MCP 2007
Vienna

5th International Conference on Multiple Comparison Procedures

08.-11. July 2007
Vienna, Austria

Editors: Martin Posch,
Jason Hsu, Franz König

www.mcp-conference.org

	Monday		Tuesday			Wednesday	
08:30	Charles Dunnett and Eckart Sonnemann Memorial Keynote Session		Theory and Foundations	Categorial Data/Non-parametrics	Bioinformatics /Genomics /Other	Bayesian Methods /False Discovery Rate	Adaptive Designs
10:00	Coffee Break						
10:30	Adaptive Designs	Bioinformatics /Genomics	Resampling Based Methods	Adaptive Designs	Screening and Selections	Bioinformatics	Others
12:00	Lunch Break						
01:30	Bioinformatics /Genomics	Adaptive Designs	Error Rates	Normal Theory, Linear/Non linear Models	Closed Testing and Partitio-ning Principle	Multiple Endpoints Problems	Error Rates
03:00	Poster Session 1			Poster Session 2		Coffee Break	
03:30	Multiple Endpoints	Bioinformatics /Genomics	Closed Testing and Partitio-ning Principle	Bioinformatics /Genomics	Response Adaptive and Optimal Designs	Clinical Trials	
05:00							

MCP 2007

Vienna

Conference Venue
University Campus - Altes AKH, Spitalgasse 2, A-1090 Vienna

Conference Website
www.mcp-conference.org

Imprint:

MCP 2007, Vienna
5th International Conference on Multiple Comparison Procedures
8-11 July, 2007
Medical University of Vienna
Core Unit for Medical Statistics and Informatics
Section of Medical Statistics
Spitalgasse 23
1090 Vienna, Austria

Editors: Martin Posch, Jason Hsu, Franz König

Production: Erwin Edtmayer, Sophie Frantal, Elisabeth Pernicka, Niklas Hack

Editorial assistance: Youlan Rao

Cover photo: The Vienna Giant Ferris Wheel built by the English engineer Walter B. Basset in the years 1896/97.
Source:
http://commons.wikimedia.org/wiki/Image:Prater_riesenrad.jpg

Print: Schwarzinger-Adressen, Heiligenstädter Str. 82, A-1190 Wien

ISBN: 978-3-200-00977-6

Table of Contents

Welcome to MCP 2007!

We welcome you to the 5th International Conference on Multiple Comparison Procedures (MCP) in Vienna, a city of coffee houses, concert halls, palaces, and Sachertorte!

In welcoming you, we must thank you for so enthusiastically responding to the conference call. The response has exceeded our original expectations by far, a very encouraging indication and one which speaks to the increasing importance of multiple comparisons procedures in applied statistics. Due to the many submissions we were forced to introduce partly a third parallel session, as well as two poster sessions. In total there are 100 talks and 26 poster presentations. The submissions very well reflect the currently most active fields of research: most numerous were submissions in the categories Bioinformatics/Genetical Statistics, Adaptive Designs, Multiple Endpoints, as well as Theory and Foundations.

The MCP meetings are organized by a loose international network of statisticians from universities and the pharmaceutical industry. The first meeting was held in 1996 in Tel Aviv and was followed by conferences in Berlin, Bethesda and Shanghai. This year's conference is located in the "Old General Hospital" in Vienna, Austria. The conference venue has a long history starting as a poor house in the 17th century that was transformed to a general hospital in the 18th century which in turn became a university campus in 1998. May the history and the charm of the city prove inspirational, as they did Beethoven, Mozart, and Klimt. We wish you a most memorable experience during MCP 2007!

Martin Posch, Jason Hsu, Franz König

Keynote

- Peter Bauer, Medical University of Vienna, Austria

Invited Speakers

- Alex Dmitrienko, Eli Lilly and Company, U.S.A.
- Sandrine Dudoit, University of California, Berkeley, U.S.A.
- Gerhard Hommel, Johannes Gutenberg University Mainz, Germany
- Mark van der Laan, University of California, Berkeley, U.S.A.
- Willi Maurer, Novartis Pharma AG, Basel, Switzerland
- Joachim Röhmel, Germany
- Joseph P. Romano, Stanford University, U.S.A.
- James F. Troendle, National Institutes of Health, U.S.A.
- Sue Jane Wang, U.S. Food and Drug Administration, U.S.A.
- Peter Westfall, Texas Tech University, U.S.A.
- Russ Wolfinger, SAS Institute, U.S.A.
- Daniel Yekutieli, Tel Aviv University, Israel

Organizing Committee

International Organizers

- Martin Posch (Co-Chair, Medical University of Vienna, Austria)
- Jason C. Hsu (Co-Chair, The Ohio State University, U.S.A.)
- Frank Bretz (Novartis Pharma AG, Switzerland)
- Guohua (James) Pan (Johnson & Johnson, U.S.A.)
- Ajit Tamhane (Northwestern University, U.S.A.)

Organizing Committee

- Peter Bauer (Medical University of Vienna, Austria)
- Yoav Benjamini (Tel Aviv University, Israel)
- Jie Chen (Merck, U.S.A.)
- Alex Dmitrienko (Eli Lilly and Company, U.S.A)
- Sandrine Dudoit (University of California, Berkeley, U.S.A.)
- Helmut Finner (German Diabetes Center, Germany)
- Anthony Hayter (University of Denver, U.S.A.)
- Chihiro Hirotsu (Meisei University, Japan)
- Ludwig Hothorn (University of Hanover, Germany)
- Armin Koch (Federal Institute for Drugs and Medical Devices, Germany)
- Sanat Sarkar (Temple University, U.S.A.)
- Peter Westfall (Texas Tech University, U.S.A.)

Local Organizers

- Franz König (Medical University of Vienna, Austria)
- Martin Posch (Medical University of Vienna, Austria)
- Andreas Futschik (University of Vienna, Austria)

Local Organizing Team (Medical University of Vienna, Austria)

- Zsuzsanna Egyedne Aranyi, Sophie Frantal, Alexandra Goll Niklas Hack, Elisabeth Pernicka, Vivian Ruschak, Karin Scholz, Sonja Zehetmayer
- Web Programming: Vivian Ruschak

Many thanks to Bernhard Lorenz, Erwin Rother and Martin Schweinberger from the IT and Finance Department of the Medical University of Vienna for their support.

Special Conference Issue

Biometrical Journal will publish a special issue with refereed articles from MCP 2007. We invite all participants to submit manuscripts to this special issue.

The intention of this special issue is to provide a topical collection of high-level scientific papers presented at the conference. All papers will be subject to a strict peer review process before publication. Any submitted manuscript should comply with the instructions for authors of Biometrical Journal.

Deadline for submission of manuscripts is **November 1, 2007.**

Manuscripts can be submitted electronically to
info@mcp-conference.org

General Information

Conference Venue:
University Campus - Altes AKH, Spitalgasse 2, Hof 2, A-1090 Vienna

Conference Office:
The conference office is located in the main hall of the conference venue. During the opening hours the office can be reached by telephone: ++43 / (0) 650 742 22 38
Opening Hours:

Sunday	7:45 – 18:30
Monday	7:45 – 15:30
Tuesday	8:15 – 15:30
Wednesday	8:15 – 15:30

Conference Website:
www.mcp-conference.org

Conference Travel Agency:
BTU, Operngasse 2, 1010 Wien, Tel: 01/516 51 54,
Fax: 01/516 51 51 54, email: s.zingaretti@btu.at

Lecture Halls:
C1, C2, Aula

Lunch:
In the restaurants on campus (see the map on the last page), daily specials are served. At the registration desk menues in English are available.

Additionally, there is a supermarket ("Billa") on campus where snacks are sold.

Internet access:
Internet access is available from Monday to Wednesday in the “EDV-Schulungsraum 2” on the first floor.

Social Events

Social Events:

- July, 08, Sunday
  5:30 pm Pre-conference Mixer
  6:30 pm Guided Walking Tour

- July, 09, Monday
  7:30 pm Reception at the City Hall

- July, 10, Tuesday
  7:30 pm Conference Dinner at a "Wiener Heurigen"

- July, 11, Wednesday
  7:00 pm Mozart Opera: "La Finte Semplice"

- July, 12, Thursday
  Guided Bus-Tour through Vienna
- July, 13, Friday
  Excursion to Wachau

Pre-conference Mixer:

July, 08, Sunday; 5:30 – 6:30 pm, Conference Venue

Guided Walking Tour:

July, 08, Sunday; 6:30 pm

Meeting point: Spitalgasse 2, 1090 Wien
Hof 2
Conference Venue

This tour offers you an initial overview of the history and the structure of the city. It takes you through the most beautiful and most elegant streets to the most famous sights in Vienna. It focuses on the Hofburg, residence of the Habsburgs for nearly 650 years and St. Stephen´s Cathedral, the landmark of Vienna. There will be no lack of tales about the Habsburgs - e.g. Sisi and Franz Joseph - the marriage and burial rites, or the stories concerning the Sachertorte, coffee houses and the oldest cake shop in Vienna.

Reception at the City Hall:

July, 09, Monday; 7:30 pm

Meeting point: City Hall, “Wappensaal”
Rathausplatz 1,1010 Wien
Entrance at Lichtenfelsgasse

Please bring the invitation you received with the registration with you. The invitation is necessary for admittance to City Hall. Each invitation is valid for one person only.

Public transport: U2, Tram 1, 2, D, J

Conference Dinner at Wiener Heuriger:

July, 10, Tuesday; 7:30 pm

Meeting point: Mayer am Pfarrplatz, Beethoven House
Pfarrplatz 2, 1190 Vienna

Please bring the conference dinner voucher with you.

The so called "Heuriger" is a speciality of Vienna and its sourrounding regions, where winemakers offer their wines directly in a pleasant atmosphere typically accompanied by traditional food and music. For more information see the enclosed folder.

Public Transport:
Tram 37 until final stop “Hohe Warte”; 5 min walk through a park (see markers).
Alternatively: Taxi (about 12€ from conference venue)

Scientific Presentations

Short Courses

Sunday, 8 July: 8:30 am – 12:30 , HC1

Introduction to Multiple Comparison Procedures
Jason Hsu

Sunday, 8 July: 8:30 am – 12:30 pm, HC2

Adaptive Designs for Clinical Trials
Werner Brannath, Frank Bretz

Sunday, 8 July: 1:30 -5:30 pm , HC1

Analysis of Multiple Endpoints
Alex Dmitrienko

Sunday, 8 July: 1:30 -5:30 pm , HC2

Analysis of Microarray Data
Katherine S. Pollard

Sessions

Monday 9 July, 8:30 – 10:00 am, HC1
Charles Dunnett and Eckart Sonnemann Memorial
Keynote Session

Multiple Inference in Medical Research – an Experience
Peter Bauer

Opening of the Conference - Martin Posch
Welcome from the Dean of the Medical University of Vienna - Wolfgang Schütz
Eulogy to Charles Dunnett - Ajit Tamhane
Eulogy to Eckart Sonnemann - Helmut Finner
Closing Remarks - Jason Hsu

Monday, 9 July, 10:30 am – 12:00 am

C1 C2

Adaptive Designs Chair: Willi Maurer	Bioinformatics/Genomics Chair: Alexander Ploner
Prospective Strategies and Challenges to Adaptively Designing Genomic Biomarker Targeted Trials in Trials Sue-Jane Wang Adaptive Model-Based Designs in Clinical Drug Development Vlad Dragalin Exploring Changes in Treatment Effects Across Design Stages in Adaptive Trials Tim Friede, Robin Henderson Estimation in Adaptive Group Sequential Design Cyrus Mehta, Werner Brannath, Martin Posch	A New Hypothesis to Test Minimal Fold Changes of Gene Expression Levels Jen-Pei Liu, Chen-Tuo Liao, Jia-Yan Dai Family-Wise Error on the Directed Acyclic Graph of Gene Ontology Jelle Goeman, Ulrich Mansmann Testing Procedures on Comparisons of Several Treatments with One Control in a Microarray Setting Dan Lin, Ziv. Shkedy, Tomasz Burzykowski, Hinrich W.H. Göhlmann, An De Bondt, Tim Perera On The Probability of Correct Selection for Large K Populations, with Application to Microarray Data Xinping Cui, Jason Wilson

Monday, 9 July, 1:30 – 3:00 pm

C1 C2

Bioinformatics/Genomics Chair: Jason Hsu	Adaptive Designs Chair: Frank Bretz
Multiple Testing Procedures with Applications to Genomics Sandrine Dudoit; van der Laan, Mark J. Involving Biological Information for Weighing Statistical Error Under Multiple Testing Anat Reiner-Benaim	**Modified Weighted Simes Tests in Group Sequential Designs** Willi Maurer Adaptive Designs with Correlated Test Statistics Heiko Götte, Andreas Faldum, Gerhard Hommel On the Use of Conventional Tests in Flexible, Multiple Test Designs Franz Koenig, Peter Bauer, Werner Brannath Flexible Group-Sequential Designs for Clinical Trials with Treatment Selection Nigel Stallard, Tim Friede

Monday, 9 July, 3:30 – 5:00 pm

C1 C2

Multiple Endpoints Chair: Chihiro Hirotsu	Bioinformatics/Genomics Chair: Jen-pei Liu
On Consequences of One-Sided Alternative Hypotheses for the Null Hypothesis Joachim Röhmel Multiple Comparisons for Ratios to the Grand Mean Ludwig A. Hothorn, G. Dilba Comparison of Methods for Estimating Relative Potencies in Multiple Bioassay Problems Gemechis Dilba Multiple hypothesis Testing to Establish Whether Treatment is "Better" than Control Aldo Solari, Salmaso Luigi, Pesarin Fortunato	Across and Down in Large SNP Studies: the MAX Test of Freidlin and Zheng vs SAS PROC CASECONTROL Dana Aeschliman, Marie-Pierre Dube Sample Size Calculation for Microarray Data Analysis using Normal Mixture Model Masaru Ushijima Estimating the Proportion of True Null Hypotheses with the Method of Moments Jose Maria Muino, P. Krajewski Knowledge-Based Approach to Handling Multiple Testing in Functional Genomics Studies Adam Zagdanski, Przemyslaw Biecek, Rafal Kustra

Tuesday, 10 July, 8:30 – 10:00 am

C1 C2 Aula

Theory and Foundations Chair: Ludwig Hthorn	Categorial Data/Nonparametrics Chair: Cyrus Mehta	Bioinformatics/Genomics/Other Chair: Nigel Stallard
Aesthetics and Power in Multiple Testing – a Contradiction? Gerhard Hommel A General Principle for Shortening Closed Test Procedures with Applications Werner Brannath, Frank Bretz FDR-Control: Assumptions, a Unifying Proof, least Favorables Configurations and FDR-Bounds Helmut Finner, Thorsten Dickhaus, Markus Roters Asymptotic Improvements of the Benjamini-Hochberg Method for FDR Control Based on an Asymptotically Optimal Rejection Curve Thorsten Dickhaus, Helmut Finner, Markus Roters	A Unified Approach to Proof of Concept and Dose Estimation for Categorical Responses Bernhard Klingenberg Multiple Testing Procedures with Incomplete Data for Rank-Based Tests of Ordered Alternatives Paul Cabilio, Jianan Peng Adjusting p-Values of a Stepwise Generalized Linear Model Chiara Brombin, Finos L., Salmaso L. A Test Procedure for Random Degeneration of Paired Rank Lists Michael G. Schimek, Peter Hall, Eva Budinska	Non-Negative Matrix Factorization and Sequential Testing Paul Fogel, S. Stanley Young, NISS Multiple Testing Procedures for Hierarchically Related Hypotheses Przemyslaw Biecek On the Conservatism of the Multivariate Tukey-Kramer Procedure Takahiro Nishiyama, Takashi Seo Distribution Theory with Two Correlated Chi-Square Variables Anwar H Joarder

