised data. We observe the changes in means and standard deviations for coefficients of canonical correlation, canonical weights and structural weights. We show that categorization significantly affects the estimates and the variability of the estimated parameters, especially in the case of 3 categories. However, all CCA parameters are not equally affected.

The Impact of Missing Values on the Stability of Canonical Correlation Analysis Results: Simulations Using ESS Data

Rok Platinovšek[1], *Rok Okorn*[2] *and Nejc Berzelak*[3]

Postgraduate students of Statistics, University of Ljubljana, Ljubljana, Slovenia
[1]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
rok.platinovsek@fdv.uni-lj.si
[2]Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia; rok.okorn@imfm.si
[3]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
nejc.berzelak@fdv.uni-lj.si

Canonical correlation analysis represents the most general form of linear model and thus serves as the basis for a range of various commonly used multivariate methods. The instability of results between samples, however, has been noted as one of its key drawbacks. This observation also suggests that the method might be sensitive to missing values. In contrast to linear models, the impact of missing values on canonical correlation analysis has not received much attention in the statistical literature. The authors attempt to address this issue through a simulation study. A progressively increasing proportion of missing values is introduced (simulated) into a European Social Survey dataset. The missing values of three different types – MCAR, MAR and NMAR – are treated with four different methods: listwise deletion, mean imputation, nearest neighbor imputation and the EM algorithm. Canonical correlation analysis is performed on each sample and the results are compared in an attempt to evaluate the relative effectiveness of the various imputation techniques under different missingness conditions. In addition, the inter-sample stability of the canonical loadings is compared to the stability of the canonical cross-loadings. The results of the simulation study partially confirm the hypothesis that the canonical cross-loadings represent a more stable alternative to the canonical loadings.

Canonical Correlation Analysis: Influence of Different Methods on Teatments of Missing Values on Simulated Data

Nina Klenovšek[1], *Aleš Toman*[2], *Sebastjan Kocar*[3] *and Romana Štokelj* [4]

Postgraduate students of Statistics, University of Ljubljana, Ljubljana, Slovenia
[1]Fakulteta za matematiko in fiziko, University of Ljubljana, Ljubljana, Slovenia;
ninaklenovsek@yahoo.com
[2]Fakulteta za matematiko in fiziko, University of Ljubljana, Ljubljana, Slovenia;
ales.toman@imfm.si
[3]sebastian_kocar@hotmail.com
[4]Inštitut za varovanje zdravja, Ljubljana, Slovenia; romana.stokelj@ivz-rs.si

So far there has been little investigation into the impact of missing data on the results of canonical correlation. The authors focus on the evaluation parameters of the multivariate normal distribution when the data are incomplete. There are only a few authors in the world who work on canonical corelation. Therefore, we explored the effect of different approaches for the generation and treatment of missing values on the results of canonical correlation analysis. Of the various methods of imputation of missing data, we selected the following four and compared them with each other: 1) deletion of units with missing data, 2) installing averages of variables, 3) insertion of random values on the distribution of variables and 4) multiple imputation. From a total population of 1414 individuals, we took samples of 400 random individuals and repeated this sampling 1000 times. This was then repeated with selected percentages of simulated data replased with one of the four methods (5%, 10%, 15%, 20%, 25% or 30% missing or replaced with avrages, etc.) We then performed canonical correlation analysis. The canonical corelation analyses of the sampling with no missing data was then compared to the canonical corelation analyses of samples when missing data had been simulated. We compared the following results: the value of canonical correlations and their standard errors, sample distribution (in 20% of imputed values), the structural weights for dependent variables and their standard errors and the structural weights of the independent variables and their standard errors. When comparing different methods of imputation, we found that methods did not differ significantly. We found the least destructure change to the data came with the multiple imputation method. Multiple imputation method best preserved the structure of the distribution throughout the sample.

Biostatistics and Bioinformatics II

Changes in Forest Species Composition in the Last 50 Years on the Snežnik Area in Slovenia Studied by Compositional Data Approach

Damijana Kastelec[1], *Milan Kobal*[2] *and Klemen Eler*[3]

[1]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia;
damijana.kastelec@bf.uni-lj.si
[2]Slovenian Forestry Institute, Ljubljana, Slovenia; milan.kobal@gozdis.si
[3]Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia;
klemen.eler@bf.uni-lj.si

Understanding the long-term shifts of forest species compositions is important both for the theory of forest dynamics and also forest management. Only the understanding of complex relations between forest ecosystem and environmental factors can give us some guidelines of proper, i.e. sustainable forest management. In this study we used Slovenian forest inventory data of the last fifty years (1954-2004) of the Snežnik area (several thousands of ha). The area is divided into forest sections each with more or less homogeneous forest structure, which is a consequence of both natural factors and human influences. The basal area (in m2 per hectare) of three tree species (beech, fir and spruce) is given for altogether 111 forest sections for the years 1954 and 2004. The data represent three component subcompositions of forest species. Additionally, some environmental (natural and antropogeneous) factors are known for each forest section: mean altitude, slope, aspect, terrain curvature and logging intensity. The objectives of the study are: (1) to evaluate the shifts in species composition in the last 50 years and link them with the environment, and (2) to find groups of forest sections with similar forest dynamics, which is important in forest management planning. The data were transformed in the sense of analysis of compositional data revealed by Aitchinson (1982), Pawlowsky-Glahn and Egozcue (2001) and Pawlowsky-Glahn (2003) and then multivariate statistical method as cluster analysis and linear model were used on compositional coordinates to get the answers. The differences between relative and absolute statistical analysis of compositions are presented.

Subgroup Discovery in Data Sets with Multi-Dimensional Responses: A Method and a Case Study in Traumatology

Lan Umek[1], *Blaž Zupan*[1], *Marko Toplak*[1], *Annie Morin*[2], *Jean-Hugues Chauchat*[3], *Gregor Makovec*[4] *and Dragica Smrke*[4]

[1]Faculty of Computer and Information Sciences, University of Ljubljana, Ljubljana, Slovenia;
[2]IRISA, Universite de Rennes, Rennes, France
[3]Universite de Lyon, ERIC-Lyon 2, Lyon, France
[4]Dept. of Traumatology, University Clinical Centre, Ljubljana, Slovenia
lan.umek@fri.uni-lj.si

Biomedical experimental data sets may often include many features both at input (description of cases, treatments, or experimental parameters) and output (outcome description). State-of-the-art data mining techniques can deal with such data, but would consider only one output feature at the time, disregarding any dependencies among them. In the paper, we propose the technique that can treat many output features simultaneously, aiming at finding subgroups of cases that are similar both in input and output space. The method is based on k-medoids clustering and analysis of contingency tables, and reports on case subgroups with significant dependency in input and output space. We have used this technique in explorative analysis of clinical data on femoral neck fractures. The subgroups discovered in our study were considered meaningful by the participating domain expert, and sparked a number of ideas for hypothesis to be further experimentally tested.

Class Imbalance Problem for Multivariate Classifiers Using High-Dimensional Data

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

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

One of the possible goals of high-throughput experiments is to develop a multivariate classifier to predict the class of the samples from the measurements (features) derived from the experiment. Data sets that are used to develop or test the classifiers can be imbalanced, i.e., the number of samples in each class is not necessarily the same and this can negatively affect the performance of the classifiers. To our knowledge no systematic study was conducted to evaluate the performance of different classifiers for imbalanced high-dimensional data sets. We considered some of the most commonly used classifiers (k-nearest neighbors, linear and quadratic diagonal discriminant analysis (LDDA and QDDA), random forests (RF), support vector machines, nearest shrunken centroid (PAM) and penalized logistic regression) and conducted a simulation study to assess their predictive accuracy (PA, overall and class-specific) varying the proportion of samples from each class in training and in test set. We considered different magnitudes of between-class differences, different normalization methods (sample or feature centering) and possible solutions to improve the performance of the classifiers for imbalanced data sets (downsizing, oversampling and "vote-weighting"). In brief, simulations results showed that, as expected, the PA of classifiers decreased as imbalance in training set increased and that the decrease of PA was smaller when the differences between classes were larger; LDDA was the most insensitive to class imbalance problem, feature centering worsened the PA for all classifiers while downsizing proved to be beneficial.

Network Analysis

Online Dictionary of the Social Sciences as a Network

Anja Žnidaršič

Faculty of Organizational Sciences, University of Maribor, Kranj, Slovenia;
anja.znidarsic@fov.uni-mb.si

Every dictionary can be presented as a network where terms are nodes of a network. A relation between two nodes in a network is present if there is a second term mentioned in the first term's description. The Online Dictionary of the Social Sciences has approximately one thousand entries covering the disciplines of sociology, criminology, political science and women's study. The dictionary already includes links to other terms and also offers other possible ways of connecting several terms. I will present comparison between the basic and the extended network and some interesting parts of the network. For example the longest path of researching dictionary terms, terms with the largest number of other terms in it's own description, the most easily accessed terms, important nodes (for example hub and authorities) and cohesive groups like cores and islands.

Dynamics of Scientific Co-authorship Networks of Slovenian Researchers

Luka Kronegger

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
luka.kronegger@fdv.uni-lj.si

The complex phenomenon of modern research practices, has been studied from the perspective of the individual scientists, research institutions and national and cross-national research policies. For the science system (including the social sciences and humanities) scientific collaboration presents the major interaction mechanism between actors both at the micro-level of individual scientists as well at the macro level of national science systems. Given that the nature of interaction in complex systems is decisive for the emergence of innovations it is not surprising that the study of collaboration has attracted a lot of attention in the last decades. From this perspective, the cooperation among researchers has in recent years become a typical subject of studies focusing on questions regarding ways of cooperation measurement in general and in various special forms, such as co-authorship in writing articles, monograph publications, etc.
The presented research is based on dataset of co-authorship networks of Slovenian researchers. The time frame used for acquiring analysed data spans the period from 1986 until 2006. The work is currently at exploratory and descriptive stage of analysis and visualization of co-authorship networks through time. For this we are using methods and models, implemented in the PAJEK program (Batagelj and Mrvar, 2003; de Nooy, Mrvar, and Batagelj, 2005; Doreian, Batagelj, and Ferligoj, 2005). We are also applying a stochastic actor oriented model approach implemented in the program SIENA (Snijders, 1996; 2001; 2005; Steglich et al., 2006) and statistical modeling of co-evolution of networks and behaviour (background variables linked to attributes) through time. With such models we are explaining the changes in modelled network (emergence of deletion of ties) by local structural features of the network, by the attributes of actors involved in the tie (constant or changing through time), by the interaction with other networks (or similarities between actors), and explain the changes in attributes of the actors that are controlled by other attributes and network variables.

