International Conference

APPLIED STATISTICS

2012

PROGRAM and ABSTRACTS

September 23 - 26, 2012
Ribno (Bled), Slovenia

http://conferences.nib.si/AS2012

International Conference

APPLIED STATISTICS 2012

PROGRAM and ABSTRACTS

September 23 – 26, 2012

Ribno (Bled), Slovenia

http://conferences.nib.si/AS2012

Organized by

Statistical Society of Slovenia

Supported by

Slovenian Research Agency (ARSS)

Statistical Office of the Republic of Slovenia

ALARIX

RESULT d.o.o.

The word cloud on the cover was generated using www.wordle.net. The source text included the abstracts of the talks; the fifty most common words were displayed, and greater prominence was given to words that appeared more frequently.

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

311(082)

INTERNATIONAL Conference Applied Statistics (2012 ; Ribno)
Program and abstracts / International Conference Applied Statistics 2012, September 23–26, 2012, Ribno (Bled), Slovenia ; organized by Statistical Society of Slovenia ; [edited by Lara Lusa and Janez Stare]. - Ljubljana : Statistical Society of Slovenia, 2012

Dostopno tudi na: http://conferences.nib.si/AS2012/AS2012-Abstracts.pdf

ISBN 978-961-92487-8-2
ISBN 978-961-92487-9-9 (pdf)
1. Applied Statistics 2. Lusa, Lara 3. Statistično društvo Slovenije
263187968

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
Jacques Esteve, France
Tamas Rudas, Hungary
Albert Satorra, Spain
Hans Waege, Belgium

Organizing Committee

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

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
Produced using: generbook R package
Circulation: 200

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	Developments in Statistics	
	13.40 – 15.00	Lunch	
	15.00 – 16.20	Education I	Measurement
	16.20 – 16.40	Break	
	16.40 – 17.40	Education II	Sampling Techniques and Data Collection
	19.00	Reception	
Monday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Biostatistics and Bioinformatics I	
	11.40 – 12.00	Break	
	12.00 – 13.00	Biostatistics and Bioinformatics II	Statistical Applications - Economics I
	13.00 – 14.30	Lunch	
	14.30	Excursion	
Tuesday	9.10 – 10.00	Invited Lecture	
	10.00 – 10.20	Break	
	10.20 – 11.40	Social Science Methodology	
	11.40 – 12.00	Break	
	12.00 – 13.20	Modeling and Simulation I	Statistical Applications - Economics II
	13.20 – 15.00	Lunch	
	15.00 – 16.20	Econometrics	Mathematical Statistics
Wednesday	9.10 – 10.30	Modeling and Simulation II	Design of Experiments and Statistical Applications
	10.30 – 10.50	Break	
	10.50 – 12.10	Modeling and Simulation III	Statistical Applications - Biostatistics
	12.10 – 12.30	Closing of the conference	
	12.30 – 14.00	Lunch	
	14.00 – 18.00	Workshop	

10.30–11.00 **Registration**

11.00–11.10 **Opening of the Conference**

11.10–12.00 **Invited Lecture** (Hall 1) *Chair: Katarina Košmelj*

1. **Symbolic Data Analysis: Are Distributions the Numbers of the Future? An Illustrative Answer**
   Lynne Billard

12.00–12.20 **Break**

12.20–13.40 **Developments in Statistics** (Hall 1) *Chair: Lynne Billard*

1. **Mallows' Distance in Symbolic Data Analysis**
   Katarina Košmelj and Lynne Billard

2. **Recent Developments in the Statistical Analysis of Interval Data**
   Ulrich Poetter, Georg Schollmeyer, Thomas Augustin, Marco Cattaneo and Andrea Wiencierz

3. **Outlier Detection in Overdispersed Proportions - A Comparison of Four Main Approaches**
   Gaj Vidmar and Rok Blagus

4. **Data Analysis using R in Extreme Value Theory**
   Helena A. Penalva, Manuela C. Neves and Sandra D. Nunes

13.40–15.00 **Lunch**

15.00–16.20 **Education I** (Hall 1) *Chair: Matevž Bren*

1. **The Determinants of Academic Success and Failure in a Competing Risks Approach**
   Renata Clerici, Anna Giraldo and Silvia Meggiolaro

2. **The Statistical Textual Analysis Applied to the Italian University Offering Database**
   Claudia Caruso

3. **Possibilities of Application of Statistical Analysis in the System of Higher Education Institution Quality**
   Zorana Lužanin and Valentina Sokolovska

4. **Statistical Approaches to Inner Design of Primary School Classroom**
   H.Derya Arslan, Kerim Çınar and Pınar Dinç

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

1. **Time Series of Macroecomic Indicators for the Czech Republic 1970 - 1990**
   Jaroslav Sixta and Jakub Fischer

2. **The Development of Labour Productivity in the Czech Republic in the Period between 1970 and 1989**
   Kristyna Vltavska and Jaroslav Sixta

3. **Methodological Manual on Purchasing Power Parities: A Useful Tool for Regional Price Levels?**
   Petr Musil and Jana Kramulova

4. **A Demographic Island. Peculiar Features of People Distribution in Sardinia**
   Maurizio Brizzi and Alessia Orrù

16.20–16.40 **Break**

16.40–17.40 **Education II** (Hall 1) *Chair: Andrej Blejec*

1. **Team Teaching with Students: Experimental Study Part 2**
   Jerneja Šifrer, Žala Žvab and Matevž Bren

2. **Internal and External Grading at Slovene Matura Exam: Differences in Distributions Shown by Ordinal Dominance Graphs**
   Darko Zupanc, Gašper Cankar and Matevž Bren

3. **R-Excel Tools for Homework Assignment and Exam Preparation**
   Lara Lusa

16.40–18.20 **Sampling Techniques and Data Collection** (Hall 2) *Chair: Maurizio Brizzi*

1. **Challenges in Measurement of Sustainable Development in the Czech Republic**
   Jana Kramulova, Lenka Hudrlikova and Jan Zeman

2. **Sampling Coordination of Business Surveys**
   Maruša Štanek and Boro Nikić

3. **SPAMIA: Spam Filtering by Quantitative Profiles**
   Marian Grendár, Jana Škutová and Vladimír Špitalský

4. **An Algorithm for Computing a Combined Interestingness Measure in Association Analysis**
   Derya Ersel and Süleyman Günay

5. **Ratio-type Estimator of Population Total with Minimum MSE: A Computational Approach**
   Mohammadreza Faghihi and Zahra Noori

19.00 **Reception**

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

1. **Multi-state Models: A Variety of Uses**
   Vern Farewell

10.00–10.20 **Break**

10.20–11.40 **Biostatistics and Bioinformatics I** (Hall 1) *Chair: Vern Farewell*

1. **Development of a Flexible Excess Hazard Model with Shared Frailty, Non-Linear and Time-Dependent Effects of Covariates**
   Hadrien Charvat, Laurent Remontet, Nadine Bossard, Laurent Roche, Jacques Estève and Aurélien Belot

2. **On Comparison of Measures of Explained Variation in Survival Analysis**
   Nataša Kejžar, Delphine Maucort-Boulch and Janez Stare

3. **Properties of a Net Survival Estimator**
   Maja Pohar Perme

4. **Dynamic Survival Model for Clustered Multiple Failure Time Data**
   Amirhossein Jalali, Firoozeh Haghighi, Zahra Rezaei Ghahroodi and Shirin Moghaddam

11.40–12.00 **Break**

12.00–13.00 **Biostatistics and Bioinformatics II** (Hall 1) *Chair: Maja Pohar Perme*

1. **Spatio-Temporal Modelling of Daily Rainfall**
   Ana F. Militino, Maria D. Ugarte and Manuel Garcia-Magarińos

2. **Benefits and Caveats of Massive Microarray Data Production**
   Ana Rotter, Urška Tajnšek, Kristina Gruden, Helena Motaln and Tamara Lah Turnšek

3. **Improved Shrunken Centroid Classifiers for High-Dimensional Data**
   Rok Blagus and Lara Lusa

12.00–13.00 **Statistical Applications - Economics I** (Hall 2) *Chair: Delphine Maucort-Boulch*

1. **Testing Export-Led Growth Hypothesis: The Case of Turkey, 1961-2010**
   Mustafa Murat Arat

2. **Water Utilization, Water Quality in the Endogenous Economic Growth: Reflections On Fixed Effects**
   Souha El Khanji and John Hudson

3. **SMEs and Fast-Growing Companies in Survey Estimates**
   Aleša Lotrič Dolinar, Rudi Seljak and Mojca Bavdaž

13.00–14.30 **Lunch**

14.30 **Excursion**

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

1. **Data Analysis: Best Practices and Future Directions**
   Hadley Wickham

10.00–10.20 **Break**

10.20–11.40 **Social Science Methodology** (Hall 1) *Chair: Hadley Wickham*

1. **Advances in Methods for Causal Inference of Observational Studies**
   Ana Kolar and Vasja Vehovar

2. **Post-Stratification Weighting Issues in Cross-National Survey Research**
   Ana Slavec and Vasja Vehovar

3. **Internet Forums: What Information is Out There?**
   Vanja I. Erčulj

4. **Clustering Macroeconomic Variables**
   Chiara Perricone

11.40–12.00 **Break**

12.00–13.20 **Modeling and Simulation I** (Hall 1) *Chair: Jacques Esteve*

1. **Generating Random Numbers from Empirical Distributions**
   Katja Rostohar, Anja Žnidaršič, Jelka Šuštar Vozlič and Andrej Blejec

2. **A Proposition of a Hybrid Stochastic Lee– Carter Mortality Models**
   Leslaw Socha and Agnieszka Rossa

3. **Bootstrap Confidence Intervals of the Difference between Two Process Capability Indices for Half Logistic Distribution**
   Somchit Wattanachayakul and Wararit Panichkitkosolkul

4. **Defective Rate Control Charts for Zero-Inflated Processes**
   Chanaphun Chananet and Piyachat Leelasilapasart

12.00–13.20 **Statistical Applications - Economics II** (Hall 2) *Chair: Giovanni Millo*

1. **The Log-Periodic Power Law model for financial bubbles with ARMA/GARCH errors**
   Špela Jezernik Širca

2. **On Longevity Risk Models in the Romanian Annuity Market and Pension Funds**
   Iulian Mircea and Mihaela Covrig

3. **Comparison of Artificial Neural Networks, Autoregressive Model and Multiple Linear Regression for Monthly Streamflow Estimation**
   Meral Büyükyıldız and Tezel Gülay

4. **Comparing the Markov Switching AR, Nonlinear Additive AR, Self-Exciting Threshold AR and Logistic Smooth Transition AR models for Analysis Time Series Data With Dramatic Jumps**
   Masoud Yarmohammadi

13.20–15.00 **Lunch**

15.00–16.20 **Econometrics** (Hall 1) *Chair: Aleša Lotrič Dolinar*

1. **Insurance Markets in the Long Run**
   Giovanni Millo and Gaetano Carmeci

2. **On the Proportional Retention Reinsurance Problem**
   Mihaela Covrig, Daniela Todose, Emilia Titan and Simona Ghita

3. **Hybrid Fuzzy Mortality Models of LC-type: A Simulation Study**
   Andrzej Szymański and Agnieszka Rossa

4. **A Modified Weighted Symmetric Estimator for a Gaussian First-Order Autoregressive Model with Additive Outliers**
   Wararit Panichkitkosolkul and Patarawan Sangnawakij

15.00–16.20 **Mathematical Statistics** (Hall 2) *Chair: Nataša Kejžar*

1. **Confidence Intervals for a Ratio of Binomial Proportions Based on Direct and Inverse Sampling**
   Thuntida Ngamkham, Kamon Budsaba and Araya Chaemchan

2. **Generalized Confidence Interval for the Difference between Normal Population Variances**
   Wichitra Phonyiem and Sa-att Niwitpong

3. **The Comparison of Tests for Equality of Coefficients of Variation for Data with Outliers**
   Piyachat Leelasilapasart and Chanaphun Chananet

4. **Random Effect One-Way ANOVA Model when Sampling from a Finite Population of Treatment Groups**
   Kamon Budsaba, Teerawat Simmachan and John J. Borkowski

9.10–10.30 **Modeling and Simulation II** (Hall 1) *Chair: Rok Blagus*

1. **Smooth Bootstrap Inference for Parametric Quantile Regression**
   Tatjana Kecojevic and Peter Foster

2. **Bayesian Model Selection Criteria for Generalized Linear Models with Data Missing Not at Random**
   Zeynep Kalaylioglu

3. **Estimation of Daily Evaporation Using Different Modeling Methods**
   Gülay Tezel and Meral Büyükyıldız

4. **An Application of Optimization Algorithms for Generating Correlated Multivariate Random Samples**
   Anamai Na-udom and Jaratsri Rungrattanaubol

9.10–10.30 **Design of Experiments and Statistical Applications** (Hall 2) *Chair: Aurélien Belot*

1. **Enhancement of Search Algorithms for Constructing Optimal Latin Hypercube Designs**
   Jaratsri Rungrattanaubol and Anamai Na-udom

2. **An Analysis of Turkey Demographic and Health Survey 2008 Data with Co-Plot Method**
   Yasemin Kayhan Atilgan and Süleyman Günay

3. **Estimation of Curvature and Displacement Ductility in Reinforced Concrete Buildings**
   M. Hakan Arslan

4. **Individual Control Treatments in Designed Genetical and Agricultural Experiments**
   Stanislaw F. Mejza and Iwona Mejza

10.30–10.50 **Break**

10.50–12.10 **Modeling and Simulation III** (Hall 1) *Chair: Gaj Vidmar*

1. **Statistical Methods for Processing and Analysis of Digital Images: Determining Compressive Strength of Concrete**
   Gamze Cankaya, M. Hakan Arslan and Murat Ceylan

2. **Bayesian Estimation of Odds Ratios: An Application**
   Deniz Taşçı and Süleyman Günay

3. **Fast Robust Kernel Density Estimation**
   Kourosh Dadkhah

4. **Bayesian Test of Homogeneity of Transition Model for Analyzing Longitudinal Ordinal Data**
   Sajad Noorian and Mojtaba Ganjali

10.50–11.50 **Statistical Applications - Biostatistics** (Hall 2) *Chair: Ana Rotter*

1. **Plant Life Forms and Ecological Indices of Lake Provala**
   Katarina J. Čobanović, Ljiljana M. Nikolić and Slobodan C. Ničin

2. **Statistical Aspects of Evaluation of Environmental Policy in Slovenia**
   Žiga Kotnik and Maja Klun

3. **Data Transformations and Normality: Example from Wheat Drought Stress Trial**
   Miroslav Z. Zorić , Emilija B. Nikolić-Djorić and Dragan S. Djorić

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

12.30–14.00 **Lunch**

14.00–18.00 **Workshop** (Hall 1)

1. **Creating Effective Visualizations**
   Hadley Wickham

ABSTRACTS

Invited Lecture

Symbolic Data Analysis: Are Distributions the Numbers of the Future? An Illustrative Answer

Lynne Billard

Department of Statistics, University of Georgia, Athens, GA, United States of America
lynne@stat.uga.edu

Massively large data sets are routine and ubiquitous given modern computer capabilities. What is not so routine is how to analyse these data. One approach is to aggregate the data sets according to some scientic criteria. The resultant data are perforce symbolic data, i.e., lists, intervals, histograms, and so on. Applications abound, especially in the medical and social sciences. Other data sets (small or large in size) are naturally symbolic valued, such as species data, data with measurement uncertainties, confidential data, and the like. Unlike classical data which are points in p-dimensional space, symbolic data are hypercubes or Cartesian products of distributions in p-dimensional space. We describe such data and how they arise. We look brifley at some of the differences between classical and symbolic data and their respective methodologies, through illustrations.