Tuesday, 10 July, 10:30 am – 12:00 am

C1 C2 Aula

Resampling based methods Chair: Sandrine Dudoit	Adaptive Designs Chair: Tim Friede	Screening and Selection Chair: Juliet Shaffer
Resampling-Based Control of the False Discovery Rate Under Dependence Michael Wolf, Joseph Romano, Azeem Shaikh To Model or Not to Model Jason Hsu, Violeta Calian, Dongmei Li False Discovery Proportion Control under Dependence Yongchao Ge Resampling-Based Empirical Bayes Multiple Testing Procedure for Controlling the False Discovery Rate with Applications to Genomics Houston Gilbert, Sandrine Dudoit, Mark J. van der Laan	Confidence Sets Following a Modified Group Sequential Test Hans-Helge Müller, Nina Timmesfeld Unbiased Estimation after Modification of a Group Sequential Design Nina Timmesfeld, Schäfer, Helmut, Müller, Hans-Helge Homogeneity of Stages in Adaptive Designs Andreas Faldum Controversy? What Controversy? - An Attempt to Structure the Debate on Adaptive Designs Marc Vandemeulebroecke	On Estimates of R-Values in Selection Problems Andreas Futschik Screening for Partial Conjunction Hypotheses Ruth Heller, Benjamini, Yoav Exact Simultaneous Confidence Bands for Multiple Linear Regression over an Ellipsoidal Region Shan Lin, Wei Liu Stepwise Confidence Intervals for Monotone Dose-Response Studies Jianan Peng, Chu-In Charles Lee, Karolyn Davis

Tuesday, 10 July, 1:30 – 3:00 pm

C1 C2 Aula

Error Rates Chair: Ajit Tamhane	Normal theory, Linear/Non Linear Models Chair: Anthony Hayter	Closed Testing and Partitioning Principle Chair: Gunnar Stefansson
Control of Generalized Error Rates in Multiple Testing Joseph P. Romano Comparing Mutliple Tests for Separating Populations Juliet Shaffer A Leave-p-Out Based Estimation of the Proportion of Null Hypotheses in Multiple Testing Problems Alain Celisse Repeated Significance Tests Controlling the False Discovery Rate Martin Posch, Sonja Zehetmayer, Peter Bauer	Simultaneous Inference for Ratios David Hare, John Spurrier Neglect of Multiplicity in Hypothesis Testing of Correlation Matrices Burt Holland Minimum Area Confidence Set Optimality for Confidence Bands in Simple Linear Regression Wei Liu, A. J. Hayter Schéffe Type Multiple Comparison Procedure in Order Restricted Randomized Designs Omer Ozturk, Steve MacEachern	The multiple confidence procedure and its applications Tetsuhisa Miwa Gate-keeping testing without tears David Li, Mehrotra, Devan An Application of the Closed Testing Principle to Enhance One-Sided Confidence Regions for a Multivariate Location Parameter Michael Vock A Procedure to Multiple Comparisons of Diagnostic Systems Ana Cristina Braga, Lino A. Costa e Pedro N. Oliveira

Tuesday, 10 July, 3:30 – 5:00 pm

C1 C2 Aula

Closed Testing and Partitioning Principle Chair: Russel Wolfinger	Bioinformatics/Genomics Chair: Lawrence Gould	Response Adaptive and Optimal Designs Chair: Sue-Jane Wang
Multiple Testing of General Contrasts: Truncated Closure and the Extended Shaffer-Royen Method Peter H. Westfall, Tobias, Randall D. Powerful Short-Cuts for Gatekeeping Procedures Frank Bretz, Gerhard Hommel, Willi Maurer Simultaneous Confidence Regions Corresponding to Holm's Stepdown Multiple Testing Procedure Olivier Guilbaud Compatible Simultaneous Lower Confidence Bounds for the Holm Procedure and Other Closed Bonferroni Based Tests Klaus Strassburger, Frank Bretz	Detecting Differential Expression in Microarray Data: Outperforming the Optimal Discovery Procedure Alexander Ploner, Elena Perelman, Stefano Calza, Yudi Pawitan Flexible Two-Stage Testing in Genome-Wide Association Studies André Scherag, Helmut Schäfer, Hans-Helge Müller Sequential Genome-Wide Association Studies for Pharmacovigilance Patrick Kelly FDR Control for Discrete Test Statistics Anja Victor, Scheuer C, Cologne J, Hommel G	Multiple Treatment Comparison Based on a Non-Linear Binary Dynamic Model Brajendra Sutradhar, Vandna Jowaheer On Multiple Treatment Effects in Adaptive Clinical Trials for Longitudinal Count data Vandna Jowaheer, Brajendra Sutradhar Multi-Treatment Optimal Response-Adaptive Designs for Continuous Responses Atanu Biswas, Saumen Mandal On Identification of Inferior Treatments Using the Newman-Keuls Type Procedure Samuel Wu, Weizhen Wang, David Annis

Tuesday 10 July, 5:10 pm, HC1
1st Meeting of Austro-Swiss and German regions' working group on "Adaptive and Multiple Testing Procedures"

Wednesday, 11 July, 8:30 – 10:00 am

C1 C2

Bayesian Methods/False Discovery Rate Chair: James Pan	Adaptive Designs Chair: Werner Brannath
Bayesian Adjusted Inference for Selected Parameters Daniel Yekutieli A Bayesian Spatial Mixture Model for FMRI Analysis Brent Logan, Maya P. Geliazkova, Daniel B. Rowe, Prakash W. Laud A Bayesian Screening Method for Determining if Adverse Events Reported in a Clinical Trial are Likely to be Related to Treatment A Lawrence Gould Exact Calculations of Expected Power for the Benjamini-Hochberg Procedure Deborah Glueck, Anis Karimpour-Fard, Lawrence Hunter, Jan Mandel and Keith E. Muller	Estimating the Interesting Part of a Dose-Effect Curve: When is a Bayesian Adaptive Design Useful? Frank Miller Sample Size Re-Estimation and Hypotheses Tests for Trials with Multiple Treatment Arms Jixian Wang, Franz Koenig Adaptive Design in Dose Ranging Studies Based on Both Efficacy and Safety Responses Olga Marchenko, Prof. R. Keener, University of Michigan, Ann Arbor Adaptive Seamless Designs for Subpopulation Selection Based on Time to Event Endpoints Emmanuel Zuber, Werner Brannath, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, Amy Rac

Wednesday, 11 July, 10:30 am– 12:00 am

C1 C2

Bioinformatics Chair: Katherine S. Pollard	Others Chair: Sanat Sarkar
Ranks of True Positives in Large Scale Genetics Experiments Russell D. Wolfinger, Zaykin, Dmitri; Zhivotovsky, Lev; Czika, Wendy; Shao, Susan Multi-Stage Designs Controlling the False Discovery or the Family Wise Error Rate Sonja Zehetmayer, Peter Bauer, Martin Posch Two-Stage Designs for Proteomic and Gene Expression Studies Applying Methods Differing in Costs Alexandra Goll, Bauer Peter Some Insights Into FDR and k-FWER in Terms of Average Power and Overall Rejection Rate Meng Du	A Weighted Hochberg Procedure Ajit Tamhane, Lingyun Liu Multiple Testing in Change-Point Problem with Application to Safety Signal Detection Jie Chen Sequentially Rejective Test Procedures for Partially Ordered Sets of Hypotheses David Edwards, Jesper Madsen Simultaneous Confidence Intervals by Iteratively Adjusted Alpha for Relative Effects in the One-Way Layout Thomas Jaki, Martin J. Wolfsegger

Wednesday, 11 July, 1:30 – 3:00 pm

C1 C2

Multiple Endpoints Problems Chair: Jie Chen	Error Rates Chair: Klauss Strassburger
Stepwise Testing of Multiple Dose Groups Against a Control With Ordered Endpoints James Francis Troendle A New Method to Identify Significant Endpoints in a Closed Test Setting Carlos Vallarino, Joe Romano, Michael Wolf, Dick Bittman Proportion of True Null Hypotheses in Non High-Dimensional Multiple Testing Problems: Procedures and Comparison Mario Walther, Claudia Hemmelmann; Rüdiger Vollandt An Exact Test for Umbrella Ordered Alternatives of Location Parameters: the Exponential Distribution Case Parminder Singh	Procedures Controlling Generalized False Discovery Rate Sanat Sarkar, Wenge Guo Effects of Dependence in High-Dimensional Multiple Testing Problems Kyung In Kim, Mark A. van de Wiel A Semi-Parametric Approach for Mixture Models: Application to Local FDR Estimation Jean-Jacques Daudin, A. Bar-Hen, L. Pierre, S. Robin Two New Adaptive Multiple Testing Procedures Etienne Roquain, Gilles Blanchard

Wednesday, 11 July, 3:30 – 5:00 pm

C1

Clinical Trials Chair: Peter Westfall
Multi-Stage Gatekeeping Procedures with Clinical Trial Applications Alex Dmitrienko, Tamhane, Ajit A Unifying Approach to Non-Inferiority, Equivalence and Superiority Tests Chihiro Hirotsu Multiplicity-Corrected, Nonparametric Tolerance Regions for Cardiac ECG Features Gheorghe Luta, S. Stanley Young, Alex Dmitrienko Comparing Treatment Combinations with the Corresponding Monotherapies in Clinical Trials Ekkehard Glimm, Norbert Benda

Monday, 9 July, 3:00 – 3:30 pm

Poster session 1
Simultaneous Confidence Intervals for Overdispersed Count Data Daniel Gerhard, Frank Schaarschmidt, Ludwig A. Hothorn
Approximative Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions and Poly-3-Adjusted Tumour Rates Frank Schaarschmidt, Martin Sill, Ludwig A. Hothorn
Multiplicity Adjusted Location Quotients Gemechis Dilba, Frank Schaarschmidt, Bichaka Fayissa
Forecasting Monthly Temperature and Relative Humidity Using Time Series Analysis Inderjeet Kaushik, PR Maiti
Application of Multiple Comparison Procedures for Analysis of Naltrexone and Fluoxetine Effects for Treatment of Heroin Dependence Elena V. Verbitskaya, Evgeny M. Krupitsky, Edwin E. Zvartau, Marina V. Tsoi-Podosenin, MD, Valentina Y
Performance of Multiple Testing Procedures for Genomic Differences in Groups of Papillary Thyroid Carcinoma Analysed by Array CGH Herbert Braselmann, Eva Malisch, Kristian Unger, Gerry Thomas, Horst Zitzelsberger
Adjusting for Multiple Testing Mohamed Moussa, Nil
Optimal Allocation of Sample Size in Two-Stage Association Studies Shu-Hui Wen, CK Hsiao
Statistical Method for Finding Protein-Binding Sites from ChIP-Chip Tiling Arrays Taesung Park, Haseong Kim, Jae K. Lee
Maximum Contrast Tests and Model Selection under Order Restriction Xuefei Mi, L.A. Hothorn
Lets ROC on Microarrays Carina Silva-Fortes, Maria Antónia Amaral Turkman, Lisete Sousa
Biotechnology as Chance for Food Safety

Kakha Nadiradze

Nonparametric Tolerance Bounds for Gene Selection
S. Stanley Young, Gheorghe Luta

Tuesday, 10 July, 3:00 – 3:30 pm

Poster session 2
Estimation of Parameters in Unconditional Categorical Regression Azam Kamal, Grami A.
Methodological Issues in the Design and Sample Size Estimation of a Cluster Randomized Trial to Evaluate the Effectiveness of Clinical Sara Marchisio, Massimiliano Panella, Manzoli Lamberto, DiStanislao Francesco
Parametric Multiple Contrast Tests in the Presence of Heteroscedasticity Mario Hasler, Ludwig A. Hothorn
Adjustment Method to Address Type I Error and Power Issues with Outcome Multiplicity and Correlation Richard Blakesley, Sati Mazumdar, Patricia Houck
A Simulation Study on the Gain in Power of Multiple Test Procedures by using Information on the Number of True Hypotheses Claudia Hemmelmann, Andreas Ziegler, Rüdiger Vollandt
Quantile Curve Estimation and Visualization for Non-Stationary Time Series Dana Draghicescu, Serge Guillas, Wei Biao Wu
Scale and Suitable Analysis Fumihiko Hashimoto
On Orthogonal Series Estimation Methods Mei Ling Hunag, Percy Brill
Inequalities for Multivariate Normal Probabilities of Nonsymmetric Rectangles

Vered Madar

Study on Statistical Analysis for Adverse Drug Reaction in Korea
Hyeon Jeong Kim, Eunhee Kim, Mun Sin Kim, Junghoon Jang, Bong Hyun Nam

Bayesian Classification and Label Estimation via EM Algorithm: A Comparative Study
Marilia Antunes, Lisete Sousa

Testing Equality of Two Mean Vectors with Uniform Covariance Structure when Missing Observations Occur
Kazuyuki Koizumi, Toshiya Iwashita,Takashi Seo

Controlling the Number of False Positives using the Benjamini-Hochberg Procedure
Paul Somerville

Talks

Abstracts are sorted by session.

The code at the bottom of each abstract denotes the session and is of the form: Day (M,T,W), Session (am1, am2, pm1, pm2), Lecture Hall (C1, C2, Aula), Talk (T1-T4).

Keynote

Multiple Inference in Medical Research – an Experience

Peter Bauer

Medical University of Vienna, Austria

The history of applications of multiple comparisons procedures in medical research over the last four decades is given from a personal perspective. Some milestones are sketched. The reaction to new statistical methodology in the medical literature is investigated by comparing how multiplicity issues have been handled in medical journal articles twenty years ago and today. Some of the multiplicity issues in the rapidly developing area of gene expression and gene association studies are discussed. They have provoked new concepts and have helped to establish multiplicity as an important point to consider in the scientific community. Some arguments a consulting medical statistician may expect from his clients will be sketched followed by some pragmatic concluding comments.

Mam1

Prospective Strategies and Challenges to Adaptively Designing Genomic Biomarker Targeted Trials in Trials

Sue-Jane Wang (Invited Speaker)

Center for Drug Evaluation and Research, U.S.A.

In recent decades, translational research has increasingly aimed to uncover the genomic biomarker signature that has the potential to differentiate the therapeutic effect among patients who are predicted to be good versus poor signatures. The advent of genomic biomarker gradually brings the awareness that phenotypically homogeneous patients may be heterogeneous at the genomic level. In this talk, I will present adaptive designs that prospectively account for genomic heterogeneous patient subpopulations and diagnostics test performance characteristics. The statistics concept of biomarker qualification and the efficiency of the pharmacogenomics clinical trials in view of personalized medicine will be exemplified. Aside from alpha allocation strategies and adaptive multiple hypotheses testing, challenges to adaptively designing pharmacogenomics targeted trials within conventional clinical trials will be elucidated via typical examples.

Mam2C1T1

Adaptive Model-Based Designs in Clinical Drug Development

Vlad Dragalin

Wyeth Research Wyeth, U.S.A.