Clustering of Discrete Distributions: New R Package and Application on US Patent 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 and Physics, University of Ljubljana, Ljubljana, Slovenia;
vladimir.batagelj@fmf.uni-lj.si

Often in clustering of a large number of units a non-hierarchical method is combined with a hierarchical clustering method. The non-hierarchical method allows clustering of large number of units and the hierarchical clustering method builds dendrogram from the obtained nonhierarchical clusters that makes the task of determining the most 'natural' final clustering(s) much easier. Usually compatible (based on the same criterion function) clustering methods are used.
Standard k-means and Ward's clustering methods are both based on the squared Euclidean distance as the error function. In clustering of discrete distributions it turns out that it favors steep unimodal patterns of distributions. Therefore clusterings obtained with these methods usually do not give the 'expected' results.
To get better results we developed adapted leaders methods and compatible agglomerative hierarchical clustering methods using several alternative error functions. Proposed error functions are based on relative error measures between clustered units and a cluster representative - leader, which needs not be defined in the same space.
In this work we present an implementation, as an R package, of the adapted methods from the paper Kejzar et al.: Clustering of discrete distributions: A case of patent citations (submitted). To compare the proposed methods we apply them on citation distributions based on the US patents data set from 1980 to 1999.

"Please Name the First Two People You Would Ask for Help": The Effect of the Limitation of the Number of Alters on Network Composition

Tina Kogovšek[1] *and Valentina Hlebec*[2]

[1]Faculty of Arts and Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
tina.kogovsek@guest.arnes.si
[2]Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia;
valentina.hlebec@fdv.uni-lj.si

Social network indicators (e.g., network size, network structure, network composition) and the quality of their measurement may be affected by different factors such as measurement method, type of social support, the limitation of the number of alters, context of the questionnaire, question wording, personal characteristics such as age, gender or personality traits and others.
In this paper we are focusing on the effect of the limitation of the number of alters on network composition indicators (e.g., percentage of kin, friends etc.), which are often used in substantive studies on social support. Often social networks are only one of the many topics measured in such large studies, therefore the limitation of the number of alters that can be named is often used directly (e.g., International Ssocial Survey Programme) or indirectly (e.g., General Social Survey) in the network items.
The analysis was done on several comparable data sets from different years. Data were collected by the name generator approach by students of University of Ljubljana as part of different social science methodology courses. Network composition on the basis of direct use (i.e., already in the question wording) of the limitation of the number of alters is compared to the composition on the full network data (collected without any limitations) and indirect use of the limitation (i.e., limiting the number of alters in the analysis, but not in question wording).

Biostatistics and Bionformatics III

Susceptibility to Environment of Barley DH Lines Examined in a Series of Trials

Tadeusz Adamski [1], *Zygmunt Kaczmarek*[2], *Stanisław Mejza*[3] *and Maria Surma*[4]

[1]Institute of Plant Genetics, Poznań, Poland; tada@igr.poznan.pl
[2]Institute of Plant Genetics, Poznań, Poland; zkac@igr.poznan.pl
[3]Poznań University of Life Sciences, Poznań, Poland; smejza@up.poznan.pl
[4]Institute of Plant Genetics, Poznań, Poland; msur@igr.poznan.pl

Infection of plants by pathogens is a biotic environmental stress. Barley plants are infected, among others, by Fusarium culmorum - a pathogen affecting seedling, head, root and stem. The infection can result in reduction of yield and grain quality. The aim of the studies was to compare the reaction of barley doubled haploids (DH) inoculated with F. culmorum and non-inoculated to various environments. Thirty two genotypes were inoculated with an isolate of F. culmorum. Experiment was carried out over six years. Spike infection score, kernel weight per spike, 1000-kernel weight and percentage of plump kernels were examined in control and inoculated plants. Genotype-by-environment interaction (GE) and its structure with reference to the environments and genotypes was analysed Additional information about the sensitivity of healthy and infected genotypes to environments was determined by the regression analysis. Statistical computation was made using the SERGEN software. Lines were considered as unstable when their GE interaction was significant at P=0.05. Unstable genotypes were classified as intensive or extensive according to the results of regression analysis. It was found that infected plants were more susceptible to environments. Comparison of classification of healthy and infected lines based on their main effects and GE interaction permitted to select lines of less susceptible both to biotic and abiotic stresses.

A Multivariate Approach to the Evaluation of Oilseed Rape Doubled Haploids on the Basis of a Series of Line x Tester Experiments

Zygmunt Kaczmarek[1], *Stanisław Mejza*[2], *Elzbieta Adamska*[3], *Teresa Cegielska-Taras*[4] *and Laura Szała*[5]

[1]Institute of Plant Genetics PAS, Poznań, Poland; zkac@igr.poznan.pl
[2]Poznań University of Life Sciences, Poznań, Poland; Mejza@up.poznan.pl
[3]Institute of Plant Genetics PAS, Poznań, Poland; eada@igr.poznan.pl
[4]Plant Breeding and Acclimatization Institute, Poznań, Poland; tceg@nico.ihar.poznan.pl
[5]Plant Breeding and Acclimatization Institute, Poznań, Poland; lsal@nico.ihar.poznan.pl

A multivariate statistical approach was proposed to estimation and testing genetic and breeding parameters of winter oilseed rape genotypes. The statistical methods were used to evaluate parental forms (DH lines) on the basis of observations of progenies from line x tester mating system. The aim of the study was the classification and choice the best oilseed rape genotypes with regard to seed yield and three fatty acids: oleic (C18:1), linoleic (C18:2) and linolenic (C18:3). A series of experiments with the set of (7 x 4 =28) progenies was carried out over 3 years. The presented approach involves the use of MANOVA and other multivariate techniques for estimating parameters and testing hypotheses of interest for breeders. These methods allowed estimation of the mean effects of general (GCA) and specific (SCA) combining abilities of DH lines and also their interactions with environments.

Comparison of Ozone-Caused Tobacco Leaf Injury Degree Between Rural and Urban Exposure Sites

Anna Budka[1], *Klaudia Borowiak*[2], *Dariusz Kayzer*[3] *and Janina Zbierska*[4]

[1]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; budka@up.poznan.pl
[2]Department of Ecology and Environmental Protection, Poznań University of Life Sciences, Poznań, Poland; klaudine@up.poznan.pl
[3]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; dkayzer@up.poznan.pl
[4]Department of Ecology and Environmental Protection, Poznań University of Life Sciences, Poznań, Poland; jzbier@up.poznan.pl

Tropospheric ozone is one of the most phytotoxic air pollutant. Its concentration has doubled over the last few decades. Ozone may cause visible injuries on plants and biomass losses, which is especially dangerous for crop plants. Bioindicator plants have been used for determination of the tropospheric ozone level. Tobacco plants reveal visible injuries in the occurrence of ozone in the ambient air. Two tobacco cultivars were used in presented studies – sensitive and resistant for tropospheric ozone. Experiment was performed at two exposure sites – rural and urban – at Poznan city and surroundings area. Plants were exposed for two weeks series in 2003-2006 from the middle of June till the beginning of September. The results obtained have been presented with the use multivariate analysis of variance. The aim of this paper was the comparison of visible leaf injury degree in individual exposure series at two chosen sites.

Analysis of Hybrids Obtained by Diallel Crossing Scheme Based on Categorical Trait

Anita Biszof[1] *and Stanisław Mejza*[2]

Poznań University of Life Sciences, Poznań, Poland
[1]biszof@up.poznan.pl
[2]smejza@up.poznan.pl

The paper deals with experiment performed in breeding programs breeders. We wish to compare the performance of inbred lines and in particular which crosses would be the most profitable. Usually the comparison is based on continuous traits. In this case some genetical characteristics as general combining ability, specific combining ability, reciprocal effect and heterosis effect are very useful in the selection process. Those characteristics are well defined in the continuous trait case. In the paper we propose some methods being analogue to that of continuous trait case. In particular, we propose the way of comparing the hybrids obtained by diallel crossing system on the basis of the categorical trait distribution. Additionally, using simultaneous test procedure we can find lines having similar distribution. The theoretical considerations are illustrated with an example on winter oil seed rape.

Sampling Techniques and Data Collection

Optimal Compression of Statistical Data by Minimization of Information Cost Function

Igor Grabec

Amanova doo, Technology Park, Ljubljana, Slovenia; igor.grabec@fs.uni-lj.si

Experimental characterization of complex physical phenomena by a system comprised of a set of sensors and memory cells is treated. The joint probability density function (PDF) of signals from sensors is utilized as a basic statistical tool for characterization. It can be experimentally estimated by samples of data over kernel estimator. Since the number of experimental samples can increase without limit, while the number of memory cells in a measurement system is generally limited, a question appears how to avoid this discrepancy. For this purpose we introduce a model comprised of representative data and probabilities related to them. In order to develop an algorithm for adaptation of these data to the observed phenomenon we employ the information entropy at the definition of the discrepancy between the PDFs stemming from the experiment and model as well as at the definition of the representative data redundancy. By the sum of discrepancy and redundancy the information cost function of the model is defined. An optimal model can be found by looking for the minimum of the information cost function. In the article an iterative method is proposed for this purpose. During application of this method a new experimental datum can cause either just an adaptation of existing probabilities, or also a creation of a new prototype datum. The action that yields a lower information cost function is considered as a proper one. It is characteristic that resulting model quite normally contains much lower number of prototype data as is the number of experimental samples. Consequently the proposed method can be used to properly compress overwhelming experimental data provided by automatic data-acquisition systems. In the presentation the method is demonstrated on the compression of traffic data from the Slovenian roads network. In this example the traffic rate measured at some characteristic point during a particular day is described by a vector of 24 components. The set of 365 measured vectors from a year is after adaptation properly represented by just 4 representative vectors and related probabilities. The representative vectors approximately correspond to normal working days and days comprising weekends or holidays, while the related probabilities correspond to the relative frequencies of these days.

Performing Descriptive Statistical Analysis on Large Univariate Numerical Datasets Using Spreadsheets

Ilija S. Hristoski

Faculty of Economics, Prilep, Republic of Macedonia; i.hristoski@t-home.mk

The necessity for acquisition and analysis of large numerical datasets is growing lately. This is a result of ever increasing demands for handling vast amount of data, collected or provided either by instrument measuring, experiments, log data analysis, surveys, polling, historical records, observations or artificially generated by algorithms and applied both in technical and social sciences, as well. The large numerical datasets turn into a basis for further scientific examination, like knowledge discovery and data mining techniques, principal data analysis, clustering, regression analysis etc. One of the basic statistical analyses performed on univariate numerical datasets is the descriptive analysis, often utilized in various fields of economics. The paper explores the possibilities for efficient utilization of Microsoft Excel spreadsheets for storing large numerical datasets and performing descriptive statistical analysis on the univariate data contained within spreadsheet cells by implementing basic algorithms as VBA program code.