Developments in Statistics

Mallows' Distance in Symbolic Data Analysis

Katarina Košmelj[1] *and Lynne Billard*[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]University of Georgia, Athens, USA
katarina.kosmelj@bf.uni-lj.si, lynneb@stat.uga.edu

Countries described by population pyramids can be regarded as symbolic data objects with two random variables, one presenting age for males and one for females. In the literature, several distances for histogram-type data can be found. We have decided to apply the Mallows' distance for several reasons. It is a well defined metric; its calculation in the histogram setting is simple, even when the number and length of histograms' subintervals differ. This distance allows the constructions of a barycentric histogram which is an "optimal" cluster representative. It also allows to define a measure of total inertia which can be decomposed into the within and between inertia according to the Huygens theorem. Finally, Mallows' distance can be decomposed into three terms: the location term, the size term and the shape term. These characteristics are very helpful in statistical methods based on distances, such as cluster analysis and MDS. A case study on population pyramids of East European countries in the period 1995-2015 was undertaken. The results provide an insight of the information that this distance can extract from a complex dataset.

Recent Developments in the Statistical Analysis of Interval Data

Ulrich Poetter[1], *Georg Schollmeyer*[2], *Thomas Augustin*[2], *Marco Cattaneo*[2] *and Andrea Wiencierz*[2]

[1]Department Social Monitoring, German Youth Institute, Munich, Germany
[2]Department of Statistics, University of Munich, Munich, Germany
poetter@dji.de, georg.schollmeyer@stat.uni-muenchen.de,
thomas@stat.uni-muenchen.de, cattaneo@stat.uni-muenchen.de,
andrea.wiencierz@stat.uni-muenchen.de

Survey data for variables like income or age are often grouped, heaped or rounded and thus only provide the information that the precise value lies in some interval. Since traditional methods for interval-censored data strongly depend on optimistic assumptions on the coarsening process, an alternative methodology is currently getting strong momentum. Questioning the implicit paradigm that imprecise data should nevertheless produce precise estimates, the corresponding methods look for an optimal *set* of models, naturally reflecting the extent of data imprecision and describing the best that can reliably be learned from the data. Typically the results are still informative enough to provide valuable insights into the underlying subject matter questions. We compare likelihood-based imprecise regression (e.g., Cattaneo/Wiencierz , IJAR to appear, 2012) with approaches in the spirit of partial identification (like Rohwer/Pötter, Juwenta, 2002; Manski/Tamer Econometrica, 2002,Vol. 70, p. 519-546; Beresteanu/Molchanov /Molinari, Econometrica, 2011,Vol. 79, p. 1785-1821), which, in essence, determine the envelopes of estimates arising from all potential data completions. The comparison contrasts the methodological background, areas of application, and the potential for different generalizations, including the ability to utilize weak background information and paradata. We investigate performance in a simulation study and present some illustrative applications using ALLBUS data and Swiss drug data.

Outlier Detection in Overdispersed Proportions - A Comparison of Four Main Approaches

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

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

Outlier detection among overdispersed proportions is a major issue in healthcare quality control. We had previously introduced control limits for the double-square-root control-chart based on prediction-intervals from regression-through-origin and compared our approach to common outlier tests. Here, we compare it to three other approaches: Laney's p'-chart for cross-sectional data, Spiegelhalter's regression-modelling approach – multiplicative (generalized linear model) and additive (random-effects), and Carling's resistant outlier rule (median rule).
Comparisons were performed on real and simulated data. The real data comprised hospital readmissions data from UK (analysed by Spiegelhalter and Laney) and data on business indicators of healthcare quality (officially monitored in all hospitals in Slovenia). Simulations resembled business indicators of healthcare quality, which are same-quantity random-denominator ratios. We generated small (below 0.2; right-skewed) and large (between 0.5 and 1, more symmetrically distributed) proportions, 1000 under each experimental condition. Samples of size N=10, 30, 50 and 100 were drawn from 3-parameter-loglogistic distribution with no or one outlier added (the largest value, drawn with location parameter x2 or x6). Various measures of diagnostic performance were examined.
In the simulations, Spiegelhalter's approach yielded very high false-alarm rates, except the multiplicative version in tiny samples. Laney's approach produced fewest false alarms, but could not detect the outlier for N=10 with small proportions and failed to detect the outlier regardless of sample size with large proportions. Median rule performed similarly (better with small proportions for N=10 but with higher false-alarm rate for N=30 or more). Our approach proved probably the best overall: similar to the median rule (just less liberal) with small proportions (though unable to detect the outlier for N=10) and the only generally useful one with large proportions.
Further research might explore theoretical relations between Spiegelhalter's, Laney's and our approach, and applicability of tests and models for overdispersed proportions to statistical quality control.

Data Analysis using R in Extreme Value Theory

Helena A. Penalva[1], *Manuela C. Neves*[2] *and Sandra D. Nunes*[1]

[1]Department of Economics and Management , College of Business Administration, Polytechnic Institute of Setubal, Setubal, Portugal
[2]Department of Mathematics, Higher Institute of Agronomy, Technical University of Lisbon , Lisboa, Portugal
helena.penalva@esce.ips.pt, manela@isa.utl.pt, sandra.nunes@esce.ips.pt

Extreme value theory has emerged as one of the most important statistical areas in several applied sciences, such as insurance, risk assessment, telecommunications, environment and biology. In analysis of extreme values it is of great importance the model assumptions on the tail of the underlying distribution function to the sample data. Statistical inference about extreme events is interested in the estimation of the probability of occurrence of events more extremes than any that have already been observed. There are a few parameters whose estimation is of major importance: the extreme value index which is the basis of all parameters of extreme events and is directly related with the heaviness of the tail of the underlying distribution; the probability of exceedance of a high level; the return period of a high level; the right endpoint of an underlying model and a high quantile of probability 1-p, with p small. Accurately modelling extreme events has become more and more exigent and the analysis require tools that must be simple to use but also should consider complex statistical models in order to produce valid inferences. A set of steps for performing a data analysis of extreme values in R (R Development Core Team, 2012) environment will be presented. This work intends to introduce first a brief background to some models that form the basis for the theory of statistical extremes. Afterwards, and using some data sets, a review of some packages and functions included in R will be shown.

Education I

The Determinants of Academic Success and Failure in a Competing Risks Approach

Renata Clerici , Anna Giraldo and Silvia Meggiolaro

Department of Statistical Sciences, University of Padova, Padova, Italy
renata.clerici@unipd.it, anna.giraldo@unipd.it, meg@stat.unipd.it

Obtaining an university degree results in important outcomes for subsequent life course. However, choosing to start the path of university does not guarantee that the student will actually graduate: university withdrawal is one of the major problems. In fact, highly complex educational histories are observed in the learning process. This paper aims at examining the factors influencing the different outcomes of the university path (withdrawal, course changes, delay, and degree completion) in three-year degree courses in a big Italian University (Padova). Data from the university administrative archives allow to have information on about 32,000 students enrolled from 2002/03 to 2005/6 academic years in 84 undergraduates courses. The analyses are conducted considering the temporal dimension within the methodological approach of survival analysis using individual longitudinal data for cohorts of first-entering students. A discrete-time method for competing risks event history analysis is used to study the determinants of academic outcomes. Preliminary results confirm descriptive findings showing the important role of some background characteristics (such as gender, residence, and the nationality) and of some characteristics of secondary school career (the type of school, the grade, and the regularity) for successful university path.

The Statistical Textual Analysis Applied to the Italian University Offering Database

Claudia Caruso

Evaluation Office, University Federico II , Naples, Italy
c.caruso@unina.it

Improving the outreach activities as the third mission for university, besides research and teaching, has become one of the main goals of the university system. In Italy, the national regulation has redefined the role of the university system with the aim to enhance the interaction between knowledge and work and professional world. So, for planning the university degrees, it has become compulsory to consult with the local market representatives in advance, to monitor and update the offering. In order to give accountability about the implementation of the process, the university courses have to fill in specific forms for reporting the summary of these consultations. The outcome of this process has been made public in a national database, available on line, on the Ministry of Education website, about the offering of the new course of studies, where there is a specific section with the summary of the consultations for each degree. In this working paper, a textual analysis has been carried out on this specific section of the national database, to provide the measures of lexical richness, based on the ratio of type, token and the proportion of infrequent word and so on. To describe the lexical words used, a cooccurrence analysis has been performed to identify the key words of the corpora using the chi-square test and identify the thematic clusters.

Possibilities of Application of Statistical Analysis in the System of Higher Education Institution Quality

Zorana Lužanin[1] *and Valentina Sokolovska*[2]

[1]Faculty of sciences, Novi Sad, Serbia
[2]Faculty of Philosophy, Novi Sad, Serbia
zorana@dmi.uns.ac.rs, valentina.sokolovska25@gmail.com

The presentation considers the possibilities of applying statistical analysis in the research within the scope of higher education institution quality. The issues which specifically stand out are the sample problems and the problems of instrument defining. Special attention is devoted to the research conducted by the Committee for Quality Assurance and Internal Evaluation at the University of Novi Sad in 2012. The sample comprises 13 faculties. By using two questionnaires, 8,500 students and 860 teachers/professors have been tested. On the one hand, the aim of the research is to explore to what extent the students are familiar with the manner in which the ECTS points are determined, as well as with the compatibility of the defined points with their load. On the other hand, it has been analysed whether the students and professors are introduced, and to what extent, to the "Student at the centre of learning" concept; whether the students' questionnaires on quality of teaching (grading of teachers) have influence on the increase of quality of studying, etc. The collected data of this research will be analysed by the chosen statistical techniques for the analysis of categorical data.

Statistical Approaches to Inner Design of Primary School Classroom

H.Derya Arslan[1], *Kerim Çınar*[1] *and Pınar Dinç*[2]

[1]Selcuk University, Konya, Turkey
[2]Gazi University, Ankara, Turkey
kolderya@selcuk.edu.tr, kcinar@selcuk.edu.tr, pdinc@gazi.edu.tr

This experimental and analytical study was conducted with primary school students and classroom teachers to define the components of setting and the design factors of the classroom setting in which primary students will willingly study, and which setting and design factors are most suitable for effective learning. The visuals of 20 different classroom settings – the physical properties of which were pre-defined in detail by the control group (composed of academician architects) - were used in the scope of the study. Directed at primary school classroom settings, this study was conducted on primary school 2nd grade students (n=189) and university 4th grade (8th semester) students, attending the Classroom Teaching Department (n=100), randomly selected from the regions representing low and high socio-economic groups of Turkey. Fourteen academicians lecturing in the architecture faculties of various universities were selected into the "control group". A questionnaire form was used to collect user evaluations on the visuals, collected data were analyzed by using SPSS. Before the statistical analysis procedure, data related to concepts in question were subjected to reliability analysis. This study proves the hypothesis that the "classroom setting perceptions of users differ in terms of setting and design factors". It was also determined that findings and users' perception of classroom setting differ from each other in terms of environment and design factors. These variances also indicated that age and education factors are two parameters significant for perception studies.

Measurement

Time Series of Macroecomic Indicators for the Czech Republic 1970 - 1990

Jaroslav Sixta and Jakub Fischer

University of Economics in Prague, Prague, Czech Republic
sixta@vse.cz, fischerj@vse.cz

The paper deals with the construction of time series of main macroeconomic indicators for the Czech Republic covering the period 1970 - 1990. It describes basic transformation procedures of indicators based on socialist Material Product System (MPS) into the System of National Accounts (SNA). It is focused on the expenditure approach to gross domestic product covering household and government consumption expenditures, gross capital formation and net export. The paper briefly describes how originally published data were adjusted to fit current statistical requirements given by SNA. These adjustments cover mainly imputed rent and non-productive services consumed by households. Beside that specific issue was represented Czech – Slovak foreign trade within Czechoslovak federation. Presented time series cover both nominal indicators and volume indices based on the prices of previous' year.

The Development of Labour Productivity in the Czech Republic in the Period between 1970 and 1989

Kristyna Vltavska and Jaroslav Sixta

University of Economics, Prague, Prague, Czech Republic
kristyna.vltavska@vse.cz, sixta@vse.cz

As the historical time series of the gross domestic product of the Czech Republic using SNA 1993 were published other analyses can be made. The dataset contains data of gross value added (in constant prices and current prices) and total employment based on classification NACE rev. 2 so that we are able to estimate the labour productivity in Czech industries. This paper presents the analysis of labour productivity as the ratio of the output to the input used in the period between 1970 and 1989. Deeper analyses of employment and structure of the gross value added are also part of the paper.