The objective of a clinical trial may be either to target the maximum tolerated dose or minimum effective dose, or to find the therapeutic range, or to determine the optimal safe dose to be recommended for confirmation, or to confirm efficacy over control in a Phase III clinical trial. This clinical goal is usually determined by the clinicians from the pharmaceutical industry, practicing physicians, key opinion leaders in the field, and the regulatory agency. Once agreement has been reached on the objective, it is the statistician's responsibility to provide the appropriate design and statistical inferential structure required to achieve that goal. There is a plenty of available designs on statistician's shelf. The greatest challenge is their implementation. We exemplify this in three case studies.

Mam2C1T2

Exploring Changes in Treatment Effects Across Design Stages in Adaptive Trials

Tim Friede, Robin Henderson

Warwick Medical School, University of Warwick, United Kingdom

The recently published draft of a CHMP reflection paper on flexible designs highlights a controversial issue regarding the interpretation of adaptive trials when the treatment effect estimates differ across design stages (CHMP, 2006). In Section 4.2.1 it states "… the applicant must pre-plan methods to ensure that results from different stages of the trial can be justifiably combined. In this respect, studies with adaptive designs need at least the same careful investigation of heterogeneity and justification to combine the results of different stages as is usually required for the combination of individual trials in a metaanalysis." This suggests that a test for heterogeneity should be preplanned and in the event of a significant result the policy should be to discard observations subsequent to the interim analysis that induced changes in the treatment. In this presentation we investigate the error rates of this procedure. Furthermore, we present an alternative testing strategy which is based on change point methods to detect calendar time effects (Friede and Henderson, 2003; Friede et al., 2006). In a simulation study we demonstrate that our procedure performs favourably compared to the procedure suggested by the guideline. Tools that help to explore changes in treatment effects will be discussed.

Committee for Medicinal Products for Human Use (2006) Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. London, 23 March 2006, Doc. Ref. CHMP/EWP/2459/02.

Friede T, Henderson R (2003) Intervention effects in observational studies with an application in total hip replacements. Statistics in Medicine 22: 3725-3737.

Friede T, Henderson R, Kao CF (2006) A note on testing for intervention effects on binary responses. Methods of Information in Medicine 45: 435-440.

Mam2C1T3

Estimation in Adaptive Group Sequential Design

Cyrus Mehta, Werner Brannath, Martin Posch

Cytel Inc., U.S.A.

This paper proposes two methods for computing confidence intervals with exact or conservative coverage following a group sequential test in which an adaptive design change is made one or more times over the course of the trial. The key idea, due to Muller and Schafer (2001), is that by preserving the null conditional rejection probability of the remainder of the trial at the time of each adaptive change, the overall type 1 error, taken unconditionally over all possible design modifications, is also preserved. This idea is further extended by considering the dual tests of repeated confidence intervals (Jennison and Turnbull, 1989) and of stage-wise adjusted confidence intervals (Tsiatis, Rosner and Mehta, 1984). The method extends to the computation of median unbiased point estimates.

Mam2C1T4

A New Hypothesis to Test Minimal Fold Changes of Gene Expression Levels

Jen-pei Liu, Chen-Tuo Liao, Jia-Yan Dai

Division of Biometry, Department of Agronomy, National Taiwan University, Taiwan

Current approaches to identifying differentially expressed genes are based either on the fold changes or on the traditional hypotheses of equality. However, the fold changes do not take into consideration the variation in estimation of the average expression. In addition, the use of fold changes is not in the frame of hypothesis testing and hence the probability associated with errors for decision-making in for identification of differentially expressed genes can not be quantified and evaluated. On the other hand, the traditional hypothesis of equality fails to take into consideration the magnitudes of the biologically meaningful fold changes that truly differentiate the expression levels of genes between groups. Because of the large number of genes tested and small number of samples available for microarray experiments, the false positive rate for differentially expressed genes is quite high and requires further adjustments such as Bonferroni method, false discovery rate, or use of an arbitrary cutoff for the p-values. All these adjustments do not have any biological justification. Hence, we propose to formulate the hypothesis of identifying the differentially expressed genes as the interval hypothesis by consideration of both the minimal biologically meaningful fold changes and statistical significance simultaneously. Based on the interval hypothesis, a two one-sided tests procedure is proposed with a method for sample size determination. A simulation study is conducted to empirically compare the type I error rate and power of the traditional hypothesis among the two-sample t-test, the two-sample t-test with Bonferroni adjustment, the fold-change rule, the method of combination of the two-sample t-test and fold-change rule, and the proposed two one-sided tests procedure under various combinations of fold changes, variability and sample sizes. Simulation results show that the proposed two one-sided tests procedure based on the interval hypothesis not only can control the type I error rate at the nominal level but also provides sufficient power to detect differentially expressed gene. Numeric data from public domains illustrate the proposed methods. *Mam2C2T1*

Family-Wise Error on the Directed Acyclic Graph of Gene Ontology

Jelle Goeman, Ulrich Mansmann

Medical Center, Leiden University, The Netherlands

Methods that test for differential expression of gene groups such as provided by the Gene Ontology database are becoming increasingly popular in the analysis of gene expression data. However, so far methods could not make use of the graph structure of Gene Ontology when adjusting for multiple testing.
We propose a multiple testing method, called the focus level procedure, that preserves the graph structure of Gene Ontology (GO) when testing for association of the expression profiles of GO terms with a response variable. The procedure is constructed as a combination of a Closed Testing procedure with Holm's method. It allows a user to choose a “focus level" in the GO graph, which reflects the level of specificity of terms in which the user is most interested. This choice also determines the level in the GO graph at which the procedure has most power. The procedure strongly keeps the family-wise error rate without any additional assumptions on the joint distribution of the test statistics used. We also present an algorithm to calculate multiplicity-adjusted p-values. Because the focus level procedure preserves the structure of the GO graph, it does not generally preserve the ordering of the raw p-values in the adjusted p-values.

Mam2C2T2

Testing Procedures on Comparisons of Several Treatments with one Control in a Microarray Setting

Dan Lin, Ziv. Shkedy, Tomasz Burzykowski, Hinrich W.H. Göhlmann, An De Bondt, Tim Perera

Center for Statistics, Hasselt University, Belgium

We discuss a particular situation in a microarray experiment; when two dimensional multiple testing occurs because of comparing several treatments with a control at one hand and testing tens of thousands of genes simultaneously at the other hand. Dunnett's single step procedure (Dunnett 1995) for testing effective treatments can be used to address one dimensional question of primary interest. Dunnett's procedure was implemented within resampling-based algorithms such as Significance Analysis of Microarray (SAM, Tusher et al. 2001) and Benjamini and Hochberg False Discovery Rate (FDR, Benjamini and Hochberg 1995). To combine the two-dimensional testing problem into one testing procedure, we proposed an approach to test for *m*K* (number of genes*number of comparisons between several treatments with the control) tests simultaneously. We compared the performance of SAM and the classical BH-FDR. The method was applied to a microarray experiment with 4 treatment groups (3 microarrays in each group) and 16998 genes. Additionally a simulation study was conducted to investigate the power of the methods proposed and to investigate how to choose the fudge factor in SAM to leverage the genes with small variances.
Keywords: Dunnett's single step procedure; microarray; multiple testing; Benjamini and Hochberg false discovery rate (BH-FDR); SAM.

Mam2C2T3

On the Probability of Correct Selection for Large k Populations, with Application to Microarray Data

Xinping Cui, Jason Wilson

University of California, Riverside Department of Statistics, U.S.A.

One frontier of modern statistical research is the "multiple comparison problem" (MCP) arising from data sets with large *k* (>1000) populations, e.g. microarrays and neuroimaging data. In this talk we demonstrate an alternative to hypothesis testing. It is an extension of the Probability of Correct Selection (PCS) concept. The idea is to select the top *t* out of *k* populations and estimate the probability that the selection is correct, according to specified selection criteria. We propose "d-best" and "G-best" selection criteria that are suitable for large *k* problems and illustrate the application of the proposed method on two microarray data sets. Results show that our method is a powerful method for the purpose of selecting the "top *t* best" out of *k* populations.

Mam2C2T4

Multiple Testing Procedures with Applications to Genomics

Sandrine Dudoit, van der Laan, Mark J. (Invited Speakers)

University of California, U.S.A.

In this two-part presentation, we will provide an overview of a general methodology for multiple hypothesis testing and applications to a range of large-scale testing problems in biomedical and genomic research. Specifically, we will describe resampling-based single-step and stepwise multiple testing procedures for controlling a broad class of Type I error rates, defined as tail probabilities and expected values for arbitrary functions of the numbers of Type I errors and rejected hypotheses (e.g., generalized family-wise error rate, tail probability for the proportion of false positives among the rejected hypotheses, false discovery rate). Unlike existing approaches, the procedures are based on a test statistics joint null distribution and provide Type I error control in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses, and test statistics. The multiple testing results are reported in terms of rejection regions, parameter confidence regions, and adjusted *p*-values. A key ingredient of our proposed procedures is the null distribution used in place of the unknown joint distribution of the test statistics.

We provide a general characterization for a proper test statistics null distribution, which leads to the explicit construction of two main types of test statistics null distributions. The first null distribution is the asymptotic distribution of a vector of null shift and scale-transformed test statistics, based on user-supplied upper bounds for the means and variances of the test statistics for the true null hypotheses. The most recent proposal defines the null distribution as the asymptotic distribution of a vector of null quantile-transformed test statistics, based on user-supplied marginal test statistics null distributions. We will discuss joint resampling-based empirical Bayes procedures for controlling generalized tail probability and expected value error rates.

The approach involves specifying:

(i) a null distribution for vectors of null test statistics and

(ii) a distribution for random guessed sets of true null hypotheses.

By randomly sampling null test statistics and guessed sets of true null hypotheses, one obtains a distribution for an arbitrary guessed

function of the numbers of false positives and rejected hypotheses, for any given vector of cut-offs for the test statistics. Cut-offs can then be chosen to control tail probabilities and expected values for this distribution at a user-supplied level. Due to their generality and flexibility, our proposed multiple testing procedures are well-suited to address high-dimensional testing problems arising in different areas of application of statistics. We will conclude with an overview of applications in biomedical and genomic research, including:
the identification of differentially expressed and co-expressed genes in high-throughput gene expression experiments, such as microarray experiments;
tests of association between gene expression measures and biological annotation metadata (e.g., Gene Ontology);
sequence analysis;
the genetic mapping of complex traits using single nucleotide polymorphisms.
Our forthcoming book provides a detailed account of the theoretical foundations of our multiple testing methodology and discusses its software implementation in the R package(www.bioconductor.org)and applications in biomedical and genomic research (Dudoit and van der Laan, 2007).

Mpm1C1T1+2

Involving Biological Information for Weighing Statistical Error under Multiple Testing

Anat Reiner-Benaim

Stanford University, U.S.A.

Given a multiple testing problem, each hypothesis may be associated with some prior information, which is related to the structure of the data and its scientific basis. This information may be unique to each hypothesis, and therefore, when estimating the overall statistical error, treating the hypotheses as having the same null distributions may lead to biased results. Using the prior information for weighing the null hypothesis can improve the error estimate and may offer less conservative controlling procedure.

The emphasis of the talk will be on use of biological data as prior information. For instance, the machinery of genetic regulation is subjected to probabilistic factors. Regulation happens when a transcription factor binds to a site on the gene. Since the match level between the two is not perfect and can vary within a wide range, it can be incorporated into the error estimation as hypotheses weights.

The effect of the weights on the error estimate will be presented, given the method of computing the weights, the pattern of the weight structure and the type of error controlled. Two approaches to control the False Discovery Rate (FDR) with weights are compared – empirical Bayes per-hypothesis FDR estimation, and weighing the p-values to control the overall FDR.

Mpm1C1T3

Modified Weighted Simes Tests in Group Sequential Designs

Willi Maurer (Invited Speaker)

Novartis Pharma AG, Switzerland

Multiple hypothesis testing problems, where the joint distribution between test statistics is not fully known and different weights or priorities are given to hypotheses, will be discussed. If there is no fully hierarchical ordering among the hypotheses, a general approach consists of gatekeeping procedures that usually are based essentially on the Bonferroni inequality with unequal type I error probability allocation, applied to closed testing procedures. A possible improvement with respect to increased power can be achieved by using the Simes inequality. Such scenarios arise, e.g., in trials aiming at investigating the effectiveness of cardiovascular and diabetes treatments in preventing or delaying cardiovascular events. The primary endpoints considered in these cases are often compound variables summarizing events in time, like fatal/nonfatal MI and stroke, revascularization, hospitalization for unstable angina, etc. So called 'hard' and 'soft' endpoints are built from subsets of these events where 'hard' endpoints can be comprised of a subset of the events constituting a 'soft' endpoint. Such endpoints are usually highly correlated among themselves but may be of unknown correlation with further endpoints like time to progression of diabetes. An additional complication in such trials is that they are usually of long duration and interim decisions have to be taken in a group sequential or adaptive design setting. We will discuss in this context issues that arise with respect to the validity of the Simes inequality in the case of unequal allocation of the probability of a type I error and show that results for bi-variate test statistics in the two-sided case can be extended to the one sided case if the rejection region is slightly altered. The problem of applying Hochberg-type testing strategies in a group sequential setting is discussed and various options regarding the allocation of spending functions in the arising repeated closed test situation are compared. Much of the work and newer results presented have been done together with Werner Brannath, Frank Bretz and Sanat Sarkar.

Mpm1C2T1

Adaptive Designs with Correlated Test Statistics

Heiko Götte, Andreas Faldum, Gerhard Hommel

Institute of Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg - University Mainz, Germany

In clinical trials the collected observations are often correlated, for example: clustered data or repeated measurements. When applying adaptive designs test statistics of different stages are often also correlated in these situations so that classical adaptive designs for uncorrelated test statistics (for example Bauer/ Köhne, 1994) do not seem to be appropriate. Hommel et al. (2005) proposed the Modified Simes test for two stage adaptive designs with correlated test statistics to handle this issue. For bivariate normally distributed test statistics the significance level can be preserved. Analogously to Shih/ Quan (1999) we give the probability of type one error for the Bauer-Köhne-design in the situation of bivariate normally distributed test statistics in an explicit formula. We show that the significance level is inflated for positively correlated test statistics. The decision boundary for the second stage can be modified in a way that the type one error is controlled. The concept is expandable to other adaptive designs. The Modified Simes test is a special case. In order to use these designs the correlation between the test statistics has to be determined. For a repeated measurement setting we show how correlation can be estimated within the framework of linear mixed models. The power of Modified Simes test is compared with the power of the Bauer-Köhne-design for this situation.

This talk contains parts of the thesis of Heiko Götte.

Bauer, P., Köhne, K. (1994). Evaluation of Experiments with Adaptive Interim Analyses. Biometrics, 50:1029-1041.

Hommel G., Lindig V., Faldum A.(2005). Two-stage adaptive designs with correlated test statistics. Journal of Biopharmaceutical Statistics, 15:613-623.

Shih W.J., Quan H. (1999). Planning and analysis of repeated measures at key time-points in clinical trials sponsored by pharmaceutical companies. Statistics in Medicine, 18:961-973

Mpm1C2T2

On the Use of Conventional Tests in Flexible, Multiple Test Designs

Franz Koenig, Peter Bauer, Werner Brannath

Medical University of Vienna, Austria

Flexible designs based on the closure principle offer a large amount of flexibility in clinical trials with control of the type I error rate. This allows the combination of trials from different clinical phases of a drug development process. Flexible designs have been criticized because they may lead to different weights for the patients from the different stages when reassessing sample sizes. Analyzing the data in a conventional way avoids such unequal weighting but may inflate the multiple type I error rate. In cases where the conditional type I error rates of the new design (and conventional analysis) is below the conditional type I error rates of the initial design the conventional analysis may be done without inflating the type I error rate. This method will be used to explore switching between conventional designs for typical examples.