Invited Lecture

Creating Structured and Flexible Models: Some Open Problems

Andrew Gelman

Coumbia University, New York, U.S.A.; gelman@stat.columbia.edu

A challenge in statistics is to construct models that are structured enough to be able to learn from data but not be so strong as to overwhelm the data. We introduce the concept of ”weakly informative priors” which contain important information but less than may be available for the given problem at hand. We also discuss some related problems in developing general models for taxonomies and deep interactions. We consider how these ideas apply to problems in social science and public health.

Bayesian Statistics

Building a Cross-Classified Multilevel Structure for Credit Scorecard: A Bayesian Monte Carlo Markov Chain Approach

Alesia S. Khudnitskaya

Ruhr Graduate School in Economics, Technische Universität Dortmund, Dortmund, Germany;
khudnitskaya@statistik.uni-dortmund.de

The paper discusses the implementation of a cross-classified multilevel model to setting up a credit scorecard for forecasting probability of default. The individual applicants for a loan are cross-classified by their living environments (micro -socio-environment), occupational fields (occupation) and infrastructure of shopping facilities in the area of their residence (infrastructure). These are higher level categories that represent clustering of borrowers according to similarities in particular characteristics. This cross-classified nesting helps to model exposure to specific risk factors and category hazards that trigger default and are essential for more accurate credit worthiness assessment. The scoring model is built in two steps. First, hierarchical clustering is done and applicants for a loan are grouped into clusters within each of the three categories: microenvironment, occupation and infrastructure. On the second step, these nested structures are combined to build a non-hierarchical multilevel model with random microenvironment, occupation and infrastructure-specific effects. The Bayesian approach with Monte Carlo Markov Chain is appealing as we simulate random effects of particular microenvironment or profession and are interested in making inference of the parameters in the population of different microenvironments or population of various occupation fields. Combining prior information on the model random effects and the data conditional posterior distribution is drawn and point and interval estimates for the main parameters are calculated. The credit scoring model is fitted using WinBugs1.4.3 and WLwin2.11 with STATA10.1 generated initial values for the random, fixed effects and variance-covariance matrices. Three chains in parallel are run with chain-length equal 500.000 iterations.

Bayesian Model Averaging for Hierarchical Log-Linear Models and an Application to the Dumping Severity Data

Haydar Demirhan[1] *and Canan Hamurkaroglu*[2]

Department of Statistics, Hacettepe University, Ankara, Turkey
[1]haydarde@hacettepe.edu.tr
[2]caca@hacettepe.edu.tr

Many fields of scientific investigation include analysis of categorical data. Log-linear models are widely used to discover association structure of considered categorical variables. Each log-linear model corresponds to an association structure. Choosing the log-linear model that fits best the data is an important step in the log-linear analysis of categorical data. There is a huge number of approaches for model selection in the classical and Bayesian settings. Almost each method has its own problems. Most important problem is on the model uncertainty. When a single model is selected and inferences are conditionally based on the selected model, model uncertainty is ignored. This case is especially seen in the classical setting. This difficulty can be overcome by using Bayesian model averaging (BMA) for model determination and parameter estimation, simultaneously. In the BMA approach, posterior distribution of considered quantity is obtained over the set of suitable models, and then they are weighted by their posterior model probabilities. Model uncertainty is included in the analysis by this way.
We consider a BMA approach for hierarchical log-linear models; propose an approach for the calculation of posterior model probabilities; and apply the approach to the Dumping severity dataset. A wide range of log-linear models fits this dataset well when analyzed with classical approaches. However, only one log-linear model is determined to fit the dataset well by the BMA approach. Therefore, use of the BMA approach for this dataset is found to be more advantageous than classical model selection approaches.

Endogeneity of Store Attributes in Heterogeneous Store-Level Sales Response Models

Harald Hruschka

University of Regensburg, Regensburg, Germany;
harald.hruschka@wiwi.uni-regensburg.de

Retailing firms as a rule decide on store attributes (e.g., store size) considering an assessment of future sales of these stores. Typically, managers allocate better or more equipment to stores for which they expect higher sales. Models which ignore the fact that this behavior leads to endogeneity overestimate effects of these attributes. We consider potential endogeneity of store attributes in the sales response function by an instrumental variable approach. We also allow for heterogeneity across stores by assuming that store-level coefficients are generated by a finite mixture distribution. Models are estimated by a MCMC technique which combines two Gibbs sampling algorithms. In the empirical study both heterogeneity and endogeneity turn out to influence estimates. For a cross section of more than 1,000 gas stations credible intervals of differences of coefficients between models ignoring and models considering endogeneity indicate that models which ignore endogeneity overestimate the effects of two store attributes on sales.

Statistical Applications - Economics I

Advertising Expenditure Components in Time; A Case Study of 17 European Countries in the Period 1994-2007

Katarina Košmelj[1] *and Vesna Žabkar*[2]

University of Ljubljana, Ljubljana, Slovenia
[1]katarina.kosmelj@bf.uni-lj.si
[2]vesna.zabkar@ef.uni-lj.si

We analyze the advertising expenditure (ADSPEND) components for 17 European countries in the period 1994-2007. The components under study are: Electronic, which summarizes radio and television; Print, which includes press and outdoor; and Online, a new medium, which evolved within the period under study and is supported by the Internet. Our objective is to gain a deeper insight into how the ADSPEND components restructured after the Online component evolved. For which countries is an increase in Online made on the account of Print, on the account of Electronic or on the account of both?
The dataset consists of 17 compositional time series. For each country, we have two components (Print and Electronic) in start. The Online component did not emerge simultaneously in all the countries. The first country with a reported value for the Online in the Euromonitor database was Finland in 1996, followed by France, Great Britain, and Sweden in 1997, these Online values were very low, however increased up to 15 percent in 2007. For some countries, the Online component remained near zero up to the end of the observed period. Our statistical analysis is focused on different statistical approaches to identify clusters of similar countries.

Investigating the Relationship Between Electricity Consumption, Export and GDP in Turkey via the Multivariate Time Series Analysis

Derya Ersel[1], *Yasemin Kayhan Atilgan*[2] *and Süleyman Günay*[3]

Hacettepe University, Ankara, Turkey
[1]dtektas@hacettepe.edu.tr
[2]ykayhan@hacettepe.edu.tr
[3]sgunay@hacettepe.edu.tr

In this study, the relationship between economic growth, electricity consumption and export in Turkey is investigated over the period 1970-2007. Unit root test revals that all series after logarithmic transformation are non-stationary. GDP and eletricity consumption are integrated of order one and export is integrated of order two. Granger causality analysis indicates that there is a two way relationship between export and GDP. Besides, electricity consumption is a Granger cause of export and GDP. According to cointegration analysis, there is a long-run relationship between electricity cosumption and GDP. Also, variance decomposition results show that GDP is an important variable to explain the electricity consumption.

Nonlinearities, Weak Integration and the Globalization of Stock Markets

Rui C. Menezes

ISCTE-IUL, Lisbon, Portugal; rui.menezes@iscte.pt

This paper analyzes the stock market globalization on the basis of the price theory and the Law of One Price. This leads to the concept of market integration which, under the nonstationarity of price level variables, can be empirically tested by using cointegration techniques. An error correction model incorporating prices and returns is specified and empirically tested. The results show that the stock markets under analysis are cointegrated and there is just one cointegrating vector that rules the long-run relationship between these markets. The Portuguese and Spanish stock markets are weakly exogenous while the remaining are endogenous in the context of our system. Market integration in the sense of the Law of One Price holds for the system as a whole but full price transmission is only accepted for the pairs Portugal-UK, UK-Japan and Japan-US, for which one can say that the Law of One Price also holds. The results, therefore, show that price movements between pairwise stock markets are highly nonlinear and complex.

Parametric Estimation of the Customer Churn Risk

Sofia L. Portela[1] *and Rui C. Menezes*[2]

ISCTE - IUL, Lisbon, Portugal
[1]slportela@iscte.pt
[2]rui.menezes@iscte.pt

Profits from customer relationships are the lifeblood of firms. So, customer churn can be very damaging to the business performance. As such, researchers have recognised the importance of an in-depth study of customer defection in different industries and geographic locations. This study aims to understand and predict customer lifetime in a contractual setting in order to improve the practice of customer portfolio management. A duration model is developed to understand and predict the residential customer churn in the fixed telecommunications industry in Portugal. The models are developed by using large-scale data from an internal database of a Portuguese company which presents bundled offers of ADSL, fixed line telephone, pay-TV and home-video. The model is estimated with a large number of covariates, which includes customer's basic information, demographics, churn flag, customer historical information about usage, billing, subscription, credit, and other. The results of this study are very useful to the computation of the customer lifetime value.

Social Science Methodology

Ordered Dissimilaritiy of Students Gradings Distributions – Ordinal Statistics to Answer Ordinal Questions

Matevž Bren[1] *and Darko Zupanc*[2]

[1]Institute of Mathematic, Phisics and Mechanics, Ljubljana, Slovenia;
matevz.bren@fov.uni-mb.si
[2]National Examination Center, Ljubljana, Slovenia; darko.zupanc@guest.arnes.si

Educational measurements in grades fall into ordinal scale. In comparing students 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 grades distributions differ and the second: to what extend this is true.
In our contribution we will present ordinal statistics to answer these ordinal questions. For the first: standard MWW test and for the second: ordered dissimilarity $d = P(X > Y) - P(X < Y)$. In demonstration we will apply real data on students gradings in upper secondary education in Slovenia. We will also discuss properties of the ordered dissimilarity d.
We intend to apply this ordinal statistics in the ALA Tool i.e. Assessment of/for Learning Analytic Tool for gathering and analysing of grades in upper secondary education in Slovenia.