Methodological Manual on Purchasing Power Parities: A Useful Tool for Regional Price Levels?

Petr Musil and Jana Kramulova

University of Economics, Prague, Prague, Czech Republic
petr.musil@vse.cz, jana.kramulova@vse.cz

This contribution focuses on the possible extension of the Purchase Power Parities (PPP) methodology to the regional level. The economic development comparison of different countries is usually based on Gross Domestic Product (GDP) per inhabitant that is valued in Purchase Power Standard (PPS) in order to eliminate differences in price levels among countries. Not only national but also regional macro aggregates are valued in national prices; that means that regional differences are not taken into account during computations. The aim of this contribution is to propose an alternative approach for the regional price levels estimation. For this we chose the case study of the Czech Republic. PPS is generally based on the data (prices and weights) estimated by expenditure approach, but in the Czech Republic only production and income approaches are used for computation of the regional GDP. Calculation of regional price levels in our contribution is based on final household consumption expenditures, which represent the main (greatest) component of GDP (in the Czech Republic approximately 50 %) and for which the biggest differences are expected. EKS method (proposed in the OECD/EUROSTAT Methodological Manual on Purchasing Power Parities) is used with several adjustments. Regional indicators concerning households as well as regional GDP are recalculated. Indicators such as average income or net disposable income are adjusted to local price level in order to provide more reliable data on living conditions in regions.

A Demographic Island. Peculiar Features of People Distribution in Sardinia

Maurizio Brizzi[1] *and Alessia Orrù*[2]

[1]Dept. of Statistical Sciences, University of Bologna, Bologna, Italy
[2]University of Cagliari, Dept. of Experimental Biology, Cagliari, Italy
maurizio.brizzi@unibo.it, ceylon2005@yahoo.it

Among the 20 Italian regions, Sardinia is the most geographically isolated and has some peculiar territorial characteristics, being constituted by a main island of 24,090 square kms (slightly larger than Slovenia), located just in the middle of Mediterranean Sea, and several small isles near to the coast; all this has induced relevant consequences to the composition and distribution of people living there, therefore it results to be useful to perform a territorially detailed study of age and sex distribution of Sardinian people. We tried to do it by considering census and inter-census data, starting with the Sardinian Census held in 1844, up to the last inter-census survey of 2010. Although now Sardinia is divided in 8 provinces, we have considered the classic four-province division: Cagliari, Nuoro, Oristano and Sassari, analyzing separately their population distribution. We focused particularly our attention on the following demographic variables: - Percentage of population over 90 - Percentage of population between 25 and 65 (working age) - Sex ratio (Female/Male) at different ages We considered the oldest population sector as specifically interesting, since this region has a proper reputation of long-life land, due to some environmental factors: mild and dry climate, reduced density of population, absence of metropolitan areas etc. This reputation has been already checked by a series of studies, which suggest the presence of a restricted geographical zone (called Blue Zone), with a high concentration of centenarians in the East Central part of the region. In particular, the final part of human survival curve has been analysed and interpolated by a polynomial model, even considering the sex ratio (F/M) which is strongly increasing in the latest ages of life.

Education II

Team Teaching with Students: Experimental Study Part 2

Jerneja Šifrer, Zala Žvab and Matevž Bren

Faculty of Criminal Justice and Security, University of Maribor, Ljubljana, Slovenia
sifrer.jerneja@fvv.uni-mb.si, zala.zvab@gmail.com,
matevz.bren@fvv.uni-mb.si

Teaching can be a lonely and burdensome experience, but team teaching can transform it from a source of stress to a source of innovation and success' was the motto of studies on professor-professor or professor-student collaborative teaching that have been carried out and reported in the nineties. In our contribution an experiment study performed in the academic years 2010/11 and 2011/12 with the undergraduate students class of Statistics and professor-student collaborative teaching at the Faculty of Criminal Justice and Security, University of Maribor will be reported: comparison of this two and previous year students outcomes, the results of quantitative analysis of students questionnaire on theirs experience on team teaching, and qualitative analysis on several benefits to students, student teacher and professors. Hypothesis tested are that team teaching contribute to the better students outcomes, more collaboration and every day students' work and more positive attitude of students.

Internal and External Grading at Slovene Matura Exam: Differences in Distributions Shown by Ordinal Dominance Graphs

Darko Zupanc[1], *Gašper Cankar*[1] *and Matevž Bren*[2]

[1]National Examinations Centre, Ljubljana, Slovenia
[2]Faculty of Criminal Justice and Security, University of Maribor, Ljubljana, Slovenia
Darko.zupanc@guest.arnes.si, gasper.cankar@ric.si,
matevz.bren@fvv.uni-mb.si

Because there are important learning goals that cannot be assessed using paper and pencil tests, coursework and oral exams are also parts of Slovene Matura exam comprising 20-40% of total. But there is large variability in the achievement distributions between paper and pencil Matura exams graded externally compared to Matura coursework and oral exams in schools graded internally. We will present extremely high average scores for internal grading of Matura subjects' coursework, upward annual trends and also differences in the distributions (shown by ordinal dominance graphs) of three pairs of achievements (teachers' grades before Matura exams, externally graded paper and pencil exams and internally graded Matura subjects' coursework) for some schools in comparison to the entire country. Very different distributions for external and internal grading in schools are disturbing and iniquitous. Grading at the end of Gimnazija program must also discriminate, it must show differences among Matura candidates according to achieved goals and standards. We will show that as a rule high achievements at external parts is followed by low achievements at internal parts and also vice-versa low achievements at external parts is increased by high achievements at internal parts. The possibility for change is outlined, suggesting that internally graded coursework and oral exams should also be externally moderated.

R-Excel Tools for Homework Assignment and Exam Preparation

Lara Lusa

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
lara.lusa@mf.uni-lj.si

Preparing personalized home assignments for students or tests for written examinations can be a very time-consuming task for instructors. The preparation can be even more demanding if some stratagem has to be adopted to prevent students from cheating during homework preparation or written exams, for instance when multiple versions of same test/assigment are necessary.
This talk will present some newly developed R functions for the preparation of group and individual home assignments and present an extension of `genertest`, the R package previously developed for test preparation. `genertest` can now be invoked also from Microsoft ®Office Excel (Excel), where the end-user can specify how to generate the tests answering to twenty questions in "plain-English" contained in an Excel spreadsheet, and obtain the printable tests simply by invoking an Excel macro written in Visual Basic for Applications. To manage the connection between Excel and R we use the RExcel add-in for Excel together with the statconn server and rcom package, all of which were developed within the statconn project (Baier and Neuwirth, 2007, Excel :: Com :: R. *Computational Statistics*, 22:91–108).

Sampling Techniques and Data Collection

Challenges in Measurement of Sustainable Development in the Czech Republic

Jana Kramulova, Lenka Hudrlikova and Jan Zeman

University of Economics, Prague, Prague, Czech Republic
jana.kramulova@vse.cz, lenka.hudrlikova@vse.cz, janzeman06@gmail.com

The assessment of sustainable development (SD) is a great issue since the 70s of the 20th century. There are many indicators that have some relation to one of the three main pillars (economic, social and environmental). Then it depends just on individual, which of them are chosen for the particular analysis. There are plenty of indicators' sets issued by different institutions and scientists. One of them is also the set compiled by the Czech Statistical Office in 2008 (revised 2010) for measuring of sustainability in Czech NUTS 3 regions. In the Czech Republic there are only 14 regions at this level, therefore we decided to focus on the lower level (NUTS 4/LAU 1), where there are 77 regions. This number is much more suitable for further analyses (from their assumptions' point of view). On the other hand, the lower level is connected with the problem of non-availability of data. That was the reason, why we were forced to find different indicators available for LAU 1 level and suitable for SD analysis. Some indicators are the same as in the Czech Statistical Office publications. The aim of our contribution is to describe the process of data collection. Furthermore the data are analysed by using the multivariate statistical methods (for example cluster analysis).

Sampling Coordination of Business Surveys

Maruša Stanek and Boro Nikić

SURS, Ljubljana, Slovenia
marusa.stanek@gov.si, boro.nikic@gov.si

The Slovenian statistical office has undertaken the task of coordinating their samples for the different Business surveys based on the following motivations:

- Obligation to reduce the burden placed on respondents
- Decreasing budget
- Obtaining a centralised system for sample selection

The paper presents the current practice of the selection of samples at the Statistical Office of the Republic of Slovenia. In the paper the main idea of the coordination of samples methods and some advantages and disadvantages of the methods is discussed. We study different sampling techniques in order to reduce the response burden in business surveys at the Statistical Office of the Republic of Slovenia. The main objective of sampling coordination is to spread the response burden among the businesses by reducing the overlap between chosen samples of different surveys. This can be done quite effectively among small businesses, but there is not much space for spreading the response burden among larger businesses which have big impact on the estimates and therefore often have to be included in samples. In the year 2011 simulation of the coordinated samples were made and the response burden of coordinated sampling is compared to the burden of stratified simple random sampling in this year

SPAMIA: Spam Filtering by Quantitative Profiles

Marian Grendár , Jana Škutová and Vladimír Špitalský

Slovanet, Bratislava, Slovakia
marian.grendar@slovanet.net, jana.skutova@slovanet.net,
vladimir.spitalsky@slovanet.net

Traditional content-based approaches to spam filtering and email categorization are based on heuristic rules, naive Bayes filtering and text-mining methods, which employ bag-of-words representation of emails. In the quantitative profiles (QP) approach, an email is represented by a p-dimensional vector of numbers. On a private corpus the QP-based Random Forest classifiers attain in spam filtering comparable and in email categorization even better performance than the optimized SpamAssassin and Bogofilter. The objective of the presented work is to assess performance of new quantitative profiles on several widely used publicly available email corpuses.

An Algorithm for Computing a Combined Interestingness Measure in Association Analysis

Derya Ersel and Süleyman Günay

Hacettepe University, Ankara, Turkey
dtektas@hacettepe.edu.tr, sgunay@hacettepe.edu.tr

Association analysis is one of the descriptive models used in data mining. Aim of this analysis is to determine interesting patterns that help decision making by identifying items that are seen together in the data set. A drawback of association analysis is that many patterns emerge even if the data set is very small. Since real databases contain many items, millions of patterns can be created and most of these patterns are not interesting. Hence, the quality of resulting patterns are evaluated according to some measures and uninteresting patterns are eliminated in association analysis. To perform this evaluation, suitable measures must be determined with respect to structure of the data, structure of the pattern and expert opinion (or prior knowledge) about the data. In literature, interestingness measures are mainly seperated into two groups. The first group is "objective interestingness measures" which depend only on structure of the pattern and the data. These measures use statistics from the data and they are usually calculated depending on frequencies given by a contingency table. The second one is "subjective interestingness measures" which depend on expert opinion as well as structure of the pattern and the data. Subjective interestingness measures are generally defined over belief systems. Bayesian networks are belief systems and they can be used to identify these measures. Objective and subjective interestingness measures may not always be sufficient to determine interesting patterns. Therefore, a combined measure must be obtained. In this study, an algorithm for computing a combined measure that integrates objective and subjective interestingness measures defined over Bayesian networks is proposed to determine interesting patterns more accurately in association analysis.

Ratio-type Estimator of Population Total with Minimum MSE: A Computational Approach

Mohammadreza Faghihi and Zahra Noori

Department of Statistics, Shahid Beheshti University, Tehran, Iran
m.faghihi@sbu.ac.ir, statnoori89@yahoo.com

Using auxilary variables could be very useful in total (or mean) estimation. Ratio-type estimators are the most common estimators to estimate total. In this paper, families of the estimators for estimating population total, which use of known values of some population parameters, are investigated. For these families, the approximate mean square error (MSE) is a criterion to decide about the best estimator. In this investigation, the the optimum cases are discussed. Also some well known ratio-type estimators have been shown as particular member of these families. Finaly, a computaional approach to find the optimum vales of parameters is introduced and by using rice fields data of Amol county of Iran, an empirical study is carried out to show which estimator is the best. In this manner, since the values of (MSE) of estimators are approximate, to select the best estimator, average squares of errors (ASE) is used.

Invited Lecture

Multi-state Models: A Variety of Uses

Vern Farewell

MRC Biostatistics Unit, Cambridge, United Kingdom
vern.farewell@@mrc-bsu.cam.ac.uk

The use of multi-state models in the analysis of longitudinal data has increased substantially in recent years. This has been facilitated by the availability of computer software but also reflects their usefulness in the specification of data structures and their flexibility. The use of multi-state models for a variety of problems will be illustrated to demonstrate these characteristics. These problems will involve potentially informative observation patterns, the challenges of panel data, time to event analyses for events defined only by prolonged observation and correlated processes. The application of causal reasoning in the context of multi-state models will also be briefly discussed.

Biostatistics and Bioinformatics I

Development of a Flexible Excess Hazard Model with Shared Frailty, Non-Linear and Time-Dependent Effects of Covariates

Hadrien Charvat, Laurent Remontet, Nadine Bossard, Laurent Roche, Jacques Estève and Aurélien Belot

Service de Biostatistique des Hospices Civils de Lyon, Lyon, France
hadrien.charvat@chu-lyon.fr, laurent.remontet@chu-lyon.fr, nadine.bossard@chu-lyon.fr, laurent.roche@chu-lyon.fr, jacques.esteve01@chu-lyon.fr, aurelien.belot@chu-lyon.fr

The excess hazard approach allows the estimation of disease-specific mortality hazard when the cause of death is unavailable or unknown; this is the standard approach to analyse population-based cancer registry data. This approach relies on the assumption that the total mortality hazard observed in the study population can be described as the sum of the disease-specific mortality hazard (i.e. the excess hazard) and the other-causes mortality hazard; this latter is obtained from general population life tables. However, observed data are collected from different geographical units (here, the département) and patients from the same geographical unit may share some characteristics (availability of health resources, medical practices, etc.). In other words, they share a common "frailty" towards their disease leading to correlated times to event data. Our objective is to propose an approach to fit a flexible excess hazard model including a random effect at the département level. The baseline excess hazard was modelled using cubic regression splines and the model took into account, if necessary, non-linear and time-dependent effects of covariates, such as age at diagnosis. The random effect was assumed to follow a normal distribution. We developed a R function to calculate the likelihood using adaptive Gaussian quadrature and maximum likelihood estimates were obtained by a standard Newton-based optimisation procedure. We conducted a simulation study to evaluate the performances of the proposed approach under several conditions (number of clusters, heterogeneity of cluster sizes). We also illustrated our approach on real data from the French cancer registry network.