Mpm1C2T3

Flexible Group-Sequential Designs for Clinical Trials with Treatment Selection

Nigel Stallard, Tim Friede

Warwick Medical School, University of Warwick, United Kingdom

Most statistical methodology for phase III clinical trials focuses on the comparison of a single experimental treatment with a control treatment. Recently, however, there has been increasing interest in methods for trials that combine the definitive analysis associated with phase III clinical trials with the treatment selection element of a phase II clinical trial.

A group-sequential design for clinical trials that involve treatment selection was proposed by Stallard and Todd (Statistics in Medicine, 22, 689-703, 2003). In this design, the best of a number of experimental treatments is selected on the basis of data observed at the first of a series of interim analyses. This experimental treatment then continues together with the control treatment to be assessed in one or more further analyses. The method was extended by Kelly, Stallard and Todd (Journal of Biopharmaceutical Statistics, 15, 641-658, 2005) to allow more than one experimental treatment to continue beyond the first interim analysis. This design controls the type I error rate under the global null hypothesis, but may not control error rates under individual null hypotheses if the treatments selected are not the best performing.

In some cases, for example when additional safety data are available, the restriction that the best performing treatments continue may be unreasonable. This talk will describe an extension of the approach of Stallard and Todd that controls the type I error rates under individual null hypotheses whilst allowing the experimental treatments that continue at each stage to be chosen in any way.

Mpm1C2T4

On Consequences of One-Sided Alternative Hypotheses for the Null Hypothesis

Joachim Röhmel (Invited Speaker)

Germany

ONeill (1997) defines the primary endpoint as 'a clinical endpoint that provides evidence sufficient to fully characterize clinically the effect of a treatment in a manner that would support a regulatory claim for the treatment'. In a clinical trial with an experimental treatment and a reference the situation may occur that two (or even more) primary endpoints may be necessary to describe the experimental treatment's benefit. Sometimes effects on important secondary endpoints will influence the judgement on the experimental treatment's value. For these situations multiple testing procedures have been developed to control the rate of false claims on superiority or non-inferiority. Often multiple testing procedures focus on the aim that for at least one primary endpoint the a priori set target is met, and little attention is then given to those endpoints which failed to demonstrate the proposed effect. When taking the above definition of a primary endpoint seriously, however, assurance is needed that none of the "non-significant" endpoints is inferior. Therefore, the focus of interest in a situation with multiple primary endpoints should be the more specific minimal target to demonstrate superiority in one of them given that non-inferiority is observed in the remaining. Several proposals exist in the literature for dealing with this or similar problems, but prove insufficient or inadequate at a closer look (e.g. Bloch et al. (2001, 2006) or Tamhane and Logan (2002, 2004)). In the talk I will focus on the case of two primary endpoints. Many aspects, however, can be transferred easily to the general case. I propose a hierarchical three step procedure, where non-inferiority in both variables must be proven in the first step, superiority has to be shown by a bivariate test (e.g. Holm (1979), O'Brien (1984), Hochberg (1988), a bootstrap (Wang (1998)), or Läuter (1996)) in the second step, and then superiority in at least one variable has to be verified in the third step by a corresponding univariate test. From the above mentioned bivariate superiority tests Läuter's SS test and the Holm procedure are preferably for the reason that these have been proven to control the type I error strictly, irrespective of the correlation structure among the primary variables and the sample size applied. A

simulation study reveals that the performance regarding power of the bivariate test depends to a considerable degree on the correlation and on the magnitude of the expected effects of the two primary endpoints. The part of the talk is based on a joined work with Christoph Gerlinger (Schering), Norbert Benda (Novartis), and Jürgen Läuter (University of Magdeburg). I explore consequences for setting up null hypotheses in situations where similar problems with directional alternative hypotheses might arise, for example in stratified clinical trials.

O'Neill RT (1997). Secondary endpoints cannot validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Contr. Clin. Trials 18, 550-556.
Bloch, D.A., Lai, T.L. and Tubert-Bitter, P. (2001) One-sided tests in clinical trials with multiple endpoints. Biometrics 57, 1039-1047.
Bloch, D.A., Lai, T.L., Su, Z. and Tubert-Bitter, P. A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. Statistics in Medicine (in preview) DOI: 10.1002/sim.2611
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800-802.
Holm, S.A. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6, 65-70
Läuter, J. (1996). Exact t and F tests for analyzing clinical trials with multiple endpoints. Biometrics 52, 964-970.
O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. Biometrics 40, 1079-1087.
Tamhane, A.C. and Logan, B.R. (2002) Accurate critical constants for the one-sided approximate likelihood ratio test for a normal mean vector when the covariance matrix is estimated. Biometrics 58, 650-656.
Tamhane, A.C. and Logan, B.R. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. Biometrika 91, 715-727.
Wang, S.-J.(1998). A closed procedure based on Follmann's test for the analysis of multiple endpoints. Communications in Statistics Theory and Methods 27, 2461-2480.

Mpm2C1T1

Multiple Comparisons for Ratios to the Grand Mean

Ludwig A. Hothorn, G. Dilba

Leibniz Universität Hannover, Germany

Multiple comparison for differences to the grand mean is a well-known approach and commonly used in quality control, see the recent textbook on ANOM (analysis of means) by Nelson et al. (2005). Alternatively, we discuss multiple comparisons for ratios to the grand mean: multiple tests and simultaneous confidence intervals. Simultaneous confidence intervals represent a generalization of Fieller intervals and plugging-in the estimated correlations into the multivariate-t distribution with arbitrarily correlation matrix. A related R program will be provided using the mvtnorm package by Hothorn et al 2001. The advantage of dimensionless confidence intervals will be demonstrated by examples for comparing several mutants or different varieties for multiple endpoints.

Hothorn T et al. (2001) On multivariate t and Gauss probabilities. R New 1 (2): 27-29.
Nelson PR et al. (2005) The analysis of means SIMA

Mpm2C1T2

Comparison of Methods for Estimating Relative Potencies in Multiple Bioassay Problems

Gemechis Dilba

Institute of Biostatistics, Leibniz University of Hannover, Germany

Relative potency estimations in both multiple parallel-line and slope-ratio assays involve construction of simultaneous confidence intervals for ratios of linear combinations of general linear model parameters. The key problem here is that of determining multiplicity adjusted percentage point of a multivariate t-distribution the correlation matrix *R* of which depends on the unknown ratio parameters. Several methods have been proposed in the literature on how to deal with *R*. Among others, conservative methods based on probability inequalities (e.g., Boole's and Sidak inequalities) and a method based on an estimate of *R* are used. In this talk, we explore and compare the various methods (including the delta approach) in a more comprehensive manner with respect to their simultaneous coverage probabilities via Monte Carlo simulations. The methods will also be evaluated in terms of confidence interval width through application to data on multiple parallel-line assay.

Mpm2C1T3

Multiple Hypothesis Testing to Establish Whether Treatment is "Better" Than Control

Aldo Solari, Salmaso Luigi, Pesarin Fortunato

Department of Chemical Process Engineering, University of Padova, Italy

Experiments are often carried out to establish whether treatment is "better" than control with respect to a multivariate response variable, sometimes referred to as multiple endpoints. However, in order to develop suitable tests, we have to specify the notion of "better". To formulate the problem, let X and Y denote the k-variate responses associated with control and treatment, respectively. We may be interested in testing H0: "X and Y are equal in distribution" against H1: "X is stochastically smaller than Y and not H0" where the definition of 'stochastically smaller' is given in [1]. If a test rejects H0, then it does not necessarily follow that there evidence to support H1, unless the latter is the complement of the null hypothesis [2]. Hence we must suppose that "X is stochastically smaller than Y" is known a priori, i.e. either H0 or H1 is true. Under this assumption, we prove that testing H0 against H1 is equivalent to the union-intersection (UI, [3]) testing formulation based on marginal distributions. However, this is not the only possible formulation for the treatment to be preferred to the control. It may be appropriate to show that the former is not inferior, i.e. not much worse, on any of the enpoints and is superior on at least one endpoint, resulting in an intersection-union (IU, [4]) combination of IU and UI testing problems [5]. For both formulations of "better", we propose a multiple testing procedure based on combining dependent permutation tests [6], and an application is presented.

[1] Marshall, A. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Applications. Academic Press, New York.
[2] Silvapulle, J.S. and Sen, P.K. (2005) Constrained Statistical Inference. Inequality, Order, and Shape Restictions. Wiley, New Jersey.
[3] Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathemathical Statistics, 24:220-238.
[4] Berger, R.L. (1982) Multiparameter hypothesis testing and acceptance sampling. Tecnometrics, 24:295-300.
[5] Röhmel, J., Gerlinger, C. Benda, N. and Läuter, J. (2006) On Testing simultaneously Non-Inferiority in Two Multiple Primary Endpoints and Superiority in at Least One of Them. Biometrical Journal, 48:916-933.
[6] F. Pesarin (2001) Multivariate Permutation Tests with Applications in Biostatistics. Wiley, Chichester.

Mpm2C1T4

Across and Down in Large SNP Studies: the MAX Test of Freidlin and Zheng vs SAS PROC CASECONTROL

Dana Aeschliman, Marie-Pierre Dube

Statistical Genetics Research Group, Montreal Heart Institute, Canada

SAS PROC CASECONTROL offers the user 3 statistical tests for assessing the association of a SNP and a binary phenotype: the allele, genotype and trend tests. Three important models of genotype-phenotype association are the recessive, additive and dominant genetic models. In a large SNP study, one is faced with both "across"and "down" aspects of the multiple testing problem. The MAX test of Freidlin et al. (Freidlin et al., 2002, Zheng and Gastwirth 2006) builds on the ideas of Armitage (1955), Sasieni (1997), and Slager and Schaid (2001) and offers a way of testing for recessive, additive, and dominant models while producing one P-value for each SNP. In this report, we compare the power of the MAX test to each of the 3 tests in SAS PROC CASECONTROL. We show that the MAX test compares very favorably. We developed a program in R to simulate genetic data sets of varying complexity. We provide two SAS MACROs that use only BASE SAS. One encodes the MAX test. The second MACRO acts as a wrapper for the first and encodes a step-down resampling algorithm, Westfall and Young's (1993) Algorithm 2.8, resulting in p-values which are corrected for the correlation between test statistics. We comment on the notion of subset pivotality as applied to this situation and discuss the treatment of missing values.

1. Zheng, G. and Gastwirth, J. (2006) On estimation of the variance in Cochran_Armitage trend tests for genetic association using case-control studies. Statistics in Medicine; 25(18): 3150-3159.
2. Freidlin, B et al. (2002) Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness. Human Heredity; 53,
3. Slager, S.L. and Schaid, D.J. (2001) Case-Control Studies of Genetic Markers: Power and Sample Size Approximations for Armitage's Test for Trend. Human Heredity; 52,
4. Sasieni, P.D. (1997) From genotypes to genes: Doubling the sample size. Biometrics; 53: 1253-1261. Armitage, P. (1955) Tests for linear trends in proportions and frequencies. Biometrics; 11: 1253-1261. Westfall,
5. P.H. and Young, S.S. (1993) Resampling-Based Multiple Testing. John Wiley and Sons, Inc.

Mpm2C2T1

Sample Size Calculation for Microarray Data Analysis Using Normal Mixture Model

Masaru Ushijima

Japanese Foundation for Cancer Research, Japan

Sample size calculation is an important procedure when designing a microarray study, especially for medical research. This paper concerns sample size calculation in the identification of differentially expressed genes between two patient groups. We use a mixture model, involving differentially expressed and non-differentially expressed genes.
To calculate the sample size, parameters to be given are as follows: (1) the number of differentially expressed genes, (2) the distribution of the true differences, (3) Type I error rate (e.g. FDR, FWER), (4) statistical power (e.g. sensitivity). We propose a sample size calculation method using FDR, family-wise power proposed by Tsai et al. (Bioinformatics, 2005, 21:1502-8), and a normal mixture model. The sample sizes for two-sample t-test are computed for several settings and the simulation studies are performed.

Mpm2C2T2

Estimating the Proportion of True Null Hypotheses with the Method of Moments

Jose Maria Muino, P. Krajewski

Institute of Plant Genetics, Poland

In order to construct the critical region for the test statistic in a multiple hypotheses testing situation, it is necessary to obtain some information about the distribution of the test statistic under the null hypothesis and under the alternative, and to use this information in an optimal way to asses which tests can be declared significant. We propose how to obtain this information in the form of the moments of these distributions and the proportion of true null hypotheses (π_0) with the method of moments. As a particular case, we study the properties of the estimator π_0 when the test statistic is the mean value, and we construct a new asymptotically unbiased (as the number of test goes to infinity) estimator. Some numerical simulation are done to compare the proposed method with others.

Mpm2C2T3

Knowledge-Based Approach to Handling Multiple Testing in Functional Genomics Studies

Adam Zagdanski, Przemyslaw Biecek, Rafal Kustra

University of Toronto, Canada and Wroclaw University of Technology, Institute of Mathematics and Computer Science, Poland

We propose a novel method for multiple testing problem inherent in functional genomics studies. One novelty of the method is that it directly incorporates prior knowledge about gene annotations to adjust the p-values. We describe general methodology to perform knowledge-based multiple testing adjustment and focus on an application of this approach in Gene Set Functional EnrichmentAnalysis (GSFEA). We apply and evaluate our method using a database of known Protein-Protein Interactions to perform large-scale gene function prediction. In this study Gene Ontology Biological Process (GO-BP) taxonomy is employed as the knowledge-base standard for describing gene functions. An extensive simulation study is carried out to investigate a behaviour of the proposed adjustment procedure under different scenarios. Empirical analysis, based on both real and simulated data, reveals that our approach yields an improvement of a number of performance criteria, including an empirical False Discovery Rate (FDR). We derive theoretical connections between our method and the stratified False Discovery Rate approach proposed by [1], and also describe similarities to the weighted p-value FDR control introduced recently by [2]. Finally we show how our method can be adopted to other multiple hypothesis problems where some form of prior information about the relationships among tests is available.

[1] L.Sun, R.V. Craiu, A.D. Paterson, S.B. Bull (2006) "Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies" Genet Epidemiol. 2006 Sep, 30(6):519-30.
[2] Ch.R. Genovese, K. Roeder and L.Wasserman (2006) "False discovery control with p-value weighting" Biometrika 2006, 93(3):509-524.

Mpm2C2T4

Aesthetics and Power in Multiple Testing – a Contradiction?

Gerhard Hommel (Invited Speaker)

IMBEI, University of Mainz, Germany

It seems to be desirable that a multiple testing procedure should be as powerful as possible, given a criterion for type I error control. However, there are important additional aspects to be considered: 1. the pattern of the decisions should be logical; 2. the decisions should be conceivable, e.g., for scientific reasons or within a clinical trial; and 3. the decisions should be taken in such a way that they can be communicated also to non-statisticians. Moreover, there are also aspects of aesthetics that can be considered as relevant. Aesthetics are certainly to some extent subjective, but many people would agree that the closure test or the Bonferroni-Holm procedure (say) have an aesthetic component. I will consider in my talk different concepts related to multiple testing procedures and discuss them under the aspects above. In particular, the following issues are discussed:

Coherence and consonance;
Monotonicity of decisions (dependent on p-values);
Exchangeability;
Criteria for control of type I errors;
Concepts of power.