The Cultural Capital of Immigrant Families and the Impact on Student Performance

Alina S. Botezat

"Al. I. Cuza" University of Iasi, Romanian Academy - "Gh. Zane" Institute of Economic and Social Research, Iasi, Romania; simonaiurea@yahoo.com

The aim of this paper is twofold. It intends to show how strongly student performance depends on student cultural background as well as to explain how the intergenerational transfer of cultural capital affects academic performance. Using PISA data 2003 for Germany we provide a detailed econometric analysis of the education production function of German and immigrant students from the first and second generation. The results imply that the education of parents has a great significant impact on the student performance only in the case of German students. In the case of immigrant children the only reason which negatively affects their educational achievements is a low level education of the mother. The possession of cultural resources is a strong predictor in the case of student performance only for German students. Logistic regressions show that the intergenerational upward mobility is higher for immigrants than for German students. Based on quantile regressions, we interpret the impact of the cultural capital on the gender gap. Our findings suggest that male students from the second generation of immigrants score considerably higher in Mathematics than females, especially in the upper part of the distribution. Females perform better only in the lower quantiles of the reading literacy test. It was also proved that the socio-economic status and the cultural capital of the student are more correlated to the school type than to school performance.

Changes in the Marital Structure of the Population of Vojvodina in Respect to their National-Ethnic and Confessional Affiliation

Katarina J. Čobanović[1] *and Valentina T. Sokolovska*[2]

[1]Agricultural Faculty, University of Novi Sad, Novi Sad, Serbia; katcob@polj.ns.ac.yu
[2]Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia; valentinas@neobee.net

Generally speaking, the changes of structure and values of the modern society are reflected in the sphere of parenthood, i.e. marriage as a social and biological category. The particularities of ethnic and confessional structure of Vojvodinian population originate from specific historical development of this region. Within the existing social and spatial context, Vojvodinian population has been characterized by a certain level of multiculturalism and their capability to integrate as well as their mutual understanding. This paper examines the tendencies regarding changes in the amounts of concluded and divorced marriages in respect to the population's national-ethnic and confessional affiliations. The changes were examined for the period of 35 years (1970-2005). On the basis of the vital statistics data, time series of the numbers of concluded and divorced marriages between people of different national-ethnic and confessional affiliation were analyzed. The method used in the analysis is the moving averages method, as well as fitting the trend model to the original series data. The changes in the numbers of concluded and divorced marriages have been analyzed also with regards to mutual influences and one's national-ethnical affiliations. The paper focuses especially on the influence of one's national-ethnic identity on occurrence of homogamous and heterogamous marriages. Marital characteristics of certain combinations of marriage partners have been compared regarding the population's national-ethnic and confessional affiliation. The research on the structural characteristics of marriages in Vojvodina in respect to the ethnic and religious affiliation for the period of 35 years was intended to determine the nature and the intensity of the changes which have emerged in the second half of XX and at the beginning of XXI century.

Education

Difficulties in Teaching Statistics in Slovenian Secondary Schools

Andreja Drobnič Vidic[1] *and Simona Pustavrh*[2]

[1]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia;
andreja.drobnic@fmf.uni-lj.si
[2]Šolski center Novo mesto, Srednja elektro šola in tehniška gimnazija, Novo mesto, Slovenia;
simona.pustavrh@guest.arnes.si

Mathematics teachers who also teach statistics as a part of mathematics in Slovenian secondary schools have mainly incomplete statistics education. Mathematics' curriculum guidelines and the assessment system do not explicitly point out the need of statistical reasoning and thinking, often seen as the most important elements in statistics education at all levels of statistics teaching. On the other hand, the aims of mathematics curriculum 2008 are very ambitious in terms of statistics meeting interdisciplinary challenges and requiring teachers' cooperation across curriculum subjects in a very limited time. To find out how teachers realized mentioned aims and what problems they were faced with, we posed a questionnaire and analysed the answers. As an example that all the aims could be realized we described a realisation of project called "Energy as a value" with concrete interdisciplinary real statistical problems designed for students at secondary school. These statistical problems connect statistics with other curriculum disciplinary subjects, such as sociology, ecology, and ICT. We focus on difficulties regarding the first year realisation as well as on teachers' efforts and students' feedback.

Mathematics in Slovene General Matura – Discrepancy of Grades at Basic and Higher Level of Achievement

Alenka Hauptman

National Examinations Centre, Ljubljana, Slovenia; alenka.hauptman@ric.si

In General Matura, Mathematics is one of the compulsory subjects and it can be taken either at Basic or Higher Level of Achievement. Basic Level of Achievement is expressed by the classic five-grade scale from 1 to 5. Candidates at Higher Level of Achievement can get grade on scale from 1 to 8. Conversion of points into grades (i.e. how many points are required for each grade) on each Level is set independently, and we tried to find out if the same grade on each Level of Achievement corresponds to the same knowledge.

Both Basic and Higher Level in Mathematics include the same Part1 of the exam. The second part of the exam (Part2) is applied only to Higher Level's candidates. Part1 amounts to 80 % of the total points at Basic Level, and 53.3 % of total points at Higher Level. Those candidates get other 26.7 % of points in Part2. Oral part of the exam represents 20% of the grades at both Levels.

In this presentation we would like to show discrepancy between knowledge within the same grades of candidates at Basic and Higher Level of Achievement on an example of Mathematics exam from 2008 General Matura. Rasch's one-dimensional measurement model will be used to place item difficulties on common scale and suitability of grade conversion on both Basic and Higher Level of Achievement will be explored. The results show interesting differences in knowledge of candidates with the same grade at Basic and Higher Level of Achievement.

Does Examinee Choice in Educational Testing Work? One Example from Slovenian General Matura Examinations

Gašper Cankar

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

Achievement tests sometimes entertain examinee choice – a situation where an examinee is presented with a set of items among which s/he has to choose one (or few) that will be scored in her/his total test score. Basic assumption is that choice items are equivalent regarding both content and psychometric characteristics and therefore doesn't matter which particular item examinee selects. For ease of use choice items often also share same maximum number of points and examinee score on test is usually achieved by summing scores from all items taken by the examinee regardless of their combination. Choice items present conceptual problems like why enabling choice when items should be equivalent in the first place and methodological ones like how item scores from different combinations of items should contribute to comparable total score on test. Since some tests on Slovenian General Matura examination include choice items we will explore these difficulties through a practical case of one of the tests. Choice items are being selected by different groups of examinees taking test which makes it inappropriate to compare difficulties of the items directly and demands the use of appropriate measurement model. Rasch's one-dimensional measurement model will be used to place item difficulties on common scale and violations of basic assumption of item equivalency will be explored.

Invited Lecture

Statistics of Compositional Data

Gerald van den Boogaart

Technical University, Freiberg, Germany; boogaart@math.tu-freiberg.de

The talk gives a short introduction into the recent approach to the statistical analysis of compositional data. Data providing the amounts of different components forming a total is called compositional if the total amount is irrelevant for statistical question under consideration. These might be amounts of different elements in minerals, the amounts of different cell types in a blood samples, the relative amounts of different beetle species in ecological systems, or the money spend on different types of expenses (workforce, tax, operation costs, raw products) in companies. Almost never all relevant components are reported.
Seen from a classical view of multivariate statistics this type of data has a lot of strange properties: it can't be normally distributed because the domain is bounded to a triangle like region, variances matrices are always singular, scaled vectors correspond to the same composition, a single measurement error or missing value changes all other values, relative errors are more relevant than absolute differences, different units of measure (e.g. mass %, vol %) or different unobserved components can lead to different order relations and directions of dependence among the unaffected components, data is practically always heavily skewed.
The talk will introduce you to a solution to that problem: The principle of working in coordinates. This principle allows it to translate compositional problems into classical multivariate statistical tasks, to do a consistent analysis with well known methods and to translate the results back into a compositional context. It will show this principle at work for some classical methods like distributions, linear models, tests, principle component analysis and outlier detection. And it will show how new a specialized methodology can be built on that. The aim is to show how simple it can be to analyze compositional data in a consistent way avoiding all the paradoxes and artefacts mentioned above, when we just follow some basic rules.

Mathematical Statistics

Fuzzy Hypothesis Testing in Linear Regression

Duygu İçen[1] *and Süleyman Günay*[2]

Department of Statistics, Hacettepe University, Ankara, Turkey
[1]duyguicn@hacettepe.edu.tr
[2]sgunay@hacettepe.edu.tr

Uncertainty caused by randomness and the uncertainty caused by fuzziness exists in the analysis and handling of data in almost every area of science. Fuzzy set theory which is a tool of handling with uncertainty is introduced by Zadeh in the paper titled as "Fuzzy Sets" in 1965. Fuzzy set theory aims at modelizing imprecise information which exists in real world situations. With fuzzy regression models observational error is assumed that the gap between the data and the model is an ambiguity in the structure of the system. The aim of this study is to model the system by means of a possibility system called fuzzy regression model with estimating the parameters of model by using Buckley's approach as triangular fuzzy numbers. Then an application of Fuzzy Hypothesis Testing in Linear Regression to a real data set is given as an example.

Incongruence of Model Fit Indices and Other Evidence of Model Quality in SEM

Roman Konarski

University of Gdansk, Gdansk, Poland; roman.konarski@pbsdga.pl

Model evaluation is a complex step in the analysis of structural equation models (SEMs). Several well researched issues in model assessment are: the violation of distributional assumptions, equivalent models, specification searches, and statistical power. Moreover the methodological literature has indicated two further issues that have not been carefully examined. One problem is the possible incongruence between global fit indices and the precision of parameter estimates. For example, it is sometimes observed that a model with an adequate fit accounts for little variance in endogenous variables and possesses insignificant parameter estimates. The opposite phenomenon occurs, when models that fit poorly can explain high proportions of variance in the endogenous variables and/or possess highly significant parameter estimates. The other problem is the sometimes observed incongruence between model fit indices and residuals. For example, a model that fits poorly, as reflected by unacceptable fit indices, may yield small residuals. It has been shown that this incongruence can be expected when the common factors in the model are measured with highly reliable indicators. We demonstrate that the divergence of global fit indices and the precision of parameter estimates and/or the reliability of indicators is explained by a more fundamental issue in SEM analysis. The issue is the relationship between global model fit indices and the strength of the relationship between observed variables in the analyzed covariance matrix. We propose a model assessment framework that is based on considerations of global fit indices, residuals, the precision of parameter estimates, and the reliability of indicators.

Bibliometrics

The Use of Statistical Methods in Library and Information Science Literature

Güleda Düzyol[1] *and Sevil Bacanli*[2]

[1]Department of Information Management, Hacettepe University, Ankara, Turkey;
gduzyol@hacettepe.edu.tr
[2]Department of Statistics, Hacettepe University, Ankara, Turkey; sevil@hacettepe.edu.tr

The aim of this study is to examine the use of statistical methods in Library and Information Science research articles which are published in the journals listed under Information Science & Library Science category of the Social Sciences Citation Index. The highest impact factored journals are selected by using Journal Citation Report data and articles published in these journals between 2000-2008 are analyzed. A subjective classification of statistical methods that summarize main topics of statistics and an ordering of sub-topics in each heading is formed according to the classification schemes of statistical abstracts. By using this classification, the research articles are classified methodologically. The findings are used to explore recent statistical trends in the field of Library and Information Science. In addition, they provide a basis for comparison with a similar analysis carried out for papers published between 1991-1997.