On Comparison of Measures of Explained Variation in Survival Analysis

Nataša Kejžar[1], *Delphine Maucort-Boulch*[2] *and Janez Stare*[1]

[1]Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[2]Service de Biostatistique des Hospices Civils de Lyon, Lyon, France
natasa.kejzar@mf.uni-lj.si, delphine.maucort-boulch@chu-lyon.fr, janez.stare@mf.uni-lj.si

Papers introducing new measures of explained variation in survival analysis, and those comparing them, discuss bias in the presence of censoring as the most important property. Measures with little or no bias under censoring are considered to be useful, the rest dismissed. In our opinion the bias is misunderstood. Measures, seemingly being unbiased, are assuming something unverifiable, and other measures, considered to be biased, are easily corrected under the same assumption. Properties of measures are usually illustrated, or even studied, using simulations, and we argue that because of the above, these simulations give wrong results, which has, unfortunately led to quite some misunderstanding of the properties of such measures.

Properties of a Net Survival Estimator

Maja Pohar Perme

Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
maja.pohar@mf.uni-lj.si

Survival analysis of long term studies is often interested in mortality due to the disease in question but faced with the problem of many deaths occurring due to other causes. Furthermore, the cause of death is often unknown or unreliable. A solution to this problem is to assume that hazard due to other causes can be described by the general population mortality, which enables us to deduct the value of the disease related hazard. The methodology based on this assumption is referred to as relative survival, its most important field of usage is cancer registry data. One of the basic aims of the analysis of cancer registry data is to estimate quantities which are comparable between different regions, countries or time periods and thus not affected by the differences in other cause mortality. We have recently shown that the methods in standard use provide biased estimates and proposed a new measure of net survival that satisfies this aim. In this work, we study its properties and behaviour in practice and discuss its assumptions and interpretation. The results are illustrated using Slovene cancer registry data.

Dynamic Survival Model for Clustered Multiple Failure Time Data

Amirhossein Jalali[1], *Firoozeh Haghighi*[2], *Zahra Rezaei Ghahroodi*[3] *and Shirin Moghaddam*[1]

[1]University of Tehran, Tehran, Iran
[2]Department of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran
[3]Statistical Research and Training Center, Tehran, Iran
jalali.stat@gmail.com, Haghighi@khayam.ut.ac.ir, z_rezaee@sbu.ac.ir, shirin6539@yahoo.com

In most clinical trials such as medical or dental researches, it is often too expensive or even impossible to follow up the subjects continuously. In these cases, subjects are examined periodically at some pre-scheduled visits, so multiple failure times are obtained from the same patient or subject, which are called clustered multiple failure time data. Since in a long term controlled clinical trial, the effects of treatment or covariates on the survival time may change, the assuming the treatment or covariates effects to be constant over time is not appropriate, so the time-varying effects should also be considered. This paper extends the random effects logistic regression models by incorporating the possibly time-varying covariate effects into the model in terms of a state-space formulation. Due to complexity of the models, it is impossible to estimate the parameters of models by Maximum likelihood method. Hence, a Bayesian approach via Gibbs sampling is used to perform parameter estimation. By Deviance Information Criterion (DIC), the results of different models are compared. Also some sensitivity analyses are performed to assess robustness of the posterior estimation of the transition parameters to the perturbations of the prior parameters.

Biostatistics and Bioinformatics II

Spatio-Temporal Modelling of Daily Rainfall

Ana F. Militino, Maria D. Ugarte and Manuel Garcia-Magariños

Universidad Publica de Navarra, Pamplona, Spain
militino@unavarra.es

Spatio-Temporal modelling of climate data is a useful tool in many environmental, climatological or biological applications. Different models and estimation alternatives have been provided in the literature depending on the auxiliary information and the quality of data, but mostly oriented to study temporal trends, providing general forecast or explaining climatological changes.
When daily rainfall information is required in a specific location as an input to decision-making tools in precision agriculture, the performance of these models can be good on regular days. However, a great variability is present some days because of local storms or dramatically changing weather. In this work, we propose the use of a dynamic state-space model that incorporates daily information and specific spatial cluster modelling in locations with high precipitation values. The model will be illustrated with real data using daily precipitation in 20 years for the 80 sampled manual rainfall gauges in Navarre (Spain). The results will be checked with 60 automatic rainfall gauges in the same region but in different locations.

Benefits and Caveats of Massive Microarray Data Production

Ana Rotter, Urška Tajnšek, Kristina Gruden, Helena Motaln and Tamara Lah Turnšek

National Institute of Biology, Ljubljana, Slovenia
ana.rotter@nib.si, urska.tajnsek@nib.si, kristina.gruden@nib.si, helena.motaln@nib.si, tamara.lah@nib.si

Microarrays have become almost a standard technology in molecular biology experiments. The last 10 years have focused on improvements both on microarray technology and experiment conduction part as on the data analysis part. Alongside, various standards for microarray experiment description and data format, gene ontologies and data dissemination repositories have emerged and nowadays it is possible to share, distribute, search, reuse and reanalyze data from practically all groups conducting similar research. This is of substantial benefit in various scenarios. For example, if studying the effect of drug A on organism X, one might be interested in research and results of the effects of the same drug but using different concentrations on the same organism or on the organism Y. Also, in order to properly design the experiment, some background (re)search is often done in order to see preliminary results from other groups. Sometimes, due to financial, biological or time restrictions, data from different experiments (inhouse or from data repositories) is merged in order to have a larger sample size. Sometimes, an appropriate microarray dataset is selected in order to have real biological data that complement simulations confirming a new statistical testing procedure. Merging datasets coming from different microarray manufacturers or even different versions of the same user is anything but trivial. First of all, it is impossible to determine the quality of data without having to conduct relatively long quality control tests. Then, even when having two similar experiments done on different microarray platforms, the number of gene spots, identifiers and gene annotation differs. These benefits and caveats have been approached in dealing with transcriptome of brain tumor tissues in comparison with the nonmalignant tissue conterparts, where different platforms have been merged. The data analysis approach taken and the conclusions will be described and disscussed.

Improved Shrunken Centroid Classifiers for High-Dimensional Data

Rok Blagus and Lara Lusa

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

Nearest shrunken centroid method has been successfully applied to many classification tasks where the number of variables greatly exceeds the number of samples (high-dimensional data). It was shown however that the method has drawbacks as it shrinks each variable by the same amount. This can be problematic for high-dimensional data as they are often a mix of variables that distinguish between the classes (alternative variables) and variables that are not important for classification (null variables); two additional methods of shrunken centroids estimation were proposed to solve this issue. In shrunken centroid classifiers it is necessary to define a tuning parameter that determines the amount of shrinkage. The approach currently used in practice depends on the overall cross-validated (CV) error rate. We show that this approach combined with the effect of the prior correction leads to poor accuracy for the minority class when the number of samples in each class (or some of the classes) is different (class-imbalanced data). Based on that finding we propose a novel approach for the determination of the optimal tuning parameter. Instead of using the tuning parameter that achieves the lowest cross-validated error rate, we suggest the tuning parameter that achieves the highest cross-validated g-means; g-means is the geometric average of class-specific predictive accuracies. We use simulated and real high-dimensional data and show that our approach outperforms the original approach when data are class-imbalanced, and performs very similarly when data are class-balanced. The difference between the approaches is larger when the level of class-imbalance is large and/or when the difference between the classes is smaller.

Statistical Applications - Economics I

Testing Export-Led Growth Hypothesis: The Case of Turkey, 1961-2010

Mustafa Murat Arat

Department of Statistics, Hacettepe University, Ankara, Turkey
muratarat@hacettepe.edu.tr

The purpose of this paper is to test the export-led growth hypothesis, which identifies export growth as a major source of economic growth, for Turkey. In order to test this hypothesis, we investigate the short-run and the long-run relationship between export, import and economic growth. We employ co-integration and error-correction modeling over annual observations for the period of 1961-2010. The economic growth is represented by Gross Domestic Product (GDP) that is collected from Central Bank of Turkey. Export and import variables are obtained from the publication "Statistical Indicators 1923-2010" released by TurkStat. All the variables in domestic currencies are deflated by appropriate indexes, to obtain the real values. Before undertaking co-integration analysis, Augmented Dickey-Fuller test was conducted to see whether all the time series are stationary or not. Since all the variables are integrated of the same order, we examine long-run equilibrium relationship between the variables. Based on the trace test and maximum eigenvalue test, it is seen that one co-integration equation exists between real GDP, real export and real import. In the presence of co-integration, using Standard Granger causality will give misleading results. Therefore, we proceed to the construction of Vector Error Correction Models (VECM). Based on the VECM, we found the evidence of bi-directional causality between export and GDP, uni-directional causality running from GDP to import and bi-directional causality between export and import. In conclusion, it is clearly seen that export-led growth hypothesis is valid for Turkey between the years, 1961-2010, and we proudly suggest that promoting exports via export promotion policies will contribute to the economic growth of Turkey.

Water Utilization, Water Quality in the Endogenous Economic Growth: Reflections On Fixed Effects

Souha El Khanji and John Hudson

University of Bath, Bath, United Kingdom
sek25@bath.ac.uk, j.r.hudson@bath.ac.uk

In previous study, we estimated the effect of the rate of water utilization and water quality on endogenous economic growth, using BOD as a water quality indicator. We used the panel data analysis, panel data attractiveness here based on different factors, one of which is the gained precision in estimation, particularly from using the fixed effects estimations that allow for the unobserved individual heterogeneity which potentially correlate with the regressors. We know fixed effects reduce potential bias, or at least that is the literature view. But the differences were surprisingly large in the coefficients in OLS, FE and RE estimators and this gave cause for concern and reflection. Here we endeavour to ascertain and support the regression analysis framework that took place before. To investigate the great difference in the coefficients in the panels of regression in part (A) of the study from fixed and random effects, we reflect on the nature of fixed effects. We argue that this involves an implicit, seldom stated and never tested assumption that the impact of the country mean of a variable X is the same as the impact of deviations from that mean within a regression context. This is something we test for and suggest an alternative approach which in many respects combines fixed and random effects.

SMEs and Fast-Growing Companies in Survey Estimates

Aleša Lotrič Dolinar[1], *Rudi Seljak*[2] *and Mojca Bavdaž*[1]

[1]University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia
[2]Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia
alesa.l.dolinar@ef.uni-lj.si, rudi.seljak@gov.si,
mojca.bavdaz@ef.uni-lj.si

Large companies are usually selected into samples with certainty because of their high impact on aggregate figures. The contribution of SMEs (small and medium enterprises), on the other hand, is small with respect to their number, so data collection from SMEs may even be completely abandoned below a certain threshold. This may be problematic at least for fast-growing SMEs (the so-called gazelles) that may quickly become big players and thus have a significant impact on aggregate figures. Research shows that even if the gazelles are rare they account for most of the new jobs. The present paper will compare several sampling designs regarding the inclusion of two groups of companies, SMEs and fast-growing companies, observing the consequent impact on survey estimates (in total and by activity). Thus, it will try to answer the question whether surveys which do not pay attention to fast-growing SMEs produce biased estimates of some economic categories and their growth. The data used will be yearly and quarterly data on taxable revenue of Slovenian companies from January 2002 to December 2010. It will also be checked whether the current global economic crisis has any different impact on the performance of gazelles compared to the other enterprises.

Invited Lecture

Data Analysis: Best Practices and Future Directions

Hadley Wickham

Rice University, Houston, TX, United States of America
hadley@rice.edu

What are are best practices and what will data analysis look like in10 years time? I'll start by discussing what I think are current best practices for data analysis (combining ideas from good science and software development), then look at how things might change in the near and not-so-near future. I'll highlight today's projects that I think are really exciting (Rstudio, D3, amazon's EC2), and do a little blue-sky speculation about what's on the horizon. I'll discuss the new field of "data science" and give some hints about what technologies you should be learning next

Social Science Methodology

Advances in Methods for Causal Inference of Observational Studies

Ana Kolar and Vasja Vehovar

University of Ljubljana, Ljubljana, Slovenia
annakolar@yahoo.com, vasja.vehovar@fdv.uni-lj.si

Observational studies, which derive from non-randomized selection procedure, have always been a tough nut if causal quantities are to be estimated. It was long advised that within non-randomized settings we can only provide descriptions of observed associations (Cochran W. G., Journal of the Royal Statistical Society, 128, 1965) while not being able to talk about any causal quantities. The paper presents the most advanced methods for estimating causal effects of observational data – the Propensity Score Methods and explains why the use of different correlation/regression methods, in cases of observational data, is conceptually problematic and can thus lead into very misleading estimates of causal quantities.

Post-Stratification Weighting Issues in Cross-National Survey Research

Ana Slavec and Vasja Vehovar

Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia
ana.slavec@fdv.uni-lj.si, vasja.vehovar@fdv.uni-lj.si

One of the methods to study and correct unequal representation of different socio-economic groups among respondents in the European Social Survey (ESS) is post-stratification weighting. The size of post-stratification weights is an indicator of non-response bias and, as earlier studies show, it differs substantially across countries. Thus, it is very important to have relevant and reliable post-stratification variables which need to be used consistently across countries. The task is complex due to a large amount of surveys (five rounds, each including from 20 to 30 countries) which need to be weighted with internal and external consistency. Each country has its own national specifics (e.g. different education systems) and methodological issues (identifying proper control data, treating missing values, etc.) which make comparative research very challenging. In our paper, we present the development of the methodology to post-stratify ESS data with four variables: age, gender, education and region, where Labour force survey (LFS) was selected as a control source for several reasons. Data preparation included also modifications and recoding to account for specifics in different education systems. Our standardized procedure and strategy for post-stratification weighting considers also the problem of missing values in sample and control data. We describe and analyze ESS post-stratification weights for the current and past rounds.

Internet Forums: What Information is Out There?