To illustrate the ideas, I will consider the “fallback procedure” (Wiens, 2003; Wiens and Dmitrienko, 2005) as an example and discuss some properties of this procedure.

Wiens, B.L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. Pharm. Stat. 2, 211-215.
Wiens, B.L. and Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. J. Biopharm. Stat. 15, 929-942.

Tam1C1T1

A General Principle for Shortening Closed Test Procedures with Applications

Werner Brannath, Frank Bretz

Medical University of Vienna, Austria

The closure principle is a general, simple and powerful method for constructing multiple test procedures controlling the family wise error rate in the strong sense. In spite of its generality and simplicity, the closure principle has the disadvantage that the number of individual tests required for its completion increases exponentially with number of null hypotheses of primary interest. Hence, multiple test procedures based on the closure principle can require large computational efforts and may become infeasible for a large number of hypotheses and/or for computational intensive hypotheses tests, such as permutation or bootstrap tests.
Shortcut procedures have been proposed in the past, which substantially reduce the number of operations. In this presentation we provide a general principle for shortening closed tests. This principle provides a unified approach that covers many known shortcut procedures from the literature. As one application among others we derive a shortcut procedure for flexible two stage closed tests for which no shortcuts have been available yet.

Tam1C1T2

FDR-Control: Assumptions, a Unifying Proof, Least Favorables Configurations and FDR-Bounds

Helmut Finner, Thorsten Dickhaus, Markus Roters

German Diabetes Center, Leibniz Center, Heinrich-Heine-University Duesseldorf, Institute of Biometrics and Epidemiology, Germany

We consider multiple test procedures in terms of p-values based on a fixed rejection curveor a critical value function and study their FDR behavior. First, we introduce a series of assumptions concerning the underlying distributions and the structure of possible multiple test procedures. Then we give a short and unifying proof of FDR control for procedures (step-up, step-down, step-up-down) based on Simes' critical values for independent p-values and for a special class of dependent p-values considered in Benjamini and Yekutieli (2001), Sarkar (2002) and Finner, Dickhaus and Roters (2007). Moreover, we derive upper bounds for the FDR for non step-up procedures which can be calculated with respect to Dirac-uniform configurations. Finally, it will be shown that Dirac-uniform configurations are asymptotically least favorable for certain step-up-down procedures when the number of hypotheses tends to infinity.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29, 1165-1188.
Finner, H., Dickhaus, T. and Roters, M. (2007). Dependency and false discovery rate: Asymptotics. The Annals of Statistics, to appear.
Finner, H., Dickhaus, T. and Roters, M. (2007). On the false discovery rate and an asymptotically optimal rejection curve. Submitted for publication.
Sarkar, S. K. (2002) Some results on false discovery rate in stepwise multiple testing procedures. The Annals of Statistics 30, 239-257.

Tam1C1T3

Asymptotic Improvements of the Benjamini-Hochberg Method for FDR Control Based on an Asymptotically Optimal Rejection Curve

Thorsten Dickhaus, Helmut Finner, Markus Roters

German Diabetes Center, Leibniz Center, Heinrich-Heine-University Duesseldorf, Institute of Biometrics and Epidemiology, Germany

Due to current applications with a large number n of hypotheses, asymptotic control $(n \to \infty)$ of the false discovery rate (FDR) has become a major topic in the field of multiple comparisons. In general, the original linear step-up (LSU) procedure proposed in Benjamini & Hochberg (1995) does not exhaust the pre-specified FDR level, which gives hope for improvements with respect to power. Based on some heuristic considerations, we present a new rejection curve and implement this curve into several stepwise multiple test procedures for asymptotic FDR control. It will be shown that the new tests asymptotically exhaust the full FDR level under some extreme parameter configurations. This optimality leads to an asymptotic gain of power in comparison with the LSU procedure. For the finite case, we discuss adjustments both of the curve and of the procedures in order to provide strict FDR control.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289-300.
Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 4, 1165-1188.
Finner, H., Dickhaus, T. & Roters, M. (2007). On the false discovery rate and an asymptotically optimal rejection curve. Submitted for publication.
Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. Ann. Stat. 30, 1, 239-257.

Tam1C1T4

A Unified Approach to Proof of Concept and Dose Estimation for Categorical Responses

Bernhard Klingenberg

Williams College Williams College, United States

This talk suggests to unify dose-response modeling and target dose estimation into a single framework for the benefit of a more comprehensive and powerful analysis. Bretz, Pinheiro and Branson (Biometrics, 2006) recently implemented a similar idea for independent normal data by using optimal contrasts as a selection criterion among various candidate dose-response models. We suggest a framework in which from a comprehensive set of candidate models the ones are chosen that best pick up the dose-response. To decide which models, if any, significantly pick up the signal we construct the permutation distribution of the maximum penalized deviance over the candidate set. This allows us to find critical values and multiplicity adjusted p-values, controlling the error rate of declaring spurious signals as significant. A thorough evaluation and comparison of our approach to popular multiple contrast tests reveals that its power is as good or better in detecting a dose-response signal under a variety of situations, with many more additional benefits: It incorporates model uncertainty in proof of concept decisions and target dose estimation, yields confidence intervals for target dose estimates, allows for adjustments due to covariates and extends to more complicated data structures. We illustrate our method with the analysis of a Phase II clinical trial.

Tam1C2T1

Multiple Testing Procedures with Incomplete Data for Rank-based Tests of Ordered Alternatives

Paul Cabilio, Jianan Peng

Acadia University Acadia University, CANADA

Page (1963) and Jonckheere (1954) introduced tests for ordered alternatives in blocked experiments. Specifically, in the model with n blocks and t treatments, it is wished to test the hypothesis of no treatment effect against a specified ordered treatment effect with at least one inequality strict. Page proposed a statistic which can be expressed as the sum of Spearman correlations between each block and the criterion ranking chosen to be $(1,2,...,t)$, while Jonckheere proposed a statistic which is based on Kendall's tau correlation. These tests were extended in Alvo and Cabilio (1995) to the situation where only $k(i)$ treatment responses are observed in block i. For such incomplete blocks, the resulting extended Page statistic L^* differs from the one in the complete case in that the complete rank of a response in each block is replaced by a weight times a score which is either the incomplete rank of the response or the average rank $(k(i)+1)/2$, depending on whether or not the treatment is ranked in that block. If the null hypothesis is rejected, it is of interest to construct test procedures to identify which inequalities in the alternative are strict, and in so doing maintain the experimentwise error rate at a pre-assigned level. Our approach in this case is to modify one or more procedures that have been developed for detecting ordered means in the context of ANOVA (Nashimoto and Wright 2005.) The form of the extended Page statistic makes it possible to apply a general step-down testing procedure for multiple comparisons such as that proposed in Marcus, Peritz, and Gabriel (1976) for normal based tests. Specifically, we define a partition of the integers 1 to t into h sets of consecutive integers. For each set of integers in the partition we define an extended Page test statistic to test the sub-alternative hypothesis of ordered effects of treatments indexed by such integers. The intersection of such hypotheses over the partition can then be tested by the sum of such statistics. The procedure is to test all such hypotheses over all possible partitions. This approach may also be used for the extended Jonckheere statistic.

Tam1C2T2

Adjusting p-values of a Stepwise Generalized Linear Model

Chiara Brombin, Finos L., Salmaso L.

University of Padova, Italy

Stepwise variable selection methods are frequently used to determine the predictors of an outcome in generalized linear model (glm). Despite its widespread use, it is well know that the tests on the explained deviance of the selected model are biased. This arise from the fact that the ordinary test statistics upon which these methods are based were intended for testing pre-specified hypotheses; whereas the tested model is selected through a data-steered procedure. In this work we define and discuss a simple nonparametric procedure which corrects the p-value of the selected model of any stepwise selection method for glm. We also prove that this procedure fall in the class of weighted nonparametric combining functions defined by Pesarin [1] and extended in Finos and Salmaso [2]. The unbiasedness and consistency of the method is also proved. A simulation study also shows the validity of this procedure. Theorical differences with previous works in the same filed (Grachanovsky and Pinsker, [3]; Harshman and Lundy, [4]) are also provided. Free codes for R and Matlab are available and an application on a real dataset is presented.

[1] Pesarin, F. (2001). Multivariate Permutation tests: with application in Biostatistics. John Wiley & Sons, Chichester-New York.
[2] L. Finos, L. Salmaso (2006). Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations. Journal of Nonparametric Statistics 18, 2, 245–261.
[3] E. Grachanovsky, I. Pinsker(1995). Conditional p-values for the F-statistic in a forward selection procedure. Computational Statistics & Data Analysis 20, 239-263.
[4] R. A. Harshman, M. E. Lundy (2006). A randomization method of obtaining valid p-values for model changes selected "post hoc". http://publish.uwo.ca/~harshman/imps2006.pdf

Tam1C2T3

A Test Procedure for Random Degeneration of Paired Rank Lists

Michael G. Schimek, Peter Hall, Eva Budinska

IMI, Medical University of Graz, Austria

Let us assume two assessors (e.g. laboratories), at least one of which ranks *N* distinct objects according to the extent to which a particular attribute is present. The ranking is from 1 to *N*, without ties. In particular we are interested in the following situations: (i) The second assessor assigns each object to the one or the other of two categories (0-1-decision assuming a certain proportion of ones). (ii) The second assessor also ranks the objects from 1 to *N*. An indicator variable takes I_j=1 if the ranking given by the second assessor to the object ranked *j* by the first is not distant more than m, say, from *j,* and zero otherwise. For both situations our goal is to determine how far into the two rankings one can go before the differences between them degenerate into noise. This allows us to identify a sequence of objects that is characterized by a high degree of assignment conformity.

For the estimation of the point of degeneration into noise we assume independent Bernoulli random variables. Under the condition of a general decrease of p_j for increasing *j* a formal inference model is developed based on moderate deviation arguments implicit in the work of Donoho et al. (1995, JRSS, Ser. B 57, 301-369). This idealized model is translated into an algorithm that allows to adjust for irregular rankings (i.e. handling of quite different rankings of some objects) typically occuring in real data. A regularization parameter needs to be specified to account for the closeness of the assessors' rankings and the degree of randomness in the assignments. Our approach can be generalized to the case of more than two assessors. The class of problems we try to solve has various bioinformatics applications, for instance in the meta-analysis of gene expression studies and in the identification of microRNA targets in protein coding genes.

Tam1C2T4

Non-Negative Matrix Factorization and Sequential Testing

Paul Fogel, S. Stanley Young, NISS (possibly speaker)

Consultant, Paris, France

The “omic” sciences, transcriptomics, proteomics, metabalomics, all have data sets with *n* much lower than *p* leading to serious multiple testing problems. On the other hand, the coordination of biological action implies that there will be important correlation structures in these data sets. There is a need to take advantage of these correlations in any statistical analysis. We use non-negative matrix factorization to organize the predictors into sets. We alpha allocate over the sets and then test sequentially within each set. The within set testing is sequential so there is no need for multiple testing adjustment. We use simulations to demonstrate the increased power of our methods. We demonstrate our methods with a real data set using a SAS JMP script.

Tam1AT1

Multiple Testing Procedures for Hierarchically Related Hypotheses

Przemyslaw Biecek

Institute of Mathematics and Computer Science, Wroclaw University of Technology, Poland

In some genomic studies, the considered hypotheses are in a hierarchical relation. For example in Gene Set Functional Enrichment Analysis (GSFEA), we are confronted with a problem of testing thousands of hypotheses which correspond to different biological terms. Since the biological terms are hierarchically related, the corresponding hypotheses are also related. If biological term *f(i)* is more specific than biological term f(j), then the rejection of hypothesis $H_0(i)$ associated with the term *f(i)* implies the rejection of hypothesis $H_0(j)$ associated with the term *f(j)* (the relationship between biological attributes is defined by the Gene Ontology Biological Process (GO-BP) hierarchical taxonomy [1]). In this case, in addition to correction for number of hypotheses, we want to guarantee that testing outcomes are coherent with the relation among biological functions. Popular multiple testing procedures (eg. step-up, step-down or single step) do not guarantee the coherency. Moreover, methods designed for testing under hierarchical relation (see [2]) do not provide a control of FDR and also cannot be easily applied in the context of GSFEA. We propose a novel approach which incorporates knowledge about the relation among hypotheses. We consider an issue of testing a set of null hypotheses with a given hierarchical relation among them. The relation, represented by a directed acyclic graph (DAG), determines all possible outcomes of testing. It also leads to the two natural testing procedures (the follow up and the follow down) presented in this paper. For these procedures, we derive formulas for significance levels that provide a strong control of the three most popular error rates (FWER, PFER and FDR). We also present a simulation study for the proposed testing procedures, discuss their strengths and weaknesses and point out some applications.

[1] Harris, M. A., et al. (2004) ,,The Gene Ontology (GO) database and informatics resource.'' Nucleic Acids Res. 32(Database issue): D258–D261. doi: 10.1093/nar/gkh036
[2] Finner, H., Strassburger, K. (2002) ,,The partitioning principle: A powerful tool in multiple decision theory''. The Annals of Statistics, Vol. 30, No. 4, 1194–1213

Tam1AT2

On the Conservatism of the Multivariate Tukey-Kramer Procedure

Takahiro Nishiyama, Takashi Seo

Tokyo University of Science, Japan

We consider the conservative simultaneous confidence intervals for pairwise comparisons among mean vectors in multivariate normal distributions. The multivariate Tukey-Kramer procedure which is the multivariate version of Tukey-Kramer procedure is presented. Also, the affirmative proof of the multivariate version of the generalized Tukey conjecture of the conservativeness of the simultaneous confidence intervals for pairwise comparisons of four mean vectors is presented. Further, the upper bound for the conservativeness of the multivariate Tukey-Kramer procedure is also given in the case of four mean vectors. Finally, numerical results by Monte Carlo simulations are given.

Tam1AT3

Distribution Theory with Two Correlated Chi-Square Variables

Anwar H Joarder

King, Fahd University of Petroleum & Minerals Department of Mathemtical Sciences, Saudi Arabia

Ratios of two independent chi-square variables are widely used in statistical tests of hypotheses. This paper introduces a new bivariate chi-square distribution where the variables are not necessarily independent. Moments of the product and ratio of two correlated chi-square variables are outlined. Distributions of the sum and product of two correlated chi-squares are also derived.

AMS Mathematics Subject Classification: 60E05, 60E10, 62E15

Key Words and Phrases: Chi-square distribution, Wishart distribution, product moments, Bivariate distribution, Correlation

Tam1AT4

Resampling-Based Control of the False Discovery Rate under Dependence

Michael Wolf, Joseph Romano, Azeem Shaikh

University of Zurich, Switzerland

This paper considers the problem of testing *s* null hypotheses simultaneously while controlling the false discovery rate (FDR).
The FDR is defined to be the expected value of the fraction of rejections that are false rejections (with the fraction understood to be 0 in the case of no rejections). Benjamini and Hochberg (1995) provide a method for controlling the FDR based on p-values for each of the null hypotheses under the assumption that the p-values are independent. Subsequent research has since shown that this procedure is valid under weaker assumptions on the joint distribution of the p-values. Related procedures that are valid under no assumptions on the joint distribution of the p-values have also been developed. None of these procedures, however, incorporate information about the dependence structure of the test statistics. This paper develops methods for control of the FDR under weak assumptions that incorporate such information and, by doing so, are better able to detect false null hypotheses. We illustrate this property via a simulation study and an empirical application to the evaluation of hedge funds.