Peer-Reviews and Bibliometrical Methods: Two Sides of the Same Coin?

Primož Južnič[1], *Matjaž Žaucer*[2], *Miro Pušnik*[2], *Tilen Mandelj*[2], *Stojan Pečlin*[3] *and Franci Demšar*[3]

[1]Department of Library and Information Science and Book Studies, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia; Primoz.Juznic@ff.uni-lj.si
[2]Central Technological Library, University of Ljubljana, Ljubljana, Slovenia
[3]Slovenian Research Agency, Ljubljana, Slovenia

Our aim was to investigate, within the context of three different and specific peer review system, whether differences in terms of peer review provision affect the decisions which projects to achieve grants by the Slovenian Research Agency. In particular, the attempt is made to discern whether there is a difference in the scientific quality between different applicants, using bibliometric and scientometric indicators. The methodology employed is that of comparison using three different forms of peer reviews judging applicants and grant holders projects and quality measured by bibliometric indicators. To address the issue of quality, a causal model is proposed that gather the publication and citation data with the ability to attract funding for the research from different sources. In this way the use of bibliometric indicators is considered as post festum to a peer review, as the way how to compare results of different peer review system. All in all, 1.375 projects proposals seeking grants, for three years - 2002, 2005 and 2007 were checked by bibliometric indicators, as provided by SICRIS. 408 applicants were successful and have received grants, which gives the success rate of 33%. We have tried to find by statistical methods if relations between successful application based on peer review and bibliometric indicators diverge, regarding different peer review system used. The results show that peer review system and bibliometric indicators are related. Analysis reveal a relationship between peer review systems based decisions in grants and bibliometric indicators. Nevertheless, we can confirm that, on average, the reviewers indeed selected the slightly better performers from a relatively homogenous group of applicants. However, a comparison between approved and rejected applicants shows that these depend on different peer review systems.

Measuring the Citation Impact of Statistic Journals with Structural Equation Modelling Analysis

Güleda Düzyol[1], *Duygu İçen*[2] *and Süleyman Günay*[3]

[1]Department of Information Managemen, Hacettepe University, Ankara, Turkey;
gduzyol@hacettepe.edu.tr
[2]Department of Statistics, Hacettepe University, Ankara, Turkey;
duyguicn@hacettepe.edu.tr
[3]Department of Statistics, Hacettepe University, Ankara, Turkey; sgunay@hacettepe.edu.tr

The citation impact of a journal reflects its quality, importance and performance. It needs to be known how external factors such as journal characteristics, journal accessibility etc. influence journal citation impact to evaluate journals more comprehensively. The aim of this study is to determine the external factors affecting journal citation impact. For this purpose, Structural Equation Modelling (SEM) is used on the journals from Statistics&Probability subject category of the Science Citation Index. The findings of this study shows that certain external factors significantly influence journal citation impact also suggests that journal citation impact can be predicted by these external factors.

Econometrics I

Testing Tobler's Law in Spatial Panels: A Test for Spatial Dependence Robust Against Common Factors

Giovanni Millo

DiSES, University of Trieste and Generali R&D, Trieste, Italy;
giovanni_millo@generali.com

In the spatial econometrics literature, spatial error dependence is characterized by spatial autoregressive processes, which relate every observation in the cross-section to any other with distance-decaying intensity: i.e., dependence obeys Tobler's First Law of Geography ("everything is related to everything else, but near things are more related than distant things"). In the literature on factor models, on the converse, the degree of correlation between cross-sectional units depends only on factor loadings.
Standard spatial correlation tests have power against both types of dependence, while the economic meaning of the two can be much different; so it may be useful to devise a test for detecting "distance-related" dependence in the presence of a "factor-type" one.
Pesaran's CD is a test for global cross-sectional dependence with good properties. The CD(p) variant only takes into account p-th order neighbouring units to test for local cross-sectional dependence. The pattern of CD(p) as p increases can be informative about the type of dependence in the errors, but the test power changes as new pairs of observations are taken into account.
I propose a bootstrap test based on the values taken by the CD(p) test under permutations of the neighbourhood matrix, i.e. when "resampling the neighbours". I provide Montecarlo evidence of it being able to tell the presence of spatial-type dependence in the errors of a typical spatial panel irrespective of the presence of an unobserved factor structure.

Modeling SBITOP Stock Index Time Series Through Decomposition

Aleša Lotrič Dolinar

Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia;
alesa.lotric.dolinar@ef.uni-lj.si

We compare two different approaches to modeling Slovenian blue chip stock index (SBITOP) time series. On the one hand, we model the SBITOP series per se, while on the other hand we decompose the index and separately model the time series of those most liquid Slovenian stocks that make up the SBITOP.
All the models, the index time series model as well as the composing stocks time series models, are set up by including the nonlinear volatility component, since these time series turn out to be very volatile (like most of the financial time series). That is, we do the modelling with the ARIMA family models that are expanded by the volatility term. This is done in two ways – we include the conventional GARCH term with one and the more recent LAVE (Locally Adaptive Volatility Estimate) method with the other set of models. Then we put the component estimates back together, taking all changes in index structure and all consecutive correction factors into account. Thus we get two different index forecasts – one directly and the other through decomposition, separate component modeling and recomposition. We compare the different modeling approaches from the goodness of fit, as well as from the quality of forecasting point of view (here we consider different error measures). Besides, we also find out, which volatility expansion, GARCH or LAVE, is more efficient in our case.

Design of Experiments

On the Efficiency of Some Incomplete Split-Plot × Split-Block Designs with Control Treatments

Katarzyna Ambroży[1] *and Iwona Mejza*[2]

[1]Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poznań, Poland; ambrozy@up.poznan.pl
[2]Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Poznań, Poland; imejza@up.poznan.pl

In many agricultural experiments a problem of the comparison of the test treatments and control treatments is usually very important. It happens also that experimental material is limited and does not allow to use a complete design. Those problems are considered in the paper dealing with an incomplete split-plot × split-block (SPSB) design for three factor experiments with crossed and nested treatment structures. Two cases of the design are considered, i.e. when its incompleteness is connected with crossed treatment structure only or with nested structure only. It is assumed the factor levels connected with the incompleteness of the design are split into two parts, the first one containing test treatments and the second one - a control treatments. Two different augmented designs are used in the construction method of the SPSB design. They both lead to generally balanced SPSB designs. In a modelling data obtained from such experiments the structure of experimental material and appropriate randomization scheme of the different kinds of units before they enter the experiment are taken into account. With respect to the analysis of the obtained randomization model with seven strata the approach typical to the multistratum experiments with orthogonal block structure is adopted. The proposed statistical analysis of linear model obtained includes estimation of parameters, testing general and particular hypotheses defined by the (basic) treatment contrasts with special reference to the notion of general balance.

Graphic Analysis of Interaction in Full Factorial Designs: A Critical Study

Dulce G. Pereira[1] *and Paulo Infante*[2]

University of Évora, Évora, Portugal
[1]dgsp@uevora.pt
[2]pinfante@uevora.pt

To analyze the simultaneous effect of two or more factors on a response variable, a statistical techniques commonly used is the factorial design. When we have a full factorial experiment with replications, the response variable is influenced by more than one independent variable (factor) and we are interested in studying the effect of individual factors; that may be dependent on the level to which it is the other factor (significant interaction). Is commonly reported in the literature of the area that the effects are independent then, in graphic terms, the lines of the averages for each level of the factors are parallel, indicating that the effect of a factor is the same for different levels of another factor. When the lines cross then this means that there are combinations of factors that produce different effects in the response variable than would be expected if the factors were considered separately. However, there are situations where the lack of parallelism is not indicative of the existence of interaction and where the parallelism of lines is not indicative of absence of interaction (concluding that the interactions are not strong enough). We will analyze in a critical way different examples in this context. Using the illustration of several trying situations assess such situations, and understand statistically.

Statistical Quality Control for Business Indicators of Healthcare Quality: General Considerations and a Specific Proposal

Gaj Vidmar[1] *and Rok Blagus*[2]

[1]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana; and University Rehabilitation Institute of the Republic of Slovenia, Ljubljana, Slovenia; gaj.vidmar@mf.uni-lj.si
[2]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia; rok.blagus@mf.uni-lj.si

Within the QA activities of the Ministry of Health in Slovenia, over 100 business indicators of economy, efficiency and funding allocation are annually monitored in 26 hospitals. The essence of associated statistical analyses is outlier identification, whereby we opted for an exploratory approach using three types of methods: well-known tests for small samples (Grubbs, Dean&Dixon, Nalimov), all based on normality assumption hence applied conditionally upon results of normality tests (Kolmogorov-Smirnov-Lilliefors, Shapiro-Wilk); boxplot rule; and control charts. We focus on indicators which are same-quantity ratios and thus bound between 0 and 1 yet not appropriately treated either as proportions (like success of an operation in related studies) or fixed-denominator ratios (like mortality rates). In funnel-plots, which are nowadays a standard for monitoring hospital/physician performance-indicators, virtually all points would be proclaimed outliers in our data because of huge denominators (thousands of sq.metres, millions of Euro) yielding too-narrow confidence intervals for average proportion. Therefore, we choose Double Square Root (Shewart) Chart plotting square-root of numerator vs. square-root of difference between denominator and numerator. Control limits for such charts are traditionally based on binomial-probability-paper, but for our data such limits (having an underlying assumption of binomial distribution like in funnel-plots) are exceedingly narrow. Hence, we estimated the overall trend using linear regression through origin (since, e.g., no costs can be incurred without income), whereby our proposed innovation is control-limits estimation using 95% confidence-interval for prediction. We studied performance and agreement of the selected methods on real data (i.e., random-denominator ratios with highly-correlated numerator&denominator) and simulated data (from best-fitting distributions, bounded and non-negative), without and with simulated outliers, including the heuristics to label as outlier any datum singled out by any of the applied methods. Adequacy of warning/control/action limits determines success of statistical process control, but adequate visual display of data&analyses is also paramount. Therefore, we conclude with examples of good presentation practices for business indicators of healthcare quality.