Vanja I. Erčulj

ro sigma, Ljubljana, Slovenia
vanja.erculj@gmail.com

Various qualitative and quantitative research methods are employed for assessment of people's opinions and beliefs. In-depth interviews and focus groups approaches are the most frequently used qualitative methods, but require respondents' time and willingness to answer the questions. With increasing number of surveys being conducted and with busy life-style, people are less willing to participate in surveys examining as trivial issues as e.g. dish washing detergent usage. On the other hand, participants are quite eager to share opinions on various topics of their interest on the internet forums. Many topics being discussed on such forums range from books having been read to specific medical issues. Thus, internet forums offer an interesting, extensive and accessible data source that can be harvested prior to conducting invasive surveys and provide an excellent starting point for the latter. Information that can be gathered from this source will be presented and illustrated on discussions about dietary supplements usage that took place on 31 different Slovenian forums in 2011.

Clustering Macroeconomic Variables

Chiara Perricone

University of Rome II - TorVergata, Rome, Italy
chiara.perricone@gmail.com

Many papers have highlighted that some macroeconomic time series present structural instability. The causes of these remarkable changes in the reduced form properties of the macroeconomy is a debated argument: Bad Policy or/and Bad Luck. In literature this issue is handled with three main econometric methodologies: structural breaks, regime-switching and time-varying parameters (TVP). Nevertheless all these approaches need some ex ante structure in order to model the change. Based on the Recurrent Chinese Restaurant Process, I have specified a model for an autoregressive process and estimated via particle filter using a conjugate prior, which applied the idea of evolutionary cluster to the study of the instability in output and inflation for US after War World II. This procedure displays some advantages, in particular does not require a strong ex ante structure in order to neither detect the breaks nor manage the parameters' evolutions. The application of the cluster procedure to GDP growth and inflation rate for US from 1957 to 2011 shows a good ability in fit the data, moreover it produces a clusterization of the time series that could be interpreted in terms of economic history and it is able to recover key data features without making restrictive assumptions, as in 'one-break' or TVP models. Considering the open debate on the source of the Great Moderation, under the caveat that until now I do not study a VAR or a structural form, this approach presents conclusions closed to Cogley and Sargent, suggesting changes in both volatility and coefficients, even if the latter are less marked.

Modeling and Simulation I

Generating Random Numbers from Empirical Distributions

Katja Rostohar[1], *Anja Žnidaršič*[2], *Jelka Šuštar Vozlič*[1] *and Andrej Blejec*[3]

[1]Agricultural Institute, Crop and Seed Science Department, Ljubljana, Slovenia
[2]Faculty of Organizational Sciences, University of Maribor, Maribor, Slovenia
[3]Department of Entomology, National Institute of Biology, Ljubljana, Slovenia
katja.rostohar@kis.si, anja.znidarsic@fov.uni-mb.si,
jelka.vozlic@kis.si, andrej.blejec@nib.si

For statistical simulations one usually generates random values that follow some known distribution. If underlying theoretical distribution is not known, it can be replaced by the empirical distribution of the sampled measurements. Our goal was to generate random numbers using the distribution of a sample. We analyzed the properties of generated random samples from empirical data, where we used three methods: bootstrapp method (BSM), rejection sampling method (RSM) and inversion sampling method (ISM). In the first step, we generated the empirical data from original known distribution (normal and chi-square). In the second step, according to the generated empirical data we generated larger samples using the BSM, RSM or ISM method. In the third step, we compared the results and distributions of original data, empirical data and generated samples. We will present the characteristics of investigated methods, the simulation results and discuss the application of each method in practice.

A Proposition of a Hybrid Stochastic Lee– Carter Mortality Models

Leslaw Socha and Agnieszka Rossa

University of Lodz, Lodz, Poland
leslawsocha@poczta.onet.pl, agrossa@uni.lodz.pl

The problem of the determination of the best mortality models is one of the basic fields in the forecasting strategy of insurance companies. There are several methods of parameter estimation of mortality models. This problem was widely studied in the literature. Some researchers have observed that the estimate of parameters for the same model depend on the year, for instance, the estimate of parameters during 1930-1950 gives quite different results of parameters compared with those of 1960 -1980 [R. Giacometti, S. Ortobelli, M. Bertocchi, A stochastic model for mortality rate on Italian Data, J. Optim. Theory Appl., (2011), 149, 216-228]. Therefore we would like to propose a new philosophy of constructing of a mortality model that will take into account the changes of estimated parameters. We will use the methodology which has been already used in control theory, economics, biology, chemistry and called „stochastic dynamic hybrid or switched systems", which are dynamic systems consisting of several structures described by deterministic or stochastic differential equations. In the successive moments of the time their structures can change according to the given switching rule thereupon creates the hybrid system. To model subsystems of the proposed hybrid system the Lee–Carter model with different sets of parameters will be used. The obtained results will be illustrated by numerical calculations.

Bootstrap Confidence Intervals of the Difference between Two Process Capability Indices for Half Logistic Distribution

Somchit Wattanachayakul and Wararit Panichkitkosolkul

Thammasat University, Phathum Thani, Thailand
somjit@mathstat.sci.tu.ac.th, wararit_tu@hotmail.com

The process capability indices are important numerical measures in statistical quality control. Well-known process capability indices are constructed under the process distribution is normal. Unfortunately, this situation is rather not realistic. This paper focuses on the half logistic distribution. The bootstrap confidence intervals for the difference between two process capability indices for the mentioned distribution are proposed. The bootstrap confidence intervals considered in this paper consist of the standard bootstrap confidence interval, the percentile bootstrap confidence interval and the bias-corrected percentile bootstrap confidence interval. A Monte Carlo simulation has been used to investigate the estimated coverage probabilities and average widths of the bootstrap confidence intervals. Simulation results showed that the estimated coverage probabilities of the percentile bootstrap confidence interval and the bias-corrected percentile bootstrap confidence interval get closer to the nominal confidence level than those of the standard bootstrap confidence interval.

Defective Rate Control Charts for Zero-Inflated Processes

Chanaphun Chananet and Piyachat Leelasilapasart

King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
chanaphunc@kmutnb.ac.th, piyachatl@kmutnb.ac.ht

In this research aim to study the performance of fraction defective control charts when the probability of observing defective follows a Zero – Inflated Binomial (ZIB) distribution, i.e., a distribution which is similar to Binomial distribution but with a defective number of zeros. The Monte Carlo simulation is used to study the performance of defective control charts: p – chart, Moving Average Control Chart, and Binomial Exponential Weighted Moving Average: Binomial EWMA for processes with varying proportions of observed fraction defective zeros, namely = 0.3, 0.35, 0.4, 0.45, 0.5 and 0.6. We looked at processes in with and without fraction defectives shifts. The Monte Carlo simulation is used to study the performance of control charts by given the trajectory samples is 10,000 times and Average Run Length: ARL is used to be criteria of consideration.

Statistical Applications - Economics II

The Log-Periodic Power Law model for financial bubbles with ARMA/GARCH errors

Špela Jezernik Širca

Faculty of Mathematics and Physics, Ljubljana, Slovenia
spela.jezernik@gmail.com

The aim of this paper is to check adequacy of the Log-Periodic Power Law (LPPL) model from econometric point of view. Prior to crashes, the mean function of a stock market index price time series is characterized by a power law decorated with log-periodic oscillations, leading to a critical point that describes the beginning of the new market regime. First, we investigate the residuals of the LPPL model. Second, we apply an extended autocorrelation functions (EACF) method and Akaike's Information Criteria (AIC and BIC) for model identification of residuals of the LPPL model. We incorporate an autoregressive moving average (ARMA) and a conditional heteroskedasticity (GARCH) structure in the error term of the LPPL model. Third, we estimate the highly nonlinear models, the LPPL model and the LPPL model with ARMA/GARCH errors, using the Differential Evolution algorithm, which performs evolutionary global optimization.
The main contribution of our analysis is that we get better statistical properties of residuals using the extended LPPL model with ARMA/GARCH errors and improved estimate of the most crucial nonlinear parameter, the critical time t_c, defined as the end of the bubble and the most probable time for a crash to occur. Both the original and the extended models are fitted to different stock market indices of major financial markets, namely S&P 500, DAX, Dow Jones, NASDAQ and Hang Seng index.

On Longevity Risk Models in the Romanian Annuity Market and Pension Funds

Iulian Mircea and Mihaela Covrig

The Bucharest University of Economic Studies, Bucharest, Romania
iulian.mircea@csie.ase.ro, mihaela_covrig@yahoo.com

One of the largest sources of risk faced by life insurance companies and pension funds is the longevity risk: members of some reference population might live longer on average than anticipated. This affects their pricing and reserving calculations. In last years, many companies have closed the defined benefit retirement plans that they used to offer to their employees. In addition, some governments, among them the Romanian government, increased the retirement age by two or five years to take into account longevity improvements, population ageing and the retirement funding. It has become more important for insurance companies and pension funds to find efficient ways to transfer part of the longevity risk to reinsurers or to financial markets. As a consequence, the markets for longevity derivatives are starting to develop. In this paper we discuss models of mortality rates and pricing the longevity risk. We make some remarks on the results in forecasting mortality rates using various models. Finally, we deal with the securitization of longevity risk through the longevity bonds. In this regard, the Special Purpose Company is the instrument in splitting the interest between the annuity provider and the investors.

Comparison of Artificial Neural Networks, Autoregressive Model and Multiple Linear Regression for Monthly Streamflow Estimation

Meral Büyükyıldız and Tezel Gülay

Selcuk University, Konya, Turkey
meralbyildiz@selcuk.edu.tr, gtezel@selcuk.edu.tr

Estimation of streamflow is essential for planning, design and management of water resources system. The purpose of this study was the modeling of the monthly streamflows of the Bayburt gauging station (No:2304) on Çoruh River operated by General Directorate of Electric Power Research Survey and Development Administration EIE in Turkey. The observed monthly data are 59 years (708 months) long with observation period between 1942 – 2000. To forecast monthly streamflows were used Autoregressive (AR), two different Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) models. Monthly streamflows of Bayburt Station between 1942-1988 (564 variables) and between 1989-2000 (144 variables) were used for training and testing in ANN models, respectively. To evaluate the performance of the recommended models, various statistical measures were used, namely; mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE) and coefficients of determination (R2).
Finally, AR(4) model with a determination coefficient of 0.85 performed much better than the ANN and MLR models in monthly streamflow estimation.

Comparing the Markov Switching AR, Nonlinear Additive AR, Self-Exciting Threshold AR and Logistic Smooth Transition AR models for Analysis Time Series Data With Dramatic Jumps

Masoud Yarmohammadi

Payame Noor University, Tehran, Iran
masyar@pnu.ac.ir

Many economic time series associated with events such as financial crises, war or change in government money policy exhibit dramatic jumps in their behavior. When jumps arise in time series data, a powerful tool will up to date themselves using a change in their regime is the Markov switching models. This model will offer a better statistical fit to the data than the other models. In this research the Markov switching autoregressive model and three different time series models such as the nonlinear additive AR, self-Exciting threshold AR, logistic smooth transition AR models are introduced. These models are compared according to their performance for capturing the Iranian exchange rate series. The series has dramatic jump in early 2002 which coincides with the change in policy of the exchange rate regime. Our criteria are based on the AIC and BIC values. The results indicate that the Markov switching autoregressive model can be considered as useful model, with the best fit, to evaluate the behaviors of Iran's exchange rate.

Econometrics

Insurance Markets in the Long Run

Giovanni Millo[1] *and Gaetano Carmeci*[2]

[1]Generali Research and Development, Trieste, Italy
[2]DEAMS, University of Trieste, Trieste, Italy
giovanni_millo@generali.com, gaetano.carmeci@econ.units.it

Two questions are on top of the list for insurance market analysts: whether a country's market will grow and whether it will be profitable. The counterparts in academic literature are the ongoing debates on the income elasticity of insurance, on one side, and the "insurance cycle", i.e. the characterization of the regular pattern of ups and downs in market profitability, on the other. Empirical studies have usually drawn on panels of countries, the reliability of their outcomes being challenged by non-stationarity and heterogeneity problems and by the possibility of omitted, cross-sectional correlation inducing common factors.

Recent methodological progress in the field of panel time series, centering around the work of Hashem Pesaran and coauthors and based on the idea of Common Correlated Effects augmentation, allows consistent estimation of parameters in the presence of unit roots and cross-sectional dependence, as well as consistent testing of the stationarity hypothesis which is of the utmost importance in cointegration analysis. The availability of these tools calls for a reassessment of some established results.

We reconsider two fundamental relationships characterizing insurance markets: the first relationship is the one between the development of an insurance market in the long run with respect to GDP; the second, the long-term equilibrium between the sources of income (premiums and financial revenues) on one side, and costs (claim costs and general expenses) on the other. We draw on a long panel of countries, and on the two variants of Pesaran's Common Correlated Effects estimator: Pooled (CCEP) and Mean Groups (CCEMG). We investigate the cointegrating behaviour of the model, looking for: the existence of an equilibrium; a characterization of the insurance cyce through the speed of adjustment; cross-sectional (and possibly spatial) dependence between individual markets and its sources.

On the Proportional Retention Reinsurance Problem

Mihaela Covrig, Daniela Todose, Emilia Titan and Simona Ghita

The Bucharest University of Economic Studies, Bucharest, Romania
mihaela.covrig@csie.ase.ro, daniela.todose@csie.ase.ro,
emilia.titan@csie.ase.ro, simona.ghita@csie.ase.ro

Sharing the risk is a common practice in insurance and finance. For the insurance industry, it implies reinsurance. Proportional reinsurance is employed when the ceding company and the reinsurwer set a cession percentage for each risk in a given portfolio. The premiums and the claims are splitted between the two companies accordind to these proportions. De Finetti (Il problema dei pieni, Giornale dell'Instituto Italiano degli Attuari, 1940, vol 11, pp 1-88) solved the problem of determining the optimal retention levels with a mean-variance approach. We discuss the same problem, giving another solution which sets down which retention levels are exactly equal to 1, i. e. full risk transfer, and which are less than 1. We investigate this solution in a case study.