Tam2C1T1

To Model or Not to Model

Jason Hsu, Violeta Calian, Dongmei Li

The Ohio State University, U.S.A.

Re-sampling techniques are often used to estimate null distributions of test statistics in multiple testing. In the comparison of gene expressions of levels and in multiple endpoint problems, re-sampling is often used to take into account correlations among the observations. We describe how each of the re-sampling techniques: permutation of raw data, post-pivot of re-sampled test statistics, and re-sampling of pre-pivoted observations, each has its requirement of knowledge of the joint distribution of the test statistics for validity. Modeling is useful toward validating a re-sampling multiple testing technique. To the extent pre-pivot re-sampling is valid, for small samples it has some advantage of smoothness and stability of estimated null distributions.

Tam2C1T2

False Discovery Proportion Control under Dependence

Yongchao Ge

New York Mount Sinai School of Medicine, U.S.A.

In datasets involving the problem of multiple testing, we are interested to have statistical inferences of a) the total number m_1 of false null hypotheses, and b) the random variable false discovery proportion (FDP): the ratio of the total number of false positives to the total number of positives. The expectation of the FDP is the false discovery rate defined by Benjamini and Hochberg 1995. We describe a general algorithm to construct an upper prediction band for the FDPs and a lower confidence bound for m_1 simultaneously. This algorithm has three features:
i) resampling to incorporate the dependence information among the test statistics to improve power, ii) an appropriate normalization of the order test statistics or the numbers of false positives, and iii) carefully chosen rejection regions. Two interesting choices for normalizations are: standard normalization and quantile normalization. The former choice generalizes the maxZ procedure (Ge et al 05, Meinshausen and Rice 06) from independent to dependent data; and the latter improves the work by Meinshausen 06. The properties of these two choices of normalizations combined with other normalizations are compared with simulated data and microarray data.

Tam2C1T3

Resampling-Based Empirical Bayes Multiple Testing Procedure for Controlling the False Discovery Rate with Applications to Genomics

Houston Gilbert, Sandrine Dudoit, Mark J. van der Laan

Berkeley University of California, U.S.A.

We propose resampling-based empirical Bayes multiple testing procedures (MTP) for controlling a broad class of Type I error rates, defined as tail probabilities and expected values for arbitrary functions of the numbers of false positives and true positives [3, 4]. Such error rates include, in particular, the popular false discovery rate (FDR), defined as the expectedproportion of Type I errors among the rejected hypotheses. The approach involves specifying the following: (i) a joint null distribution (or estimator thereof) for vectors of null test statistics; (ii) a distribution for random guessed sets of true null hypotheses. A working model for generating pairs of random variables from distributions (i) and (ii) is a common marginal non-parametric mixture distribution for the test statistics. By randomly sampling null test statistics and guessed sets of true null hypotheses, one obtains a distribution for a guessed specific function of the numbers of false positives and true positives, for any given vector of cut-offs for the test statistics. Cut-offs can then be chosen to control tail probabilities and expected values of this distribution at a user-supplied level. We wish to stress the generality of the proposed resampling-based empirical Bayes approach:
(i) it controls tail probability and expected value error rates for a broad class of functions of the numbers of false positives and true positives;
(ii) unlike most MTPs controlling the proportion of false positives, it is based on a test statistics joint null distribution and provides Type I error control in testing problems involving general data generating distributions with arbitrary dependence structures among variables;
(iii) it can be applied to any distribution pair for the null test statistics and guessed sets of true null hypotheses, i.e., the common marginal non-parametric mixture model is only one among many reasonable working models that does not assume independence of the test statistics. Simulation study results indicate that resampling-based empirical Bayes MTPs compare favorably in terms of both Type I

error control and power to competing FDR-controlling procedures, such as those of Benjamini and Hochberg (1995) [1] and Storey (2002) [5]. The proposed MTPs are also applied to DNA microarray-based genetic mapping and gene expression studies in Saccharomyces cerevisiae [2].

1.) Y. Benjamini and Y. Hochberg. Contolling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 1995.
2.) R.B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci., 2005.
3.) S. Dudoit and M.J. van der Laan. Multiple Testing Procedures and Applications to Genomics. Springer, 2007. (In preparation).
4.) S. Dudoit, H.N. Gilbert and M.J. van der Laan. Resampling-based empirical Bayes multiple testing procedure for controlling the false discovery rate.
Technical report, Division of Biostatistics, University of California, Berkeley, 2007. (In preparation).
5.) J.D. Storey. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 2002.

Tam2C1T4

Confidence Sets Following a Modified Group Sequential Test

Hans-Helge Müller, Nina Timmesfeld

Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Germany

Consider the statistically monitoring of a clinical trial comparing two treatments where the confirmatory analysis is based on a carefully planned group sequential design. Let us look at the Brownian motion model with the drift parameter reflecting the treatment difference. From now on suppose that during the course of the trial a change of the group sequential design is advisable, however, that the effect size parameter measuring treatment differences can be retained unchanged. In order to control the type I error rate, it is necessary and sufficient to redesign the trial on the basis of the Conditional Rejection Probability (CRP) principle proposed by Müller and Schäfer (2004). In addition to decision making on a hypothesis testing paradigm, estimation of the effect size parameter with a confidence set is an important issue at the end of the trial.
Following a group sequential trial, the simple fixed sample confidence intervals are inadequate. Methods for the construction of confidence intervals reflecting early stopping for both, significance and futility, have been proposed, e.g. the confidence intervals by Tsiatis et al. (1984). Starting with a valid concept of estimation of confidence sets in group sequential testing, in this contribution it is shown how to accommodate with the issue of constructing confidence sets following a modified design using the flexible CRP approach. The application in clinical trials is illustrated for a survival study using the method by Tsiatis et al.. The method of transformation is discussed regarding the choice of group sequential confidence sets.

Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine 2004; 23: 2497-2508.
Tsiatis AA, Rosner GL, Metha CR. Exact confidence intervals following a group sequential test. Biometrics 1984; 40:797-803.

Tam2C2T1

Unbiased Estimation after Modification of a Group Sequential Design

Nina Timmesfeld, Schäfer, Helmut, Müller, Hans-Helge

Institut of Medical Biometry and Epidemiology, Philipps-University Marburg, Germany

It is well known that the classical group-sequential designs perform well in terms of expected sample size for various effect sizes, while the type I and type II error rates are controlled. For ethical and economical reasons such a design is chosen in many clinical trials. Although the planning of the study was carefully done, it might happen that a design change is reasonable. The design can be changed with control of the type I error rate by the method of Müller and Schäfer (2004) at any time during the course of the trial. At the end of a study additional inference is required such as confidence bounds and estimates for the effect size. In the case of group sequential designs an unbiased estimator can be obtained by the method of Liu and Hall(1999). In this talk, we will present a method to modify this estimator to keep the unbiasedness after design modifications, in particular after modification of the sample size.

Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine 2004; 23:2497–2508.
Liu A, Hall W. Unbiased estimation following a group sequential test. Biometrika 1999; 86:71–78.

Tam2C2T2

Homogeneity of Stages in Adaptive Designs

Andreas Faldum

IMBEI, Universitätsklinikum Mainz, Germany

Adaptive designs result in great flexibility in clinical trials and guarantee full control of type I error. Despite increasing interest, such designs are only hesitantly implemented in pharmaceutical trials. One possible reason is concern of the regulatory authorities. In a reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan [EMEA 06], the European Medicines Agency (EMEA) requests methods to assure comparable results of interim and end analysis. The authors point out that it might be difficult to interpret the conclusions from a trial if it is suspected that the observed discrepancies of stages are a consequence of dissemination of the interim results. EMEA states that the simple rejection of the global null hypothesis across all stages is not sufficient to establish a convincing treatment effect. In order to avoid jeopardizing the success of a trial by differing results of the stages, the probability of such discrepancies should be taken into account when planning a trial. In this talk we concentrate on two-stage adaptive designs. Boundaries for discrepant effect estimates of stages are given dependent on the p value of the first stage and the adaptive design selected. By choosing an appropriate adaptive design a rejection of the null hypothesis despite a relevantly reduced effect estimate in the second stage can be prevented. On the other hand, rejection of the null hypothesis with treatment effect estimates increasing relevantly over stages cannot reasonably be avoided. However, the probability of rejecting the null hypothesis with homogeneous effect estimates of both stages can be predetermined. The results can help to find an adaptive design, which prevents a relevant decrease of effect estimates in case of a significant trial success and reduces the probability of a random relevant increase in the effect estimate. The underlying analyses can be used as a basis for discussion with the regulatory authorities. The considerations proposed here will be clarified by examples.

EMEA (2006). Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan. CHMP/EWP/2459/02, end of consultation Sept 2006, http://www.emea.eu.int/pdfs/human/ewp/245902en.pdf.

Tam2C2T3

Controversy? What Controversy? - An Attempt to Structure the Debate on Adaptive Designs

Marc Vandemeulebroecke

Novartis Pharma AG, Switzerland

From their beginnings, concepts for consecutive analyses of accumulating data have evoked lively debate. Classical sequential analysis has been provocatively criticized, and group sequential approaches have been controversially discussed. Since recently, the merits and pitfalls of adaptive designs are passionately debated. Starting from striking examples, we will in this talk try to dissect the debate. We identify what we consider the main discussion points, sketch their scope, and ponder their relative importance. We propose to standardize and render more precisely the terminology. We hope that this can contribute to the creation of a frame of reference for the current controversy on adaptive designs.

Tam2C2T4

On Estimates of R-values in Selection Problems

Andreas Futschik

University of Vienna, Austria

In the context of selection, quantities analogous to p-values (called R-values) have been introduced by J. Hsu (1984). They may be interpreted as a measure of evidence for rejecting (i.e. not selecting) a population. As in multiple hypothesis testing when p-values are corrected for multiplicity, these R-values can be quite conservative in high dimensional settings unless the parameters are close to the least favorable configuration.
We propose estimates of R-values that are less conservative and investigate their behavior. They also lead to selection rules for high dimensional problems.

Tam2AT1

Screening for Partial Conjunction Hypotheses

Ruth Heller, Benjamini, Yoav

Department of Statistics and Operations Research, Tel-Aviv University, Israel

We consider the problem of testing the partial conjunction null, that asks whether less than *u* out of *n* null hypotheses are false. It offers an in-between approach to the testing of the global null that all *n* hypotheses are null, and the full conjunction null that not all of the *n* hypotheses are false. We address the problem of testing many partial conjunction hypotheses simultaneously, a problem that arises when combining maps of p-values. We suggest powerful test statistics that are valid under dependence between the test statistics as well as under independence. We suggest controlling the false discovery rate (FDR) on the p-values for testing the partial conjunction hypotheses, and we prove that the BH FDR controlling procedure remains valid under various dependency structures. We apply the method to examples from Microarray analysis and functional Magnetic Resonance Imaging (fMRI), two application areas where the need for partial conjunction analysis has been identified.

Tam2AT2

Exact Simultaneous Confidence Bands for Multiple Linear Regression over an Ellipsoidal Region

Shan Lin, Wei Liu

University of Southampton, UK

A simultaneous confidence band provides useful information on whereabouts of the true regression function. Construction of simultaneous confidence bands has a history going back to Working and Hotelling (1929) and is a hard problem when the predictor space is restricted in some region and the number of regression covariates is more than one. This talk gives the construction of exact one-sided and two-sided simultaneous confidence bands for a multiple linear regression model over an ellipsoidal region that is centered at the point of the means of the predictor variables in the experiment based on three methods, i.e., the method of Bohrer (1973), the algebraical method and the tubular neighborhood method. Also, it is of interest to show these three methods give the same result.

Tam2AT3

Stepwise Confidence Intervals for Monotone Dose-Response Studies

Jianan Peng, Chu-In Charles Lee, Karolyn Davis

Acadia University, Canada

In dose-response studies, one of the most important issues is the identification of the minimum effective dose (MED), where the MED is defined as the lowest dose such that the mean response is better than the mean response of a zero-dose control by a clinically significant difference. Usually the dose-response curves are monotonic. Various authors have proposed step-down test procedures based on contrasts among the sample means to find the MED. In this paper, we improve Marcus and Peritz's method (1976, Journal of Royal Statistical Society, Series B, Vol 38, 157-165) and combine Hsu and Berger's DR method (1999, Journal of the American Statistical Association, Vol 94, 468-482) to construct the lower confidence bound for the difference between the mean response of any non-zero dose level and that of the control under the monotonicity assumption to identify the MED. The proposed method is illustrated by numerical examples and simulation studies on power comparisons are presented.

Tam2AT4

Control of Generalized Error Rates in Multiple Testing

Joseph P. Romano (Invited Speaker)

Stanford University, U.S.A.

Consider the problem of testing *s* hypotheses simultaneously. The usual approach restricts attention to procedures that control the probability of even one false rejection, the familywise error rate (FWER). If *s* is large, one might be willing to tolerate more than one false rejection, thereby increasing the ability of the procedure to correctly reject false null hypotheses. One possibility is to replace control of the FWER by control of the probability of *k* or more false rejections, which is called the *k*-FWER. We derive both single-step and stepdown procedures that control the *k*-FWER in finite samples or asymptotically, depending on the situation. We also consider the false discovery proportion (FDP) defined as the number of false rejections divided by the total number of rejections (and defined to be 0 if there are no rejections). The false discovery rate proposed by Benjamini and Hochberg controls E(FDP). Here, the goal is to construct methods which satisfy, for a given γ and α, $P\{FDP > \gamma\} \leq \alpha$, at least asymptotically. In contrast to Bonferroni type methods, we construct methods (using resampling) that implicitly take into account the dependence structure of the individual test statistics in order to further increase the ability to detect false null hypotheses. This feature is also shared by related work of van der Laan, Dudoit and Pollard, but our methodology is quite different. Simulations demonstrate improved performance over currently available methods. (This talk is based on joint work with Michael Wolf of the University of Zurich.)

Tpm1C1T1

Comparing Mutliple Tests for Separating Populations

Juliet Shaffer

Berkeley University of California, U.S.A.

Most studies for comparing multiple test procedures for finding differences among populations concentrate on the number of true and false differences that are significant, the former as a measure of power, the latter or a combination of both in various forms as a measure of error. For researchers, the configuration of results, e.g. the extent to which they divide populations into nonoverlapping classes, may be as important as or more important than the actual numbers. Results that lead to separations of populations into groups, when accurate, are especially useful. The talk will discuss some new measures of such separability and compare different multiple testing methods on these measures.

Tpm1C1T2

A Leave-P-Out Based Estimation of the Proportion of Null Hypotheses in Multiple Testing Problems

Alain CELISSE

UMR 518 AgroParisTech / INRA MIA, France

A large part of the literature have been devoted to multiple testing problems since the introduction of the False Discovery Rate (FDR) by Benjamini and Hochberg (1995). In this seminal paper, authors provide a procedure that enables control of the FDR at a pre-specified level. However an improvement of the method in terms of power is possible thanks to the introduction of an estimate of the unknown proportion of true null hypotheses: π_0. We propose an estimator of this proportion that relies on both density estimation by means of irregular histograms and exact leave-p-out cross-validation.
We estimate first the density of p-values from a collection of irregular histograms among which we select the best estimator in terms of minimization of the quadratic risk. The estimate of π_0 is deduced as the height of the largest column of the selected histogram. An estimator of the risk is obtained by use of leave-p-out cross-validation. We present a closed formula for this risk estimator and an automatic choice of the parameter p in the leave-p-out. It consists in minimizing the mean square error (MSE) of the leave-p-out risk estimator.
Besides, recent papers have pointed out that the use of two-sided statistics in one-sided tests entails p-values corresponding to false positives near to 1. Whereas most of the existing estimators do not take this phenomenon into account, leading to systematic overestimation, our estimator of the proportion remains accurate in such situations.
Eventually, we compare our procedure with existing ones in simulations, showing as well how problematic false positives near 1 may be. The proposed estimator seems more accurate in terms of variability for instance. Better FDR estimations are obtained.