Data Mining

Discriminant Analysis Versus Random Forests on Qualitative Data: Contingent Valuation Method Applied to the Seine Estuary Wetlands

Salima Taibi[1] *and Dimitri Laroutis*[2]

Esitpa, Mont saint aignan, France
[1]staibi@esitpa.org
[2]dlaroutis@esitpa.org

The contingent valuation method constitutes an economic method quantifying monetarily the set of values which individuals allot to a given environmental good. At the center of this method we find a questionnaire aiming to reveal the willingness to pay of the individuals for the preservation, in our study, of the seine estuary wetlands. Our objective is to build a model making it possible to be able to predict the CAP. The predictors for the majority are qualitative variables (nominal or ordinal), we used a procedure which allowed to transform them into quantitative variables. We carried out the analysis of the multiple correspondences of the predictors i.e. the analysis of the correspondences of the disjunctive table. The p selected explanatory variables X1, X2 ,..., Xp are replaced by the co-ordinates on q factorial axes (q??p) by weighting allowing to preserve the importance of the components. Two methods of classification were implemented, the discriminating analysis and the method of the random forests. The goal is to compare their performances in terms of classification.

Using Decision Trees for Classification in Actuarial Analysis

Damla Barlas [1] *and Omer Esensoy*[2]

Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey
[1]dbarlas@hacettepe.edu.tr
[2]esensoy@hacettepe.edu.tr

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in miscellaneous ways that are both understandable and useful to the data owner. Statistical Methods are used during data mining process. Data mining can be used for credibility, rate making, reinsurance and risk classification in insurance sector. Actuaries develop risk models by segmenting large populations of policies into predictively accurate risk groups, each with its own distinct risk characteristics(Apte et. All, 1998). Due to the fact that they can decrement mistakes and make confidential accounting in their calculations. In this study we focus on grouping a charitable fund members using decision trees which is commonly used in data mining algorithms. First of all we must arrange our data set. Then we apply Factor Analysis to the data set. Factor Analysis is reduce the number of variables and to detect structure in the relationships between variables, that is to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method. Secondly, we use decision trees for classification. So the data set can be classified homogenous class. We use SPSS CLEMENTINE 11.1 in our study.

Analysis of the eDonkey Data – In Search of Pedophilia

Aleš Žiberna[1], *Vasja Vehovar*[2] *and Matej Kovačič*[3]

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

We will present an analysis of three weeks of log files of an eDonkey server provided by the MAPAP project (http://antipaedo.lip6.fr/). The main aim of the analysis was to identify potentially pedophilic files, IPs (people) and (unknown) words. We have started our analysis by selecting the words that will be used as a criteria for "tagging" IPs and files as potentially pedophilic.
In the next step we identified potentially pedophilic IPs and files using these words. Here we will present some characteristics of the files and IPs appearing in the eDonkey network with an emphasis on those that were identified as potentially pedophilic. One of the most interesting discoveries was that most IPs only appear as as "seekers" or "hosts". As a consequence, the IPs that seek pedophilic files are usually not the same as those that host them.
Based on the words that the potentially pedophilic files and IPs use (in addition to those that identify them as potentially pedophilic) we have identified other words that might indicate pedophilic content and evaluated our identification of potentially pedophilic IPs and files.
We also explored the networks of potentially pedophilic files, IPs and (selected) words. While in this talk we will focus on how files and IPs are connected, we will also present some other networks. In most cases we focused on networks containing only two types of nodes, however we have also explored all of them together.

Supervised Learning for Automatic Target Recognition

Gerard Brunet

IUT-STID, Niort, France; gerard.brunet@univ-poitiers.fr

The aim of this paper is to perform recognition of zones in a sequence of aerial images. User can select two zones of interest in a set of candidate locations after feature extraction and edge detection. Given an image of the scene, a set of parameters are extracted and two sets of potential targets are constituted, corresponding to the two zones selected by the user. Euclidian distance between zones from the two sets and color information are used to solve the correspondence problem. To implement a practical system, concurrent programming uses threads with Java language and semaphores for synchronisation of processes. Real images have been tested to evaluate the performance concerning translation, rotation, zoom, and random noise addition.

Econometrics II

Management Accounting in Austrian Family Enterprises - An Empirical Research

Christine Duller

Johannes Kepler University, Linz, Austria; Christine.Duller@jku.at

In spite of the high economical relevance management accounting deals with the subject family enterprises in empirical research rather little. Based on the hypothesis that family enterprises aim at humane objectives to a greater extent and use instruments of management accounting to a lesser extent than nonfamily enterprises, results of an empirical study for the region Upper Austria are presented. The conclusion is that apart from the extent of the return on equity the objectives of responding family enterprises do not differ much from those of non-family enterprises. Contrariwise a high level of professionalization of applied instruments of management accounting can be identified for family enterprises.
The research method is based on a standardized questionnaire. All enterprises in Upper Austria with 50 or more employees were invited to take part in the survey, the usable return was 20%.
It is accepted that family enterprises aim to achieve a combination of financial and nonfinancial goals. This research will proof the following hypothesis:

- Family enterprises aim at humane objectives to a greater extent than nonfamily enterprises.
- Family enterprises aim at financial objectives to a lesser extent than nonfamily enterprises.
- Family enterprises demand a lesser return of equity than nonfamily enterprises.
- Family enterprises use instruments of management accounting to a lesser extent than nonfamily enterprises (strategic and operative).

Family enterprises are more often than not small enterprises, too. Therefore each hypothesis will be tested with respect to structure and size.

The Attitude of the Government to the Arab Minority in Israel: A Study of Government Fiscal Allotments to Local Authorities

Tal Shahor

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

This research examines government fiscal allotments made to Jewish local authorities in comparison to its fiscal allotments to Arab local authorities during the period 1994 – 2006. Data for this research was taken from publications issued by the Central Bureau of Statistics. This research is of particular importance because in Israel funding for many different areas of activity (education, social welfare, development, etc.) are channeled through local authorities. The first stage of this research compares the average fiscal allotments made by the Government to the Jewish local authorities with the average fiscal allotments made to the Arab local authorities. (The Druze local authorities were reviewed separately from the Arab local authorities; mixed local authorities were excluded from the database.) Findings indicate that the Arab local authorities received greater support. In the second stage, two additional factors were taken into account: (1) Size can be advantageous in the operation of local authorities. Consequently, it was anticipated that the amount of fiscal support per person in the smaller local authorities would be greater; (2) Fiscal allotments to local authorities that are weaker in socio-economic terms should be greater. Consequently, the data was analyzed according to population size, nationality and socio-economic class. Results of this analysis show that Jewish local authorities received more than their Arab counterparts, although the gap has continually decreased over the years.

Statistical Applications - Economics II

Path Analysis and Cumulative Measures Applied to the Spanish Customer Satisfaction Indices: A Marketing Application for Automobile Industry

López Caro Cristina[1], *Mariel Chladkova Petr and Fernandez Aguirre Karmele*

University of Basque Country, Bilbao, Spain [1]edplocac@bs.ehu.es

This study tries to identify the factors and structures that underlie in the Consumer Satisfaction (CS) using data from the automobile industry. The specific issue of this study is the way in which CS is extracted, since its measures are aggregated data where observations obtained in previous periods are added to the observations of current period. In this way, each variable depends not only on data obtained in certain moment but on the whole history. For this reason, adequate treatments of the cumulative structure of data are needed, apart from adequate model and method of estimation. The observed data come from the magazine Autopista that has interviewed 427 340 of its readers along 13 periods of time asking them about their satisfaction in 25 attributes of 130 models of cars. The methodology used for the treatment of data is a combination of the Structural Equation Models and the Path Analysis through the Path Diagram and the decomposition rules. Through the Path Analysis, the influence of the past and present values over the variances and covariances of the analyzed variables is studied. This methodology and models used, separate the measurement properties of the CS of one period from the satisfaction of previous periods.

Forecasting of Incurred But Not Reported Reserves with Generalized Linear Models

Tuğba Tunç[1] *and Gençtürk Yasemin*[2]

Department of Actuarial Sciences, Hacettepe University, Ankara, Turkey
[1]ttunc@hacettepe.edu.tr
[2]yasemins@hacettepe.edu.tr

Generalized Linear Models (GLMs) which are the generalization of the ordinary linear models do not rely on normality. These models use other distributions from the exponential dispersion family, including Poisson, Normal, Inverse Gaussian and Gamma distributions. In addition to this, a wide range of mean-variance relations, from constant variance to a variance proportional to the third power of the mean is described by using GLMs. GLMs are widely used to solve the problems faced in real life including actuarial problems such as survival models, risk classification, ratemaking, loss reserving etc.. As loss data are generally skewed and the variance of the observations are dependent on their mean, they can be modelled using GLMs. One of the successful application of GLMs in actuarial problems is in estimating Incurred But Not Reported (IBNR) Reserves. IBNR Reserves are provisions to be held for claims that have been incurred, but are not yet reported, or as yet not fully paid. The basic data generally consists of past payments which are broken down by policy year and by development year, and grouped into a so-called run-off triangle. It is important to project future payments in order to set up a suitable provision. We focus our interest in the use of GLMs to forecast IBNR Reserves.

Modeling and Simulation

Poisson Mixture Regression Model: Application to Financial Data

Fátima Gonçalves[1] *and Susana Faria*[2]

University of Minho, Guimaräes, Portugal
[1]fat.rod.goncalves@sapo.pt
[2]sfaria@mct.uminho.pt

In many financial applications, Poisson mixture regression models are used to analyse heterogeneous count data. These models assume that the observations of a sample arise from two or more latent classes, of unknown proportions, that are mixed, and parameter estimation is typically performed by means of maximum likelihood via the EM algorithm. In this study, we briefly discuss this procedure for fitting Poisson mixture regression models, proposing a new strategy to choose the set of initial values for the EM algorithm. We also investigate the testing of hypotheses concerning the number of components in the mixture regression via bootstrapping. We apply these models to real data for credit-scoring purposes, modelling the number of defaulted payments of clients who had obtained loans for consumption from a bank.