Hybrid Fuzzy Mortality Models of LC-type: A Simulation Study

Andrzej Szymański and Agnieszka Rossa

University of Lodz, Lodz, Poland
anszyman@math.uni.lodz.pl, agrossa@uni.lodz.pl

For a given age group x at year t the mortality rate x(t) can be expressed in the form of so-called Lee-Carter stochastic mortality model (LC). On the other hand, it can be easily seen that the LC model is the solution of the stochastic differential equation of the Black-Sholes type.
Koissi and Shapiro have formulated the fuzzy version of the LC model (FLC), where the model coefficients are assumed to be fuzzy numbers with the symmetric triangular membership function (STMF).
Simulation study gives us exact solutions of the stochastic differential equation for simulated Wiener standard process. Hence we can determine the histogram of the exact solution and then the membership function when treated the solution as a fuzzy function. Our simulation study enables us to verify the assumption that membership function in FLC model is of the type STMF. We have applied the logistic function to approximate empirical membership function obtained during the simulation process.
Using this model for the data from Poland for the period of years 1990-2009 we have observed parameter changes during different time subperiods. We applied the methodology used in control theory called stochastic dynamic (or switched) systems, which are dynamic systems with changing structure. To estimate parameters of the proposed stochastic hybrid system we use one of the literature methods.
To make more precise and elegant inferences from the improved FLC we apply two Banach algebras of fuzzy numbers, first one called OFN-algebra introduced by Kosiński et al. and the second one called C*-algebra introduced by Ishikawa. The difference between these two algebras is in the multiplication operation definition. It will allow us to create two mortality models called AFLC and C*FLC.
On the other hand Lee-Carter model can be generalized to the so-called hybrid (or switching) Lee-Carter model (HFLC). Combining ideas of FLC and HFLC we propose the bybrid fuzzy mortality model and verify its switching rule in the simulation study.

A Modified Weighted Symmetric Estimator for a Gaussian First-Order Autoregressive Model with Additive Outliers

Wararit Panichkitkosolkul and Patarawan Sangnawakij

Thammasat University, Phathum Thani, Thailand
wararit@mathstat.sci.tu.ac.th,

It is well-known that the effect of outliers may cause serious bias in estimating autocorrelations, partial correlations, and autoregressive moving average parameters. This paper presents a modified weighted symmetric estimator for a Gaussian first-order autoregressive (AR(1)) model with additive outliers. We apply the recursive median adjustment based on an exponentially weighted moving average (EWMA) to the weighted symmetric estimator. We consider the following estimators: the weighted symmetric estimator (W), the recursive mean adjusted weighted symmetric estimator (RW), the recursive median adjusted weighted symmetric estimator (RDW), and the weighted symmetric estimator using adjusted recursive median based on EWMA (RD-EWMA). Using Monte Carlo simulations, we compare the mean square error (MSE) of estimators. Simulation results have shown that the RD-EWMA estimator, provides a MSE lower than those of W, RW, and RDW estimators for almost all situations.

Mathematical Statistics

Confidence Intervals for a Ratio of Binomial Proportions Based on Direct and Inverse Sampling

Thuntida Ngamkham, Kamon Budsaba and Araya Chaemchan

Thammasat University, Pathumthani, Thailand
thuntida@mathstat.sci.tu.ac.th, kamon@mathstat.sci.tu.ac.th,
araya@mathstat.sci.tu.ac.th

A general problem of the interval estimation for a ratio of two proportions according to data from two independent samples is considered. Each sample may be obtained in the framework of direct or inverse binomial sampling. For each type of sampling scheme we construct asymptotic confidence interval based on unbiased estimations of success probabilities and also their logarithms. Various methods of confidence intervals construction in the situations when values for the both samples are obtained for identical sample schemes (for only direct or only inverse binomial sampling) were already developed and well known, so the main subject of our investigation is a construction of confidence intervals in two cases that correspond to different sampling schemes. In this situation it is possible to plan the sample size for the second sample according to the number of successes in the first and this, as it is shown by the results of statistical modelling, provides the intervals with confidence level that close to the nominal. Our goal is to show how reliable are normal approximations for the distributions of estimates of the ratio of proportions and their logarithms for a construction of confidence intervals. It is shown the preference of the scheme of inverse binomial sampling with planning of the size in the second sample. Main probability characteristics of intervals corresponding to all possible combinations of sampling schemes are investigated by the Monte-Carlo method. Estimations of coverage probability, expectation and standard deviation of intervals length has the form in tables and some recommendations for an application of each of the intervals obtained is presented.

Generalized Confidence Interval for the Difference between Normal Population Variances

Wichitra Phonyiem and Sa-att Niwitpong

King Mongkut's University of Technology North Bangkok , Bangkok, Thailand
wpy@kmutnb.ac.th, snw@kmutnb.ac.th

This paper we present a new confidence interval for the difference between two normal population variances based on the generalized confidence interval of Weerahandi [S. Weerahandi, Generalized Confidence intervals. Journal of the American Statistical Association, 1993, 88(423): 899-905.]. Monte Carlo simulation results indicate that the proposed confidence interval gives a better coverage probability than that of the existing confidence interval.

The Comparison of Tests for Equality of Coefficients of Variation for Data with Outliers

Piyachat Leelasilapasart and Chanaphun Chananet

King Mongkut's University of Technology North Bangkok , Bangkok, Thailand
piyachatl@kmutnb.ac.th, chanaphunc@kmutnb.ac.th

The coefficient of variation (CV), which is defined as a ratio of standard deviation to mean, is a dimensionless measure of dispersion generally used in the applied sciences and social sciences. It enables to compare the variability among populations with different tendencies and among populations with measurements carrying different units. The objective of this study is to compare testing the equality of coefficient of variation for two normal populations with outliers five methods; the likelihood ratio test (Lohrding,1975), Squared rank test (Miller, 1991), Modified Miller's asymptotic test (Feltz and Miller,1996), A generalized approach test (Jafari and Behboodian, 2010), and the new approximation (Jafari and Behboodian, 2010). The Monte Carlo simulation is used to study the efficiency of Type I Error and the power of the test.

Random Effect One-Way ANOVA Model when Sampling from a Finite Population of Treatment Groups

Kamon Budsaba[1], *Teerawat Simmachan*[1] *and John J. Borkowski*[2]

[1]Thammasat University, Pathumthani, Thailand
[2]Montana State University, Bozeman, United States of America
kamon@mathstat.sci.tu.ac.th, teerawat@grad.sci.tu.ac.th,
jobo@math.montana.edu

This study consists of two parts: the theoretical part and the computational part. The main focus of the theoretical part is to determine the expected value of the mean square error and the expected value of the treatment mean square of the random effect one-way ANOVA model assuming a finite population of treatment groups. For balanced data, both the expected values for the finite population are the same as that for the infinite population. For unbalanced data, the expected value of the treatment mean square for the finite population is different from that for the infinite population. because of some different multiplier values. The main purpose of the computational part is to assess the impact of sampling from a finite population of treatment groups on hypothesis testing. The results suggest that when the null hypothesis is true, the F-ratio will still follow a (central) F-distribution. However, if the number of levels in population of treatment groups is not large enough relative to the number of randomly selected treatments, it can have a large value of the Type II Error. It also appears that, when the null hypothesis is false, the F-ratio will not follow a non-central F-distribution.

Modeling and Simulation II

Smooth Bootstrap Inference for Parametric Quantile Regression

Tatjana Kecojevic[1] *and Peter Foster*[2]

[1]University of Central Lancashire, Preston, United Kingdom
[2]University of Manchester, Manchester, United Kingdom
tkecojevic@uclan.ac.uk, peter.foster@manchester.ac.uk

Assessing the accuracy of the τth (τ[0,1]) quantile parametric regression function estimate requires valid and reliable procedures for estimating the asymptotic variance-covariance matrix of the estimated parameters. This covariance matrix depends on the reciprocal of the density function of the error evaluated at the quantile of interest which, particularly for heteroscedastic non-iid cases, results in a complex and difficult estimation problem. It is well-known that the construction of confidence intervals based on the quantile regression estimator can be simplified by using a bootstrap.
To construct confidence intervals in quantile regression we propose an effective and easy to apply bootstrap method based on the idea of Silverman's (1986) kernel smoothing approach. This proposed bootstrapping method requires the estimation of the conditional variance function of the fitted quantile.
After fitting the τth quantile function, we obtain the residuals, which are squared and centered to zero. Estimating the conditional mean function of the centered squared residuals gives the conditional variance function of the errors about the estimated τth quantile. Using an estimate of the conditional variance function allows the standardisation of the residuals which are then used in Silverman's (1986) kernel smoothing bootstrapping procedure to make inferences about the parameters of the τth quantile function.
To estimate the conditional variance function we consider the adaptation of GLMs as well as non-parametric regression based estimation. These different approaches have been assessed under various data structures and compared to several existing methods. The simulation studies show good results in terms of coverage probability and the spread of the constructed parameters confidence intervals when compared with existing methods.
This methodology is also applicable to a wider class of regression models with heteroscedastic errors where the transformation to normality is difficult to achieve or maybe undesirable given a need to preserve the original data scale.

Bayesian Model Selection Criteria for Generalized Linear Models with Data Missing Not at Random

Zeynep Kalaylioglu

Middle East Technical University, Ankara, Turkey
kzeynep@metu.edu.tr

We provide an evaluation of the performances of deviance information criterion and weighted L measure for comparison among a set of candidate nonignorable missingness models. New DIC and WL extensions that take direct account of the missingness models are proposed. A Monte Carlo simulation experiment is designed to assess the performances of these model selection criteria in generalized linear models under different scenarios for missingness amounts.

Estimation of Daily Evaporation Using Different Modeling Methods

Gülay Tezel and Meral Büyükyıldız

Selcuk University, Konya, Turkey
gtezel@selcuk.edu.tr, meralbyildiz@selcuk.edu.tr

Evaporation is an important tool in the hydrologic cycle and used for hydrologic water balance, water resources planning and management. There are many meteorological factors affects the evaporation such as solar radiation, vapor pressure, temperature. In this study, daily evaporation data of Seydi?ehir in Turkey was used. This observed data which contains solar radiation, vapor pressure, temperature, wind, humidity and atmospheric pressure parameters is 10 years long with an observation period between 2000 and 2010. After realized homogeneity test, it was investigated which factor(s) are mostly effects using support vector machine (SVM), support vector machine regression, Self Organizing Map and Multiple Linear Regression analysis (MRA). Also different statistical techniques (root mean square error, mean square error, determination coefficient, etc.) are evaluated of these modeling methods to compare effective factors on evaporation. In addition, various equations used for estimation evaporation were applied and contrasted their results. Thus, the performance of all models for estimation of evaporation using meteorological variables has been illustrated in this study. The applications were realized utilizing different input parameters. As a result, the best achievement was generally obtained with SVM.

An Application of Optimization Algorithms for Generating Correlated Multivariate Random Samples

Anamai Na-udom[1] *and Jaratsri Rungrattanaubol*[2]

[1]Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand
[2]Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
anamain@nu.ac.th, jaratsrir@nu.ac.th

The methods of generating random samples from multivariate distribution are widely used in many fields. The correlated multivariate random samples are frequently required in the context of risk analysis. This paper presents an application of optimization algorithms for generating multivariate random samples where the marginal distribution and correlation matrix are specified. Two popular optimization algorithms namely Columnwise-pairwise (CP) and Simulated Annealing (SA) are employed in this study. The proposed method starts from generating each univariate random sample and the validity of the sample structure is checked through the possible boundary of the correlation values. Then these two popular optimization algorithms CP and SA are applied for rearranging the elements in each marginal distribution until the achieved correlation values are as close to the target values as possible. The results indicate that CP performs very well and is comparable to SA while the structure of CP is simpler and more transparency to use than SA.

Design of Experiments and Statistical Applications

Enhancement of Search Algorithms for Constructing Optimal Latin Hypercube Designs

Jaratsri Rungrattanaubol[1] *and Anamai Na-udom*[2]

[1]Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
[2]Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand
jaratsrir@nu.ac.th, anamain@nu.ac.th

Currently computer simulated experiments (CSE) have been extensively used in sciences and engineering applications. Selecting a proper design to run CSE is very critical for reliability of output responses. The output responses from computer simulated experiments are normally deterministic. Hence the space filling designs which focus on spreading the design points over a design space are required. Latin hypercube designs (LHD) are normally practiced in the context of computer simulated experiments. The optimal LHD for a given dimensional problem is obtained by using a search algorithm under a pre-specified optimality criterion. Usually the search process takes a long time to terminate, especially when the dimension of the problem is large. This paper presents the methods to enhance the performance of the search algorithms which are widely used in the context of computer simulated experiments. The comparative studies are also employed based on a range of dimensions of problem and the optimality criteria. The choice of the best search algorithms for particular dimension of the problem is presented and discussed.

An Analysis of Turkey Demographic and Health Survey 2008 Data with Co-Plot Method

Yasemin Kayhan Atilgan and Süleyman Günay

Hacettepe University, Ankara, Turkey
ykayhan@hacettepe.edu.tr, sgunay@hacettepe.edu.tr

Classical multivariate data analysis methods such as multi dimensional scaling, principal component analysis and cluster analysis, analyze observations and variables separately. However, examining the observations and variables on the same map provides great advantages to the researcher. Co-Plot method developed as an extension of multi dimensional scaling aims to address this problem. It consists of two graphs superimposed on each other. First graph represents n observations into a two dimensional space. The second graph consists of p arrows and each arrow represents a variable. By means of this merged graph, variables and observations are analyzed simultaneously. As a result, this method gives opportunity to make more detailed comments about the multivariate data set with a single map. This technique is especially suitable for few observations of many variables. In this study, Co-Plot method is briefly explained and then applied to some suitable variables chosen from "Turkey Demographic and Health Survey, 2008" data set. "This survey is a nationally representative sample survey designed to provide information on levels trends on fertility, infant and child mortality, family planning, and meternal and child health". Obtained results are interpreted with respect to coefficient of alienation, average correlation and map of Co-Plot. Superiority of Co-Plot method about visually interpreting the data is emphasized.