Tpm1C1T3

Repeated Significance Tests Controlling the False Discovery Rate

Martin Posch, Sonja Zehetmayer, Peter Bauer

Medical University of Vienna, Austria

When testing a single hypothesis repeatedly at several interim analyses, adjusted significance levels have to be applied at each interim look to control the overall Type I Error rate. There is a rich literature on such group sequential trials investigating the choice and computation of adjusted critical values. Surprisingly, if a large number of hypotheses are tested controlling the False Discovery Rate, we can show that under quite general conditions no adjustment of the critical value for multiple interim looks is necessary. This holds asymptotically (for a large number of hypotheses) under all scenarios but the global null hypothesis where all null hypotheses are true. Similar results are given for a procedure controlling the per-comparison error rate.

Tpm1C1T4

Simultaneous Inference for Ratios

David Hare, John Spurrier

University of Louisiana, Monroe, U.S.A.

Consider a general linear model with p-dimensional parameter vector β and i.i.d. normal errors. Let K_1, ..., K_k, and L be linearly independent vectors of constants such that $L^T\beta \neq 0$. We describe exact simultaneous tests for hypotheses that $K_i^T\beta/L^T\beta$ equal specified constants using one-sided and two-sided alternatives, and describe exact simultaneous confidence intervals for these ratios. In the case where the confidence set is a single bounded contiguous set, we describe what we claim are the best possible conservative simultaneous confidence intervals for these ratios - best in that they form the minimum k-dimensional hypercube enclosing the exact simultaneous confidence set. We show that in the case of $k = 2$, this "box" is defined by the minimum and maximum values for the two ratios in the simultaneous confidence set and that these values are obtained via one of two sources: either from the solutions to each of four systems of equations or at points along the boundary of the simultaneous confidence set where the correlation between two t variables is zero. We then verify that these intervals are narrower than those previously presented in the literature.

Tpm1C2T1

Neglect of Multiplicity in Hypothesis Testing of Correlation Matrices

Burt Holland

Temple University, U.S.A.

Many social science journals publish articles with correlation matrices accompanied by tests of significance that ignore multiplicity. A highly cited article in Psychological Methods recommended use of an MCP when testing correlations but promoted MCP procedures that are inapplicable to correlations. We discuss viable options for handling this problem.

Tpm1C2T2

Minimum Area Confidence Set Optimality for Confidence Bands in Simple Linear Regression

Wei Liu, A. J. Hayter

S3RI and School of Maths University of Southampton, United Kingdom

The average width of a simultaneous confidence band has been used by several authors (e.g. Naiman, 1983, 1984, Piegorsch, 1985a) as a criterion for the comparison of different confidence bands. In this paper, the area of the confidence set corresponding to a confidence band is used as a new criterion. For simple linear regression, comparisons have been carried out under this new criterion between hyperbolic bands, two-segment bands, and three-segment bands, which include constant width bands as special cases. It is found that if one requires a confidence band over the whole range of the covariate, then the best confidence band is given by the Working & Hotelling hyperbolic band. Furthermore, if one needs a confidence band over a finite interval of the covariate, then a restricted hyperbolic band can again be recommended, although a three-segment band may be very slightly superior in certain cases.

Tpm1C2T3

Schéffe Type Multiple Comparison Procedure in Order Restricted Randomized Designs

Omer Ozturk, Steve MacEachern

The Ohio State University, Ohio, U.S.A.

Ozturk and MacEachern (2004) introduced a new design, the order restricted randomized design (ORRD), for the contrast parameters in a linear model. This new design uses a restricted randomization scheme that relies on subjective judgment ranking of the experimental units based in their inherent heterogeneity (or homogeneity). The process of judgment ranking creates a positive correlation structure among within set units and the restricted randomization on these ranked units translates this positive correlation into a negative one when estimating a contrast. Hence, the design serves as a variance reduction technique for treatment contrasts.

In this talk, we first develop a test for the generalized linear hypothesis based on an ORRD and discuss how this test can be used to test the treatment effects. We then develop a Schéffe-type multiple comparison procedure for all possible contrasts of the treatment effects. We show that the coefficients of contrasts depend on the design matrix and the underlying covariance structure of the judgment ranked observations. A simulation study shows that the multiple comparison procedure is robust against wide range of underlying distributions.

Tpm1C2T4

The Multiple Confidence Procedure and its Applications

Tetsuhisa Miwa

National Institute for Agro-Environmental Sciences, Japan

In 1973 Takeuchi proposed a multiple confidence procedure for multiple decision problems in his book “Studies in Some Aspects of Theoretical Foundations of Statistical Data Analysis” (in Japanese). This procedure is based on the partition of the parameter space. Therefore it is closely related to the recent development of the partitioning principles. In our talk we first review the basic concepts of Takeuchi's multiple confidence procedure. Then we discuss some applications and show the usefulness of the procedure.

Tpm1AT1

Gate-Keeping Testing without Tears

David Li, Mehrotra, Devan

Merck Research Labs, U.S.A.

In a clinical trial, there are one or two primary endpoints, and a few secondary endpoints. When at least one primary endpoint achieves statistical significance, there is considerable interest in using results for the secondary endpoints to enhance characterization of the treatment effect. Because multiple endpoints are involved, regulators may require that the trial-wise type-I error rate be controlled at a pre-set level. This requirement can be achieved by using "gate-keeping" methods. However, existing methods suffer from logical oddities such as allowing results for secondary endpoint(s) to impact the likelihood of success for the primary endpoint(s). We propose a novel and easy-to-implement gate-keeping procedure that is devoid of such deficiencies. Simulation results and real data examples are used to illustrate efficiency gains of our method relative to existing methods.

Tpm1AT2

An Application of the Closed Testing Principle to Enhance One-Sided Confidence Regions for a Multivariate Location Parameter

Michael Vock

Institute of Mathematical Statistics, University of Bern, Switzerland

If a one-sided test for a multivariate location parameter is inverted, the resulting confidence region may have an unpleasant shape. In particular, if the null and alternative hypothesis are both composite and complementary, the confidence region usually does not resemble the alternative parameter region in shape, but rather a reflected version of the null parameter region.
We illustrate this effect and show one possibility of obtaining confidence regions for the location parameter that are smaller and have a more suitable shape for the type of problems investigated. This method is based on the closed testing principle applied to a family of nested hypotheses.

Tpm1AT3

A Procedure to Multiple Comparisons of Diagnostic Systems

Ana Cristina Braga, Lino A. Costa e Pedro N. Oliveira

University of Minho, Portugal

In this work, a method for the comparison of two diagnostic systems based on ROC curves is presented. ROC curves analysis is often used as a statistical tool for the evaluation of diagnostic systems. For a given test, the compromise between the False Positive Rate (FPR) and True Positive Rate (TPR) can be graphically presented through a ROC curve. However, in general, the comparison of ROC curves is not straightforward, in particular, when they cross each other. A similar difficulty is also observed in the multi-objective optimization field where sets of solutions defining fronts must be compared in a multi-dimensional space. Thus, the proposed methodology is based on a procedure used to compare the performance of distinct multi-objective optimization algorithms. Traditionally, methods based on the area under the ROC curves are not sensitive to the existence of crossing points between the curves. The new approach can deal with this situation and also allows the comparison of partial portions of ROC curves according to particular values of sensitivity and specificity, of practical interest. For illustration purposes, real data from Portuguese hospital was considered.

Tpm1AT4

Multiple Testing of General Contrasts: Truncated Closure and the Extended Shaffer-Royen Method

Peter H. Westfall (Invited Speaker), Randall D. Tobias

Texas Tech University, U.S.A.

Powerful improvements are possible for multiple testing procedures when the hypotheses are logically related. Closed testing with alpha-exhaustive tests provides a unifying framework for developing such procedures, but can be computationally difficult and can be "nonmonotonic in p-values". Royen (1989) introduced a "truncated" closed testing method for the case of all pairwise comparisons in the ANOVA that is monotonic in p-values. Shaffer (1986) developed a similar truncated procedure for more general comparisons, but using Bonferroni tests rather than alpha-exhaustive tests, and Westfall (1997) extended Shaffer's method to allow alpha-exhaustive tests. This paper extends Royen's method to general contrasts and proves that it is equivalent to the extended Shaffer procedure. For large number k of contrasts, the method generally requires evaluation of $O(2^k)$ critical values corresponding to subset intersection hypotheses, and is computationally infeasible for large k. The set of intersections is represented using a tree structure, and a branch-and-bound algorithm is used to search the tree and reduce the $O(2^k)$ complexity by obtaining conservative "covering sets" that retain control of the familywise type I error rate (FWE). The procedure becomes less conservative the deeper the tree search, but computation time increases. In some cases with logical relations, even the more conservative covering sets provide much more power than standard methods.

The method is general, computable, and often much more powerful than commonly-used methods for multiple testing of general contrasts, as shown by applications to pairwise comparisons and response surfaces. In particular, with response surface tests, the method is computable with complete tree search, even when k is large, and can make many more "discoveries" than the standard FDR-controlling method.

The Extended Shaffer-Royen method has recently been hard-coded

in the SAS/STAT procedure PROC GLIMMIX; syntax and output will be shown.

Westfall, P.H. and Tobias, R.D. (2007). Multiple Testing of General Contrasts: Truncated Closure and the Extended Shaffer-Royen Method, to appear in Journal of the American Statistical Association.

Tpm2C1T1

Powerful Short-Cuts for Gatekeeping Procedures

Frank Bretz, Gerhard Hommel, Willi Maurer

Novartis Pharma AG, Switzerland

We present a general testing principle for a class of multiple testing problems based on weighted hypotheses. Under moderate conditions, this principle leads to powerful consonant multiple testing procedures. Furthermore, short-cut versions can be derived, which simplify substantially the implementation and interpretation of the related test procedures. It is shown that many well-known multiple test procedures turn out to be special cases of this general principle. Important examples include gatekeeping procedures, which are often applied in clinical trials when primary and secondary objectives are investigated, and multiple test procedures based on hypotheses which are completely ordered by importance. We illustrate the methodology with two real clinical studies.

Tpm2C1T2

Simultaneous Confidence Regions Corresponding to Holm's Stepdown Multiple Testing Procedure

Olivier Guilbaud

AstraZeneca R&D, Sweden

The problem of finding simultaneous confidence regions corresponding to multiple testing procedures (MTPs) is of considerable practical importance. Such confidence regions provide more information than the mere rejections/acceptances of null hypotheses that can be made by MTPs. I will show how one can construct simultaneous confidence regions for a finite number of quantities of interest that correspond to Holm's (1979) step-down multiple-testing procedure. Holm's MTP is an important and widely used generalization of the Bonferroni MTP. As the Bonferroni and Holm MTPs, the proposed confidence regions are quite flexible and generally valid. They are based on marginal confidence regions for the quantities of interest, and the only essential assumption for their validity is that the marginal confidence regions are valid. The estimated quantities, as well as the marginal confidence regions, can be of any kinds/dimensions. The proposed simultaneous confidence regions are of particular interest when one aims at confidence statements that will “show” that quantities belong to target regions of interest.

Tpm2C1T3

Compatible Simultaneous Lower Confidence Bounds for the Holm Procedure and other Closed Bonferroni Based Tests

Klaus Strassburger, Frank Bretz

German Diabetes Center, Leibniz-Institute at the Heinrich–Heine-University, Düsseldorf, Germany

In this contribution we present simultaneous confidence intervals being compatible with a certain class of one-sided closed test procedures using weighted Bonferroni tests for each intersection hypothesis. The class of multiple test procedures covered in this talk includes gatekeeping procedures based on Bonferroni adjustments, fixed sequence procedures, the simple weighted or unweighted Bonferroni procedure by Holm and the fallback procedure. These procedures belong to a class of short cut procedures, which are easy to implmenet. It will be shown that the corresponding confidence bounds have a straight forward representation. For the step-down procedure of Holm we illustrate the construction of compatible confidence bounds with a numerical example. The resulting bounds will be compared with those of the classical single-step procedure. Assets and drawbacks will be discussed.

Tpm2C1T4

Detecting Differential Expression in Microarray Data: Outperforming the Optimal Discovery Procedure

Alexander Ploner, Elena Perelman, Stefano Calza, Yudi Pawitan

Karolinska Institutet MEB, Sweden

The identification of differentially expressed genes among the tens of thousands of sequences measured by modern microarrays presents an obvious and serious multiplicity problem. The central role of gene expression data in molecular biology has stimulated much research in addressing this issue over the last decade; an important result of that research is the Optimal Discovery Procedure (ODP) proposed by John Storey, which generalizes the likelihood ratio test statistic of the Neyman-Pearson lemma for multiple parallel hypotheses, and which can be shown to be optimal in the sense that for any fixed number of false positive results, ODP will identify the maximum number of true positives [1].

However, the optimality result derived in [1] assumes exact knowledge of a large number of nuisance parameters that have to be estimated for any realistic application. In our talk, we will demonstrate that the practical implementation of ODP described in [2] is less powerful than a variant of the local false discovery rate we have proposed recently, which uses the distribution of the same nuisance parameters to weight conventional t-statistics [3]. We also show how a combination of the ODP test statistic with our weighting scheme can even further improve the power to detect differentially expressed genes at controlled levels of false discovery.

[1] Storey JD: The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing. UW Biostatistics Working Paper Series 2005, Working Paper 259.
[2] Storey JD, Dai JY, Leek JT: The Optimal Discovery Procedure for Large-Scale Significance Testing, with Applications to Comparative Microarray Experiments. UW Biostatistics Working Paper Series 2005, Working Paper 260.
[3] Ploner A, Calza S, Gusnanto A, Pawitan Y: Multidimensional local false discovery rate for microarray studies. Bioinformatics 2006, 22(5):556–565.

Tpm2C2T1

Flexible Two-Stage Testing in Genome-Wide Association Studies

André Scherag, Helmut Schäfer, Hans-Helge Müller

Institute of Medical Biometry and Epidemiology, Philipps-University of Marburg, Germany

Genome-wide association studies have been suggested to unravel the genetic etiology of complex human diseases [1]. Typically, these studies employ a multi-stage plan to increase cost-efficiency. A large panel of markers is examined in a subsample of subjects, and the most promising markers will also be genotyped in the remaining subjects.
Until now all proposed design require adherence to formal statistical rules which may not always meet the practical necessities of ongoing genetic research. In practice, investigators may e.g. wish to base the genetic marker selection on other criteria than formal statistical thresholds.
In this talk we describe an algorithm that enables various design modifications at any time during the course of the study. Using the Conditional Rejection Probability approach [2] the family-wise type I error rate is strongly controlled. The algorithm can deal with an extremely large number of hypotheses tests though requiring very limited computational resources. This algorithm is evaluated by simulations. Furthermore, we present a real data application.

[1] Freimer NB, Sabatti C. Human genetics: variants in common diseases. Nature. 2007 Feb 22;445(7130):828-30.
[2] Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. Stat Med. 2004 Aug 30;23(16):2497-508.