A Group Sequential Form for the Partially Grouped Log Rank Test and A Simulation Study

Yaprak Parlak Demirhan[1] *and Haydar Demirhan*[2]

[1]Quality Test and Certification Department, Undersecretariat for Defence Industries, Ankara, Turkey; ypdemirhan@ssm.gov.tr
[2]Department of Statistics, Hacettepe University, Ankara, Turkey;
haydarde@hacettepe.edu.tr

Clinical trials, which aim to investigate the effect of alternative treatments on survival rates, allow the subsequent patient entrance into the study. In these trials, one may want to compare the time to failure data that come from different treatments. It is the most convenient way to analyze the accumulated data in groups because group sequential tests provide the advantage of early stopping the trial.
In the test of significance of hazard ratio, one may claim that the differences are unimportant prior to a predetermined time t, however differences for time points exceeding the time t are important. For instance, a medicine may require a time period after which it becomes effective for the treatment of a disease. Sposto et al. (Sposto, R., Stablein, D., Carter-Campbell, S., 1997, A partially grouped log rank test, Statistics in Medicine, 16, 195 – 704) propose the partially grouped log rank (PGLR) statistic. It takes into account the required time period.
We propose a group sequential form for the PGLR test of Sposto et al. and extend it for some other variations of the log rank statistic, such as Tarone-Ware, Gehan-Wilcoxon, and Fleming-Harrington family. Then, we conduct a wide simulation study, including different proportional hazards scenarios, censoring rates, predetermined time points, sampling sizes and tied observations to analyse powers of the proposed group sequential form of PGLR test and the variations.

Measurement

Measurement and Expressions of Uncertainty of Material Characteristics

Athanasios Papargyris[1] *and Dimitrios Papargyris*[2]

[1]TEI Larissas, Larissa, Greece; papargyr@teilar.gr
[2]; edp@uth.gr

Accuracy and precision are very important in the measuring of performance. The aim of any measurement is the estimation of the its true value. However, the result of a measurement, is only an approximation of the true value of the specific quantity which is subject to measurement and is complete only when it is accompanied by a statement of its uncertainty. There are mainly two approaches for estimating uncertainty of analytical procedures in sciences, the ISO GUM approach originally intended for physical measurements, the Analytical Methods Committee inter-laboratory approach for chemical measurements and the ISO Guide and the UKAS expressions for mechanical measurements. The GUM approach has received much criticism as being costly, tedious, time consuming and analytically unrealistic, while the top-down technique can be applied when inter-laboratory exercises are available and the uncertainty could be underestimated if it is calculated from routine analysis results by a given laboratory with small bias or over estimated if the laboratory has a large bias. An excellent alternative to the above techniques are other techniques where the measurement uncertainty is evaluated from information gathered at the validation stage. More complicated of the above measurements is the measurement of uncertainty of mechanical properties due to large number of parameters which influence the measurement. The uncertainty can be expressed in various ways which could be confusing. The most commons expressions are: (a) as ? an arithmetic value, (b) as % value of foul scale deflection and (c) as % value of reading. In the present work a discussion of the above techniques is given and various examples of the way, the uncertainty can be expressed.

Measurement of Supranational Policy Level of Decision Making. A Practical Method for Helping Policy Makers

Lluis Coromina[1] *and Willem E. Saris*[2]

[1]University of Girona, Girona, Spain; lluis.coromina@udg.edu
[2]University Ramon Llull, Barcelona, Spain;

The European Union (EU) is formed by several countries, which all have their own national policy. Since these countries belong to a supranational institution, the EU, they decide jointly their policies. EU refers to the principle of subsidiarity in the sense that policy decisions should be taken as closely as possible to the citizen in order to ensure the lowest effective level of governance at the regional, national or supranational level. Our interest is to study the citizen's opinion about which policies should be decided at national or supranational level in 21 European countries. These policies are 'protecting the environment', 'fighting against organized crime', 'defense', 'social welfare', 'aid to developing countries', 'immigration and refugees', 'interest rates' and 'agriculture'. We first define an index for supranational level of decision making and then we obtained proportions for each policy in 21 countries. However, some countries are more similar than others concerning the proportion on those policies and it becomes quite complex its comparability. So, we classify these countries according their similarities on these policies using with cluster analysis. Four clusters are obtained and the interpretation of proportions for supranational levels of decision making on those policies is easier. A more sophisticated procedure we use is to study whether a cumulative scale exists on these policies in the different clusters. We use Mokken scale in order to check whether these items fulfill monotone homogeneity and non intersection requirements. A final analysis is to relate the supranational level of decision making with sociopolitical variables. We use a regression model in which the clusters where introduced as dummy coded variables including all possible interaction effects. Optimal transformations of the main and interaction variables were used.

Pitfalls and Remedies in Testing the Calibration Quality of Rating Systems

Wolfgang Aussenegg[1], *Florian Resch*[2] *and Gerhard Winkler*[3]

[1]Vienna University of Technology, Vienna, Austria; waussen@pop.tuwien.ac.at
[2]Oesterreichische Nationalbank, Vienna, Austria; florian.resch@oenb.at
[3]WU Wien - Vienna University of Economics and Business, Vienna, Austria;
gerhard.winkler@oenb.at

Testing calibration quality by means of backtesting is an integral part in the validation of credit rating systems. Against this background this paper provides a comprehensive overview of existing testing procedures. We study their deficiencies theoretically and illustrate their impact empirically. Based on the insights gained thereof, we develop enhanced hybrid testing procedures which turn out to be superior to the commonly applied methods. We also propose computationally efficient algorithms for our calibration tests. Finally, we are able to demonstrate empirically that our method outperforms existing tests in a scenario analysis using rating data of Moody's.

Workshop

Statistics Education and Educational Research Based on Reproducible Computing

Patrick Wessa, Bart Baesens, Stephan Poelmans and Ed van Stee

K.U.Leuven Association, Integrated Faculty of Business and Economics, Leuven, Belgium;
patrick@wessa.net

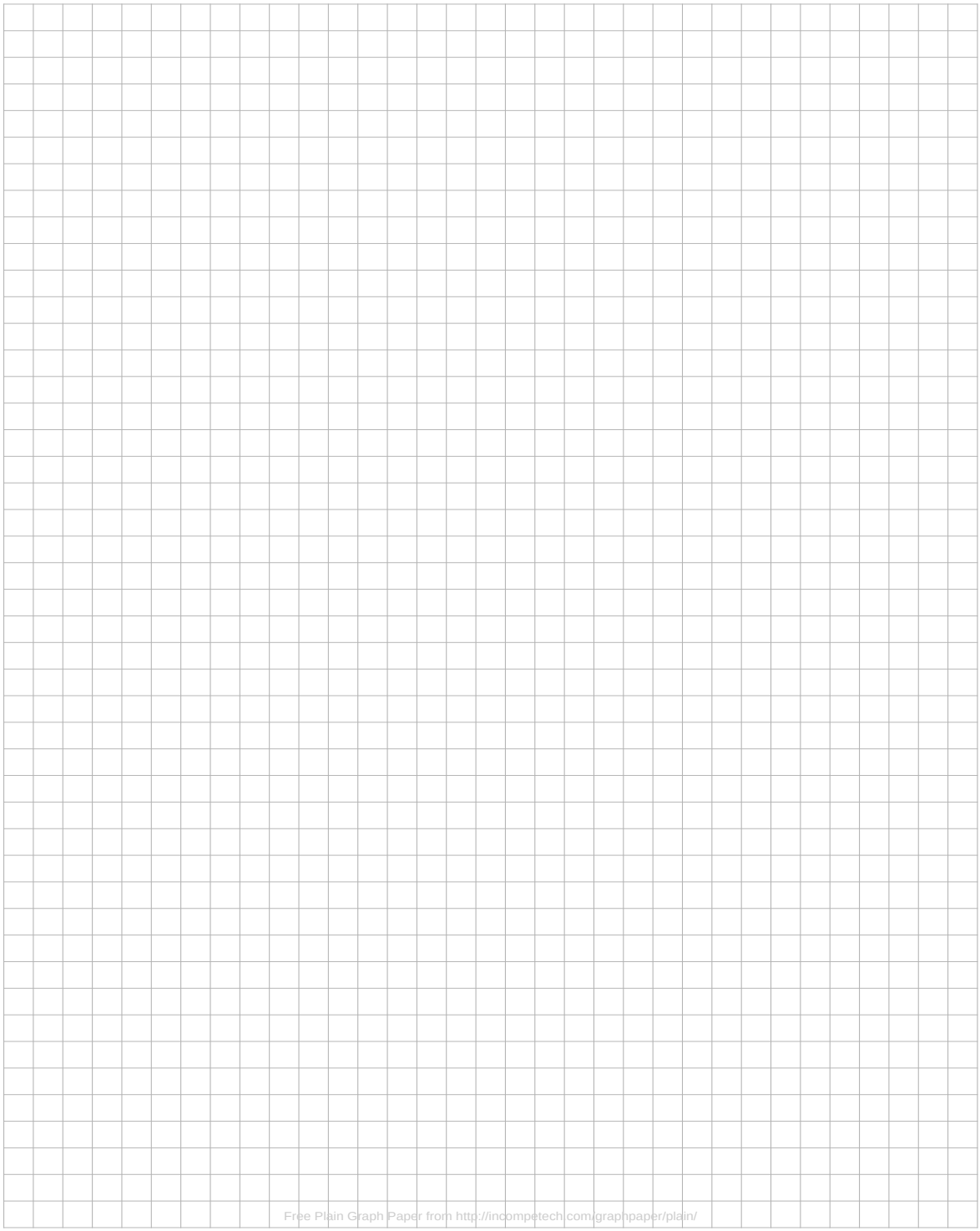
Workshop will be divided into three parts, a general one and two practical ones with limited space for 20 attendees. You have to register for parts 2 and 3 using the format that will be available at the workshop website. You will get the information after registration for the conference.
What is the focus of this workshop?
This workshop may be useful for anyone with an interest in one of the following three topics:

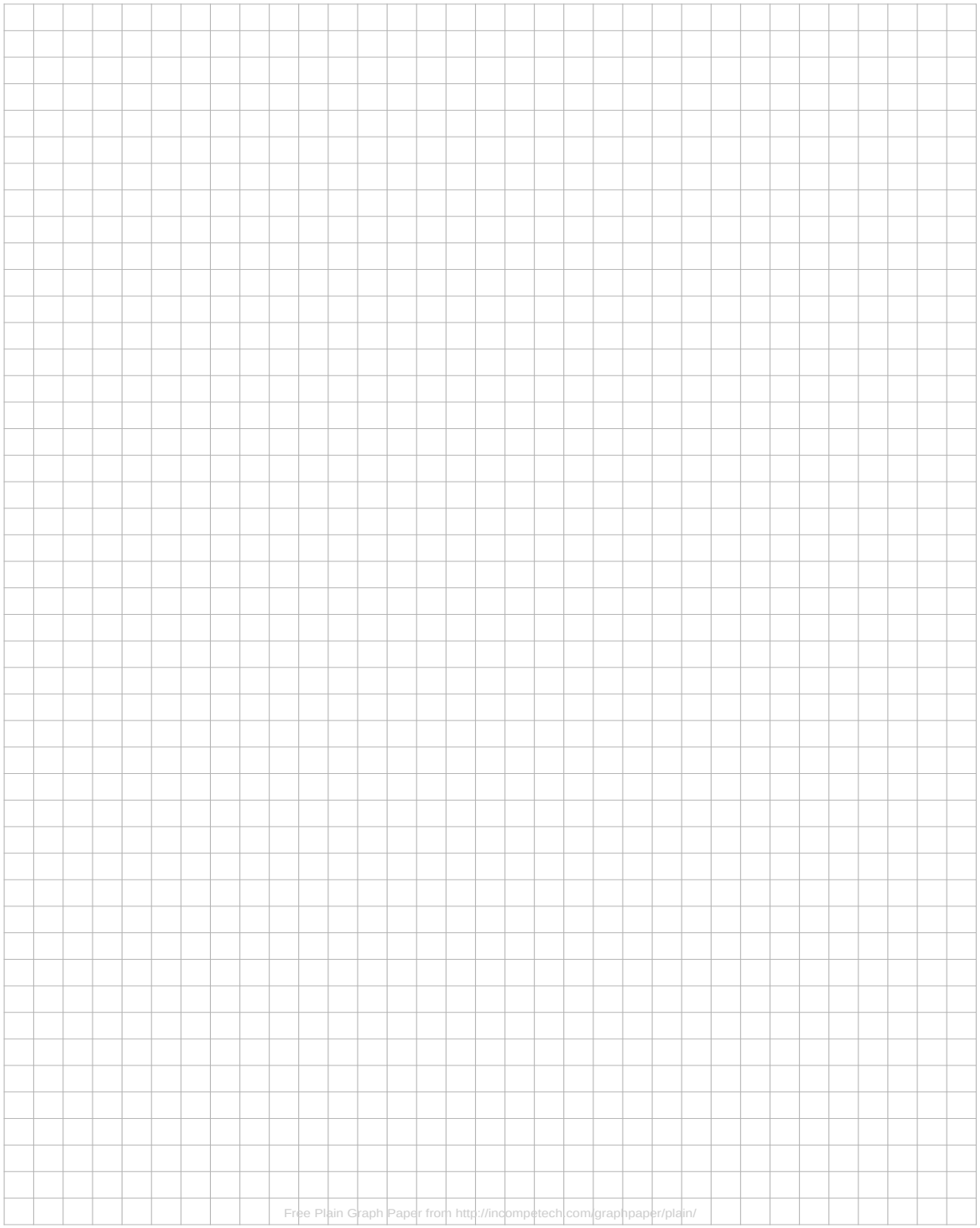
- Statistics Education within the pedagogical paradigm of social constructivism,
- Educational Research based on objectively measured learning activities which have never been available before,
- Empirical Research which is fully reproducible – this is called Reproducible Computing which supports communication, collaboration, and dissemination of research results.

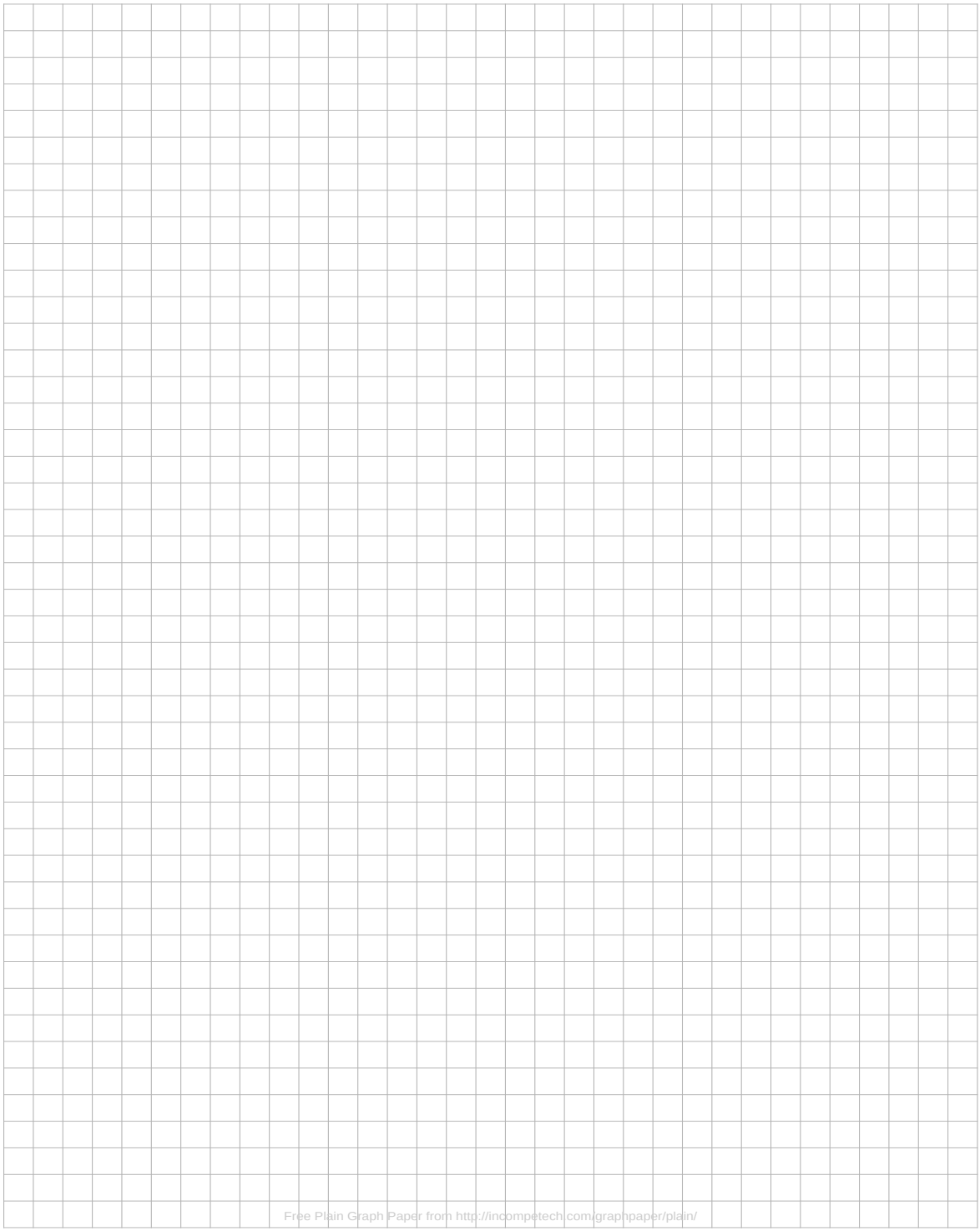
The main focus of this workshop is on Statistics Education or any type of education where students need to be able to interact with and communicate about empirical research results. In this sense, the workshop may be of interest to academics from various fields.

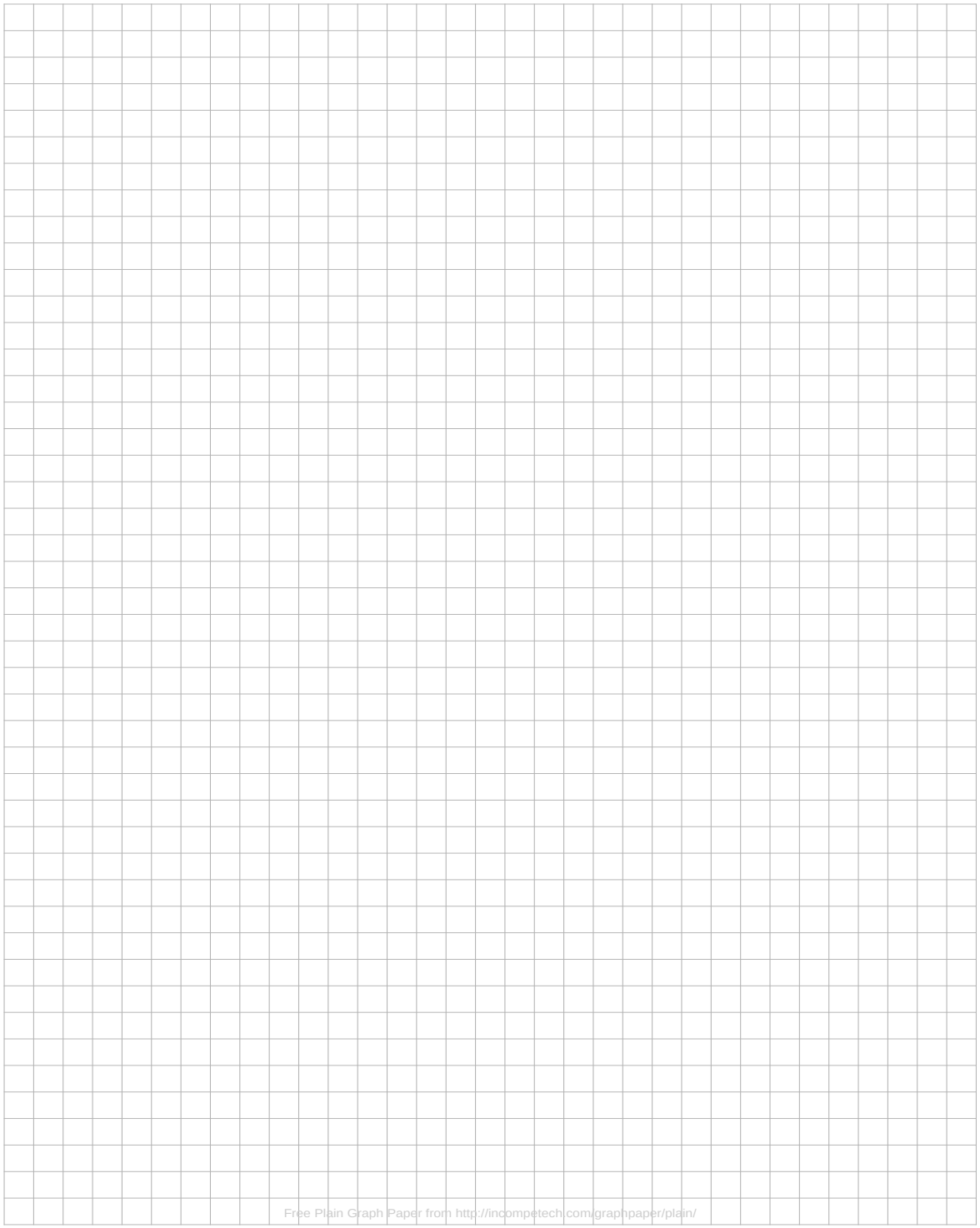
INDEX OF AUTHORS

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