Estimation of Curvature and Displacement Ductility in Reinforced Concrete Buildings

M. Hakan Arslan

Selcuk University, Konya, Turkey
mharslan@selcuk.edu.tr

Ensuring sufficient ductility in building load bearing systems and elements of the load bearing system is quite important for their seismic performance. The Seismic codes stipulate that certain requirements must be met to maintain ductility values above a certain level. The purpose of this study is to determine how ductility values of both elements and load bearing systems vary as parameters related to the conditions specified in the codes change and as estimates of these values are used. With this aim in mind, the curvature ductility in columns and beams of a four-storey reinforced concrete (RC) building differs depending on parameters that include the axial load level, longitudinal reinforcement, transverse reinforcement, compression bar ratio and concrete strength. The value of the curvature ductility was found to vary according to the number of parameters and variance range, which was found to be 60 and 135 in the beam section and column section, respectively. Later, a pushover analysis was applied to 540 different statuses of the sample RC system for the same parameters, and the ratio variations and respective displacement (global) ductility of the frames were calculated. The relationship between obtained ductility values with the parameters, as well as the accuracy of the established model, were estimated using regression analyses (Multi-linear and nonlinear regression (MLR, NLR)) and 11 various artificial neural networks (ANN) methods. According to the estimation methods, it was found that the test parameters that significantly affect curvature ductility values are not sufficient to explain the displacement ductility values. On the other hand, it was seen that the estimation strength of ANNs proved to be greater than MLR in both curvature ductility and displacement ductility. Outcomes also indicated that the NLR model exhibits superior performance for estimating displacement ductility.

Individual Control Treatments in Designed Genetical and Agricultural Experiments

Stanislaw F. Mejza and Iwona Mejza

Poznan University of Life Sciences, Poznan, Poland
smejza@up.poznan.pl, imejza@up.poznan.pl

A common aim of genetical and agricultural experiments is to compare the test treatments with an individual control (standard) treatment. Two kinds of experiments are considered, namely; 1) - non-replicated genetical experiments performed at early stage breeding program and 2) - the factorial experiments with crossed and nested structures of factors. By unreplicated genetical experiment we mean one in which examined genotypes are replicated only once. The use of unreplicated design is only one possible way to carry out an evaluation (inference) of the lines. Additionally, to control the real or potential heterogeneity of experimental units, control (check) plots are arranged in the trial. Some plots (check plots) with a control variety are usually placed between the plots with the lines. There are two main problems that have to be considered in the experiment, i.e. density of check plots and arrangement of them, random or systematic. In the article a response surface methodology is proposed for the analysis of nonreplicated breeding experiments. First, estimates of the yield response surface based on check plots as supporting points are obtained. Then the treatment (genotype, hybrid) effect is estimated as the difference between the observation obtained for the treatment and the response surface forecast. The consequences of density and arrangements of controls (check plots) on statistical inference using both simulation and uniformity trials are investigated. Factorial experiments with nested and crossed factorial structure (split block designs, split plot designs) are considered in detail. In particular arrangements of individual controls in the incomplete split plot designs and incomplete split block designs are considered. Two aspects of these experiments, namely constructing methods leading to optimal designs and design efficiency, are examined. The Kronecker and the so-called semi-Kronecker product of designs are applied to generate new designs with desirable properties.

Modeling and Simulation III

Statistical Methods for Processing and Analysis of Digital Images: Determining Compressive Strength of Concrete

Gamze Cankaya, M. Hakan Arslan and Murat Ceylan

Selcuk University, Konya, Turkey
gamze@selcuk.edu.tr, mharslan@selcuk.edu.tr, mceylan@selcuk.edu.tr

Determination of compressive strength of concrete used in a reinforced concrete (RC) is important so as to estimate real behaviour of the buildings under loading such as earthquake and vertical loads. In the literature, there are many methods given to determine concrete compressive strength. In this study, it is aimed to explain new application area titled image processing techniques (IPT) on determining concrete compressive strength. From this motivation, image processing techniques (IPT) which are almost new in construction technology has been investigated firstly. After that, numerical and experimental part of this study, 23 different images of concrete cube samples was evaluated by means of IPT method. The input data and the corresponding output data (the targets of artificial neural network (ANN) have been defined. The digitized concrete sample photographs have been used as the input data and the compressive strength of samples have been taken as the output (target) data. Therefore, in this study, determination of compressive strength of concrete is realized with two phase: feature extraction from digital image and estimation using ANN. Processing and anaylsing steps of digital images some statistical methods has been used. For determining compressive strength of concrete, statistical methods and pattern recognition applied to high definition images. Statistical models are used for describing point relation between pixels, correlation and the shape and structure of objects. Obtained statistical features from images are used in ANN phase for pattern recognition. It was seen that the accuracy rate of the results of the ANN method and experimental results is very high. Therefore it can be concluded that the prediction power of ANN which has been used to determine the compressive strength of concrete is satisfactory level. It is important that ANN, which is frequently used in the field of IPT, can be easily applied in a discipline like civil engineering specifically in concrete technology.

Bayesian Estimation of Odds Ratios: An Application

Deniz Taşçı and Süleyman Günay

Department of Statistics, Hacettepe University, Ankara, Turkey
deniztasci@hacettepe.edu.tr, sgunay@hacettepe.edu.tr

The aim of this study is to find odds ratios estimates by using Bayesian approach for contingency tables. A contingency table can contain zero cell frequencies which are caused by sampling structure. These cells are said to be sampling zeros. In this study, likelihood function and prior distribution, which are utilized for estimation of odds ratios by employing Bayesian approach, are examined for the cells with or without sampling zeros. We consider Bayesian estimations of odds ratios for contingency tables which contain sampling zeros, and apply on a real dataset about scoliosis and kyphosis disorders.

Fast Robust Kernel Density Estimation

Kourosh Dadkhah

University of Kurdistan, Sanandaj, Iran
kdadkhah@gmail.com

The classical kernel density estimation technique is the commonly used method to estimate the density function. It is now evident that the accuracy of such density function estimation technique is easily affected by outliers. To remedy this problem, Kim and Scott (2008) proposed an Iteratively Re-weighted Least Squares (IRWLS) algorithm for Robust Kernel Density Estimation (RKDE). However, the weakness of IRWLS based estimator is that its computation time is very long. The shortcoming of such RKDE has inspired us to propose new non-iterative and unsupervised based approaches which are faster, more accurate and more flexible. The proposed estimators are based on our newly developed Robust Kernel Weight Function (RKWF). The basic idea of RKWF based method is to first define a function which measures the outlying distance of observation. The resultant distances are manipulated to obtain the robust weights. This idea that the normal (clean) data appear in high probability area of stochastic model, while the outliers appear in low probability area of stochastic model, has motivated us to develop RKWF. Based on this notion, the robust weights are incorporated in the kernel function to formulate the robust density function estimation. An extensive simulation study has been carried out to assess the performance of the RKWF-based estimator. The RKDE based on RKWF performs as good as the classical Kernel Density Estimator (KDE) in outlier free data sets. Nonetheless, their performances are faster, more accurate and more reliable than the IRWLS approach for contaminated data sets. The fast performance of this method makes it applicable not only for small data sets but also in some fields such as data mining which is faced with huge data sets.

Bayesian Test of Homogeneity of Transition Model for Analyzing Longitudinal Ordinal Data

Sajad Noorian[1] *and Mojtaba Ganjali*[2]

[1]Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran
[2]Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, G. C., Tehran, Iran
sajad.noorian@gmail.com, m-ganjali@sbu.ac.ir

For analyzing the longitudinal ordinal response data, many methods are available. Since in longitudinal studies, there are a sequence of correlated responses, we have to take into account the correlation between responses. One of the methods to consider this correlation is the Markov (transition) model. In this paper we present the Bayesian test of the homogeneity assumption of Markov (transition) model for analyzing the longitudinal ordinal response data. A cumulative logistic regression model and the Bayesian method, using MCMC, are implemented for testing the null hypothesis of homogeneity. We also define what the specific homogeneous covariate is. Our approach is applied to Fluvoxamine (a treatment for deregulation of serotonin in the brain) data.

Statistical Applications - Biostatistics

Plant Life Forms and Ecological Indices of Lake Provala

Katarina J. Čobanović, Ljiljana M. Nikolić and Slobodan C. Ničin

Agricultural Faculty, Novi Sad, Serbia
kacaherodot@gmail.com, ljnik@polj.uns.ac.rs, boba.nicin@yahoo.com

It is known that plants could be supposed like bio indicators showing the characteristics of climatic, ecological and other conditions of the region where they are spread out. Environmental factors in this study are expressed as ecological indices. There were analyzed five more important ecological indices: substrate moisture, nutrient content, substrate dispersion/aeration, substrate pH and humus content.Floristic study of the lake Provala and its surrounding regions were carried out in the period 1996-2004 where it was noted 65 vascular species grouped into 7 life forms (Nikoli?, 2005). Using some graphical procedures the study is oriented to the analysis of relation between life forms and ecological indices.The paper emphasizes the graphical presentation and analysis of the environmental indices, the frequency of plant species and life forms of vascular plants, based on the use of "Variability plots" (STATISTICA 8.0).This study showed that used type of graphical presentation of analyzed categorical variables provides a detailed and comprehensive preliminary analysis . Graphical presentation using Variability plots proved to be very suitable indicating which of the life forms are closed and which differ significantly, with respect to the frequency of plant species depending on the value of ecological indices.Therefore, this way of presenting can be a contribution to the exploratory data analysis, which precedes the application of methods of grouping and classification. The study is a part of a wider investigation of relationship between the ecological indices and life forms of vascular plants using correspondence analsysis.

Statistical Aspects of Evaluation of Environmental Policy in Slovenia

Žiga Kotnik and Maja Klun

Faculty of Administration, University of Ljubljana, Ljubljana, Slovenia
ziga.kotnik@fu.uni-lj.si, maja.klun@fu.uni-lj.si

Current global financial and economic crisis exposes important question of effectiveness and efficiency of public sector. Accompanying economic circumstances oblige us to behave very rational, take necessary austerity measures, create higher added value and utilize use of limited financial resources. The expediency of public spending has to represent the basis of healthy public finance and therefore, every budget has to be properly planned and oriented towards measurable results. Performance budgeting promises such solution as this approach investigates the linkage between spent public resources and planned public policy objectives. Realization of these objectives is measured through a set of indicators, attributed to each objective.

The environmental protection has become one of the main political priorities of the United Nations and the European Union. Environmental policy is one of the areas where measurement of performance and efficiency is particularly difficult specially owing to lack of information and absence of traceability of actual effects on the environment. For this reason, environmental policy requires its own approach that will properly evaluate environmental data and use them when planning the budget. The term environmental policy includes measures and objectives of environmental protection and strategically important long-term policies aimed at protecting and preserving nature and reducing harmful consequences in the environment.

In our research we will investigate whether the amount of financial resources invested influences the efficiency and effectiveness of achieving environmental objectives, taking into account public funds for environmental policy, i.e. funds collected from environmental taxes and funds spent on public investments in the environment. These will be estimated by a specifically tailored statistical model and tested in the Slovenian case.

Data Transformations and Normality: Example from Wheat Drought Stress Trial

Miroslav Z. Zorić [1], *Emilija B. Nikolić-Djorić*[2] *and Dragan S. Djorić*[3]

[1]Institute of Field and Vegetable Crops, Novi Sad, Serbia
[2]Faculty of Agriculture, Novi Sad, Serbia
[3]Faculty of Organizational Sciences, Beograd, Serbia
crop.biometrics@gmail.com, emily@polj.uns.ac.rs, djoricd@fon.bg.ac.rs

Data normality is an important issue when analyzing the datasets by linear based statistical models. Violation from this assumption could lead erroneous data interpretation. In this study the effect of different data transformation will be illustrated on the dataset from large wheat drought stress trial. Using this dataset, Dodig et al (2008, Australian Journal of Agricultural Research 59:536–545) reports on the problem of genotype by environment interaction problem for wheat yield under drought managed stress conditions. The experimental material in this trial consisted of 100 wheat released cultivars and landraces of worldwide origin, chosen on the basis of their differences in yield and performance of several other traits in irrigated and drought stress conditions.
The deviation from normality was measured by means of statistical tests that are base on the deviation of skewness and excess kurtosis from zero, density functions and properties of ranked series. In order to have more clear insight into data, several graphical techniques were applied (Q-Q plot, P-P plot and polyplot).
Although Box-Cox power transformation is usually recommended for elimination effects of non normality, for this type of data more appropriate was Johnson transformation.

Workshop

Creating Effective Visualizations

Hadley Wickham

Rice University, Houston, TX, United States of America
hadley@rice.edu

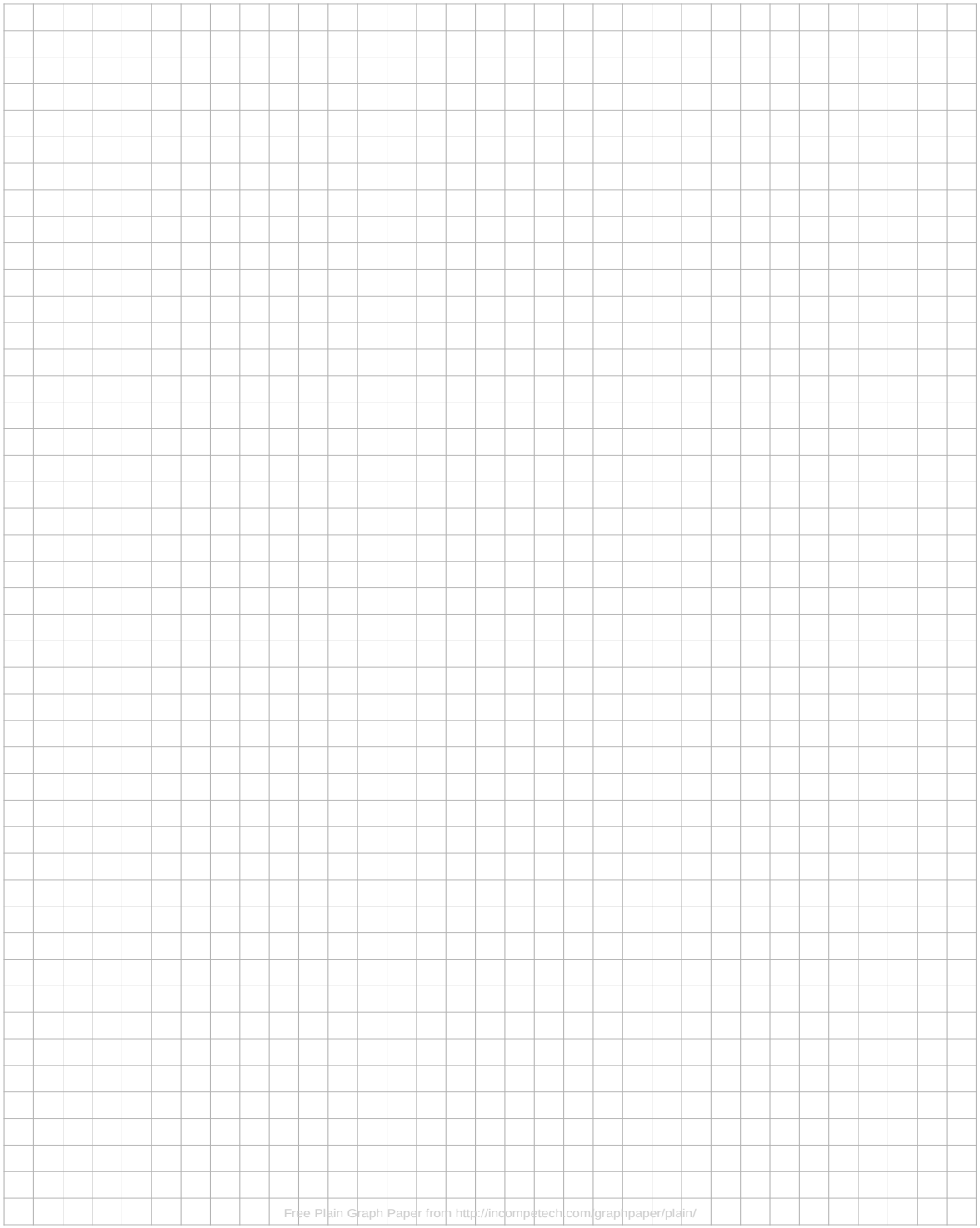
This half-day course will help you create better visualisations by teaching you about the findings from cognitive psychology that help us understand how the brain processes visual information. You'll learn important principles that underlying all visual displays and find out about some common mistakes that lead to confusion and inaccurate perception. We'll focus on four important principles:

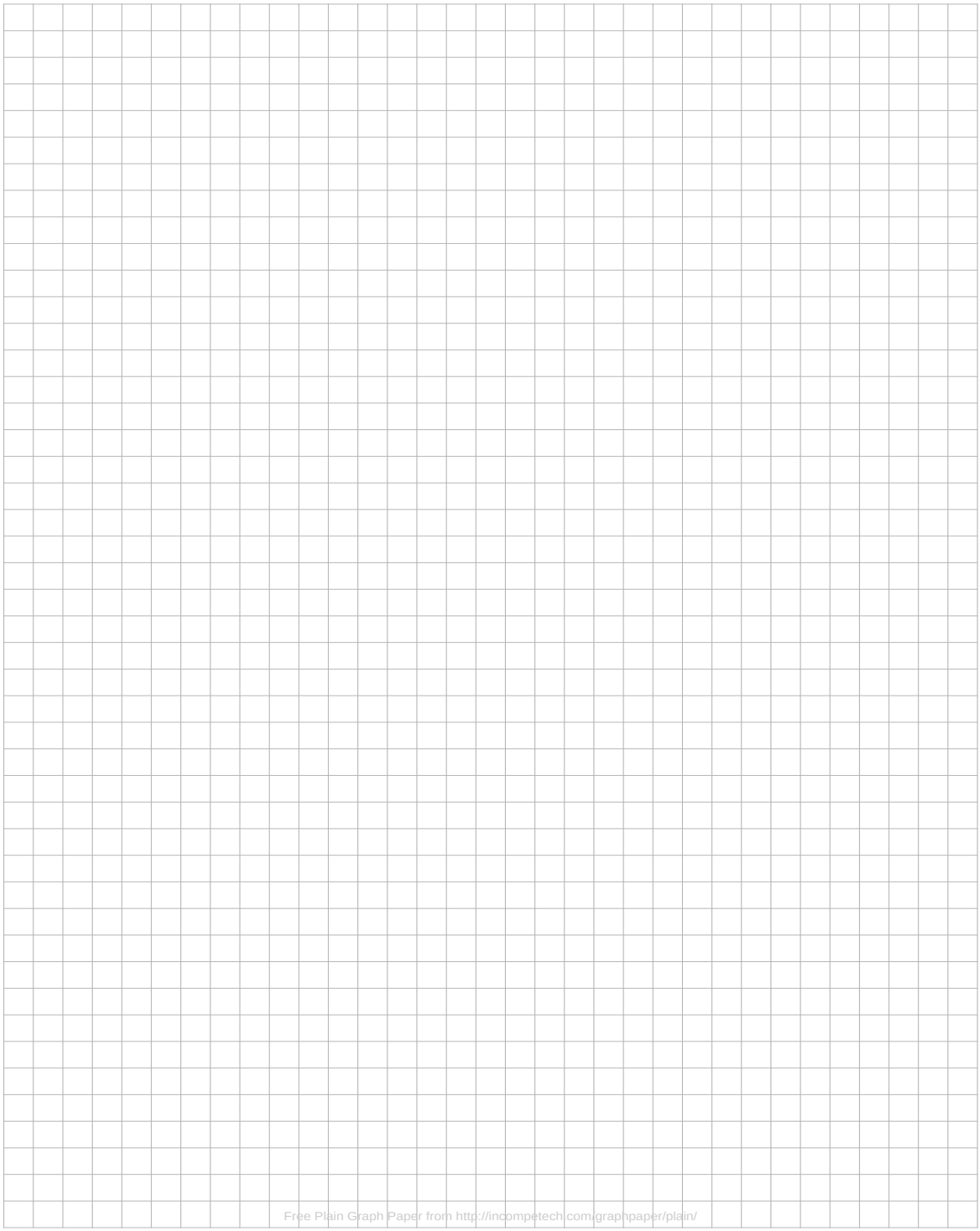
- Match perceptual and data topology
- Make important comparisons easy
- Visual connections should reflect real connections
- Beware of animation!

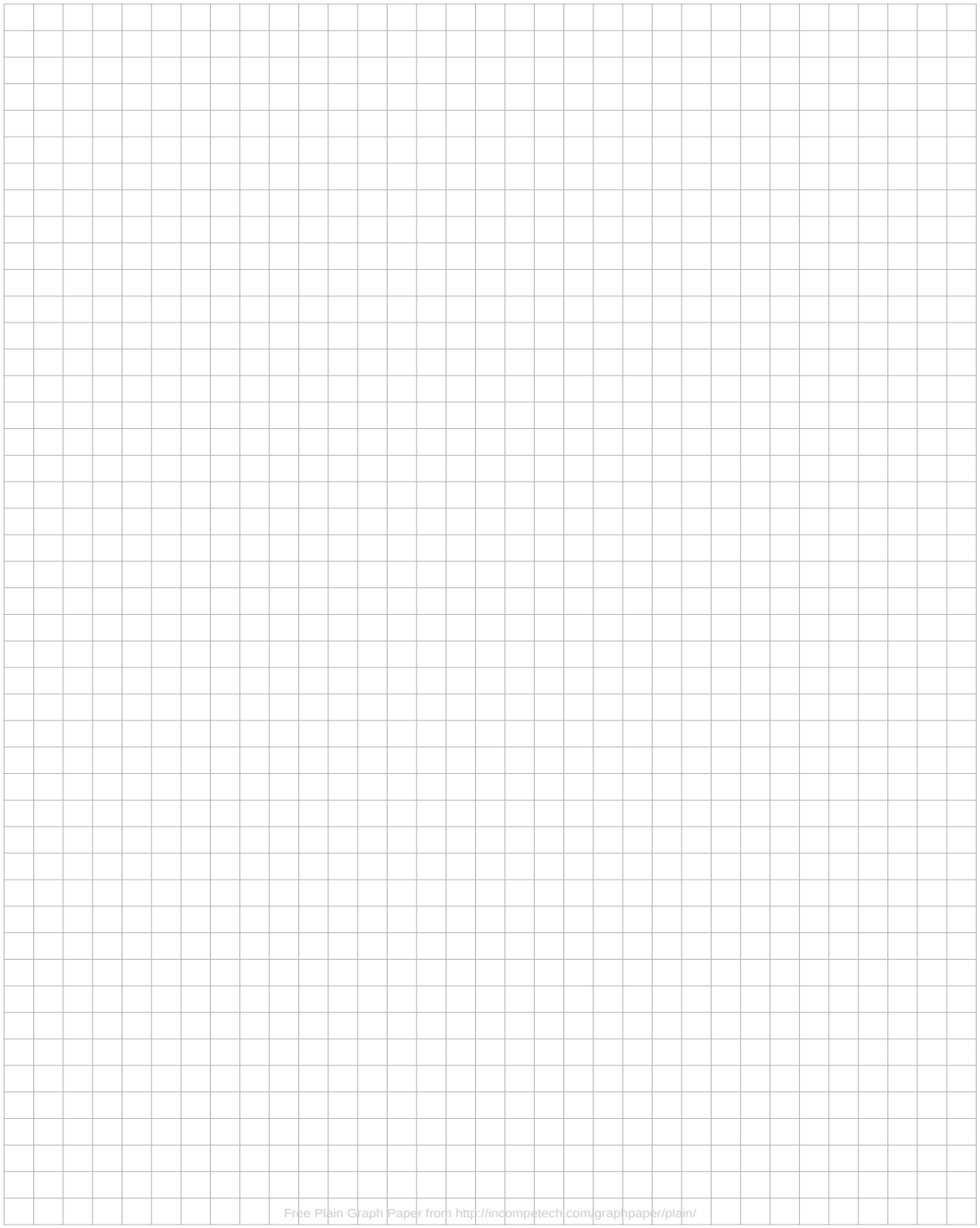
The class will be a mixture of lecture and small-group activities, where you'll apply your new skills to critique existing visualisations and suggest improvements. Bring along one or two visualisations that you've been struggling with.
While some of the examples will use the `ggplot2` R package, this course is graphics package agnostic. You'll be able to apply the skills you learn to any visualisation task, whether its in R or in another environment.

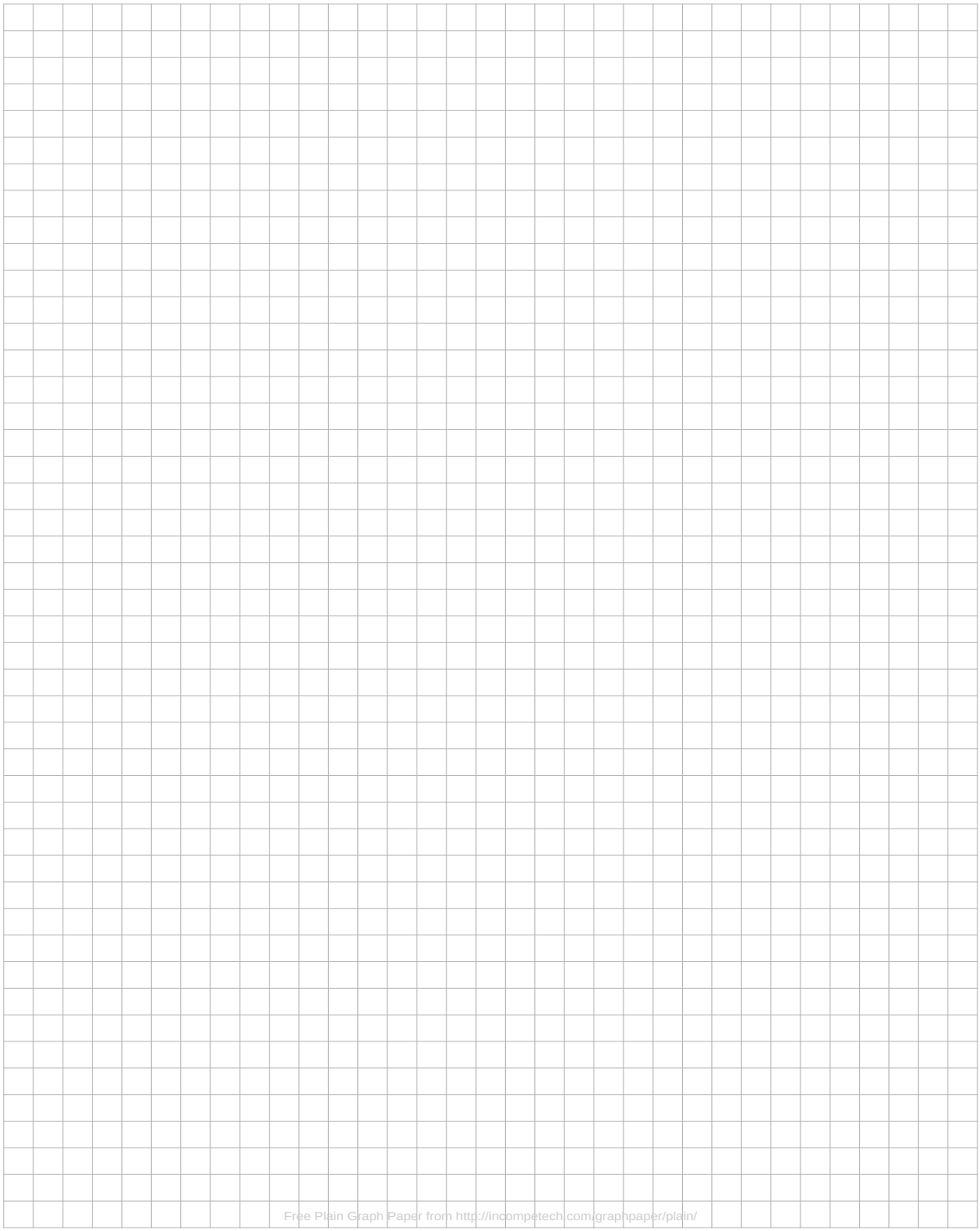
INDEX OF AUTHORS

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.alarix.si

www.result.si