Tpm2C2T2

Sequential Genome-Wide Association Studies for Pharmacovigilance

Patrick Kelly

University of Reading, UK

Pharmacovigilance, the monitoring of adverse events, is an integral part in the clinical evaluation of a new drug. Until recently, attempts to relate the incidence of adverse events to putative causes have been restricted to the evaluation of simple demographic and environmental factors. The advent of large-scale genotyping, however, provides an opportunity to look for associations between adverse events and genetic markers, such as single nucleotides polymorphisms (SNPs). It is envisaged that a very large number of SNPs, possibly over 500,000, will be used in pharmacovigilance in an attempt to identify any genetic difference between patients who have experienced an adverse event and those who have not.

This paper presents a sequential genome-wide association test for analysing pharmacovigilance data as adverse events arise, allowing evidence-based decision-making at the earliest opportunity. This gives us the capability of quickly establishing whether there is a group of patients at high-risk of an adverse event based upon their DNA. The method uses permutations and simulations in order to obtain valid hypothesis tests which are adjusted for both linkage disequilibrium and multiple testing. Permutations are used to calculate p-values because the asymptotic properties of the test statistic are unlikely to hold due linkage disequilibrium. Simulations are used to find the required nominal significance level in order to satisfy some overall type I error rate. The simulations provide a simple and easy approach for obtaining a correction for the multiple testing without having to determine how the repeated tests are correlated.

Tpm2C2T3

FDR Control for Discrete Test Statistics

Anja Victor, Scheuer C, Cologne J, Hommel G

Institute of Medical Biometry, Epidemiology and Informatics, University Clinic Mainz, Germany

In genetic association studies considering e.g. Single Nucleotide Polymorphisms (SNPs) one deals with categorical data and dependencies between SNPs may occur (because of linkage Disequilibrium, LD). Additionally genetic association studies exhibit many different study situations ranging from genomewide scans to the examination of just a few selected candidate loci. The proportions of true null hypotheses will vary greatly between these situations, which influences FDR control. We will focus on multiple testing procedures that take the categorical structure of the SNP data into account. The most popular FWER control procedure for discrete data is Tarone's procedure (Tarone 1990). However Tarone's procedure is not monotone in the a-level. Therefore Hommel & Krummenauer published an improvement (Hommel & Krummenauer 1998). Recently Gilbert (Gilbert 2005) transferred Tarone's procedure to FDR control by explorative Simes procedure (Simes 1986, Benjamini & Hochberg 1995). However in Gilbert's procedure the finally attained boundary for the p-values by Simes procedure can be higher than the boundary for the previous selection of hypotheses for the „Tarone subset", such that no rejection may occur for small p-values outside the "Tarone subset" but for larger ones inside. We discuss ideas how the Hommel & Krummenauer procedure can be extended to FDR control and how Gilbert's procedure can be improved. Additionally we examine the advantages of using test procedures adapted to discrete test statistics in genetic association studies. Therefore we compare Gilbert's FDR-controlling procedure with the Homme l& Krummenauer procedure and additionally with classical FWER controlling procedures and the classical FDR controlling procedure. Results suggest that increase in power by exploiting the discrete nature can only be achieved when the number of subjects is small. Superiority of FDR control is more prominent if a larger proportion of null hypotheses are false.

Benjamini Y., and Hochberg Y. (1995) Controlling the false discovery rate: a practical and

powerful approach to multiple testing. JRSS B, 57, 289–300.
Gilbert PB. (2005) A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. Applied Statistics 44, 143–158.
Hommel, G. and Krummenauer, F. (1998) Improvements and modifications of Tarone's multiple test procedure for discrete data. Biometrics 54,673–681.
Simes, RJ. (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.
Tarone, RE. (1990) A modified Bonferroni method for discrete data. Biometrics 46: 515–522

Tpm2C2T4

Multiple Treatment Comparison Based on a Non-Linear Binary Dynamic Model

Brajendra Sutradhar, Vandna Jowaheer

Memorial University of Newfoundland, Canada

When an individual patient receives one of the multiple treatments and provides repeated binary responses over a small period of time, the efficient comparison of the treatment effects requires to take the longitudinal correlations of the binary responses into account. In this talk, we use a non-linear binary dynamic model that allows the full range for correlations and estimate the regression effects including the treatment effects by using the GQL (generalized quasilikelihood) approach that provides consistent as well as efficient estimates. We then demonstrate how to test the treatment effects based on the asymptotic distributions of their estimators.

Tpm2AT1

On Multiple Treatment Effects in Adaptive Clinical Trials for Longitudinal Count Data

Vandna Jowaheer, Brajendra C. Sutradhar

University of Mauritius, Mauritius

In longitudinal adaptive clinical trials it is an important research problem to compare more than two treatments for the purpose of treating maximum number of patients with the best possible treatment. Recently, in the context of longitudinal adaptive clinical trials for count responses, Sutradhar and Jowaheer (2006) [SJ (2006)] introduced a simple longitudinal play-the-winner (SLPW) design for the treatment selection for an incoming patient and discussed a weighted generalized quasilikelihood (WGQL) approach for consistent and efficient estimation of the regression effects including the treatment effects. Their study however was confined to the comparison of two treatments. In this paper, we generalize their SLPW design for the two treatment case to the multiple treatment case. For the estimation of the treatment effects we provide a conditional WGQL (CWGQL) as well as an unconditional WGQL approach. Both approaches provide consistent and efficient estimates for the treatment effects, the CWGQL being simpler but slightly unstable as compared to the unconditional WGQL approach where we use the limiting weights for the treatment selection. A normality based asymptotic test for testing the equality of the treatment effects is also outlined.

Tpm2AT2

Multi-Treatment Optimal Response-Adaptive Designs for Continuous Responses

Atanu Biswas, Saumen Mandal

Indian Statistical Institute, Kolkata, India

Optimal response-adaptive designs in phase III clinical trial set up are becoming more and more current interest. Most of the available designs are not from any optimal consideration. An optimal design for binary responses is given by Rosenberger et al. (2001) and an optimal design for continuous responses is provided by Biswas and Mandal (2004). Recently, Zhang and Rosenberger (2006) provided another design for normal responses. The present paper deals with some shortcomings of the earlier works and then extends the present approach for more than two treatments. The proposed methods are illustrated using some real data.

Tpm2AT3

On Identification of Inferior Treatments Using the Newman-Keuls Type Procedure

Samuel Wu, Weizhen Wang; David Annis

University of Florida, U.S.A.

We are concerned with selecting a subset of treatments such that the probability of including ALL best treatments exceeds a prespecified level. In this paper, we provide a stochastic ordering of the Studentized range statistics under a balanced one-way anova model. Based on this result we show that, when restricted to the multiple comparisons with the best, the Newman-Keuls type procedure strongly controls experimentwise error rate for a sequence of null hypotheses regarding the number of largest treatment means.

Tpm2AT4

Bayesian Adjusted Inference for Selected Parameters

Daniel Yekutieli (Invited Speaker)

Tel Aviv University, Israel

Benjamini and Yekutieli suggested viewing FDR adjusted inference as marginal inference for selected parameters. I will explain this approach – focusing on its weaknesses. I will then argue that overlooking the problem of selective inference is an abuse of Bayesian methodology; and introduce "valid" Bayesian inference for selected parameters. I will then show that these methods are straightforward generalizations of Bayesian FDR. To make the discussion clearer I will demonstrate use of the Bayesian adjustments on microarray data.

Wam1C1T1

A Bayesian Spatial Mixture Model for fMRI Analysis

Brent Logan, Maya P. Geliazkova, Daniel B. Rowe, Prakash W. Laud

Division of Biostatistics, Medical College of Wisconsin, U.S.A.

One common objective of fMRI studies is to identify voxels or points in the brain, which are activated by a neurocognitive task. This is an important multiple comparisons problem, since typically inference (often using z- or t- tests) is performed on each of thousands or hundreds of thousands of voxels. The false discovery rate has been studied for use in this problem by several authors. Finite mixture models have also been proposed to address the multiplicity issue, where voxels are classified according to being activated or not activated by the cognitive task. Links between the false discovery rate and mixture models have been shown in the literature. One limitation to these procedures is that activation is typically expected to occur in clusters of neighboring voxels rather than in isolated single voxels; methods which do not account for this may have lower sensitivity to activation. We propose a Bayesian spatial mixture model to address these issues. Each voxel has an unknown or latent activation status, denoted by a binary activation variable. The spatial model for the binary activation indicators is induced by a latent Gaussian spatial process (a conditional autoregressive, or CAR, model), thresholded to produce the binary activation, analogous to a spatial probit model. An efficient Gibbs sampling algorithm is developed to implement the model, yielding posterior probabilities of activation for each voxel, conditional on the observed data. We apply this method to a real fMRI study, and compare its performance in simulation with other methods proposed for fMRI analysis.

Wam1C1T2

A Bayesian Screening Method for Determining if Adverse Events Reported in a Clinical Trial are Likely to be Related to Treatment

A Lawrence Gould

Merck Research Laboratories, Inc. UG-1D88, U.S.A.

Many different adverse events usually are reported in large-scale clinical trials. Most of the events will not have been identified a priori. Current analysis practice often applies Fisher's exact test to the usually relatively small event counts, with a conclusion of “safety” if the finding does not reach statistical significance. This practice has serious disadvantages: lack of significance does not mean lack of risk, the various tests are not adjusted for multiplicity, and the data determine which hypotheses are tested. This presentation describes a new approach that does not test hypotheses, is self-adjusting for multiplicity, and has well-defined diagnostic properties. The approach is a screening approach that uses Bayesian model selection techniques to determine for each adverse event the likelihood that the occurrence is treatment-related. The approach directly incorporates clinical judgment by having the criteria for treatment relation determined by the investigator(s). The method is developed for outcomes that arise from binomial distributions (relatively small trials) and for outcomes that arise from Poisson distributions (relatively large trials). The calculations are illustrated with trial outcomes.

Wam1C1T3

Exact Calculations of Expected Power for the Benjamini-Hochberg Procedure

Deborah Glueck, Anis Karimpour-Fard, Lawrence Hunter, Jan Mandel, Keith E. Muller

University of Colorado and Health Sciences Center, Denver, U.S.A.

We give exact analytic expressions for the expected power of the Benjamini and Hochberg procedure. We derive bounds for multiple dimensional rejection regions. We make assumptions about the number of hypotheses being tested, which null hypotheses are true, which are false, and the distributions of the test statistics under each null and alternative. This enables us to find the joint cumulative distribution function of the order statistics of the p-values, both under the null, and under the alternative. We thus have order statistics that arise from two sets of real-valued independent, but not necessarily identically distributed random variables. We show that the probability of each rejection region can be expressed as the probability that arbitrary subsets of order statistics fall in disjoint, ordered intervals, and that of the smallest statistics, a certain number come from one set. Finally, we express the joint probability distribution of the number of rejections and the number of false rejections by summing the appropriate probabilities over the rejection regions. The expected power is a simple function of this probability distribution. We give an example power analysis for a multiple comparisons problem in mammography.

Wam1C1T4

Estimating the Interesting Part of a Dose-Effect Curve: When is a Bayesian Adaptive Design Useful?

Frank Miller

AstraZeneca, Södertälje, Sweden

We consider the design for dose-finding trials in phase IIB of drug development. We propose that “estimating the interesting part of the dose-effect curve” is an important objective of such trials. This objective will be made more concrete and formulated in statistical terms in the talk. Having defined the objective, we can apply optimal design theory to derive efficient designs. Due to our objective, we use a customized optimality criterion and not a common optimality criterion like D-optimality. We specify both an optimal fixed design (without adaptation) and a two-stage Bayesian adaptive design. The efficiencies of these two designs are compared for several situations. We describe typical situations where you can gain efficiency from using an adaptive design but also situations where it might be better with a fixed design. Briefly, we discuss modifications of the considered adaptive design and potential advantages of these.

Wam1C2T1

Sample Size Re-Estimation and Hypotheses Tests for Trials with Multiple Treatment Arms

Jixian Wang, Franz Koenig

Novartis Pharma AG, Switzerland

Sample size re-estimation (SSRE) provides a useful tool to change a design during the conduct of a study when an interim look reveals that the original sample size is inadequate. For trials comparing an active treatment with a control, a common way to control the type I error is to construct an asymptotically normal distributed weighted test statistic combining the information before and after the interim look.

We consider sample size re-estimation methods for comparing multiple active treatments with a control, where we allow the change of sample size for one arm to depend on the interim information across all arms. We propose several ways to construct weighted statistics combining the information before and after SSRE as well as related test procedures to control the overall type I error. When the change of sample size is proportional across all treatment arms, it is possible to construct statistics so that the Dunnett test can be used as if there was no SSRE. For arbitrary SSREs, we propose other procedures including a closed test based on weighted statistics with marginally standard normal distribution and a test using a multivariate generalization of weighted test statistics in combination with the closure principle. A practical example is used to illustrate the proposed approaches. The properties of the procedures are evaluated by simulations.

Wam1C2T2

Adaptive Design in Dose Ranging Studies Based on Both Efficacy and Safety Responses

Olga Marchenko, Prof. R. Keener, Ann Arbor

i3 Statprobe, Inc, U.S.A.

Traditionally, most designs for Phase I studies gather safety information, aiming to determine the maximum tolerated dose (MTD). Then Phase II designs would evaluate the efficacy of doses in the (assumed) toxicity acceptable. It is highly desirable for many reasons to base the dose selection on efficacy and safety responses simultaneously. Recently, several different designs for dose selection have been proposed that are based on both efficacy and safety (e.g., Thall and Cook (2004), Fedorov and Dragalin (2006), Zhang et al. (2006), etc.). While a majority of designs provide appropriate, safe and efficacious dose or doses with some precision, few of them gain the sufficient information on all doses in the range studied. In this talk, I will show how a flexible, adaptive, model-based design proposed by V.Fedorov and V.Dragalin can be implemented and changed as appropriate by studying simulations similar to three case studies with different desirable responses from several therapeutic areas.

Wam1C2T3

Adaptive Seamless Designs for Subpopulation Selection Based on Time to Event Endpoints

Emmanuel Zuber, Werner Brannath, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, Amy Racine

Novartis Pharma AG, Basel, Switzerland

A targeted therapy might primarily benefit a sub-population of patients. Thus, the ability to select a sensitive patient population may be crucial for the development of such a therapy. Traditionally, one would need to start with a hypothesis generating phase II study to identify a sub-population. The specific sensitivity of that sub-population would have to be confirmed independently in a second phase II study, before a phase III study could be run in the selected target population. A formal claim of efficacy would be based on the phase III data only.

A more efficient approach is presented using an adaptive phase II/III seamless design, to combine into a single two-stage study the selection of either the full or the sub-population, with the proof of efficacy.

From a separate concomitant exploratory study, a sub-population is to be identified independently before the end of stage 1 of the combined phase II/III study. At the end of stage 1, Bayesian tools are used to confirm the hypothesis of a more sensitive sub-population. One may then decide at this step to adapt the conduct of the trial by limiting to that sub-population the further recruitment into stage 2, and by choosing the hypothesis testing strategy. Thus, the independent confirmation of the sub-population is more reliable, being made on the same clinical endpoint and in the same setting as the final phase III demonstration of efficacy. The latter is efficiently based on the combined data from stage 1 and 2, in the selected population, with an adapted testing strategy.

The use of the adaptive design methodology with a time to event endpoint relies on the asymptotic independent increment property of the logrank test statistics. The overall type I error rate is controlled thanks to the concomitant use of adaptive design methodology and of

the closed testing principle for the testing in the different populations. The use of Bayesian decision tools such as predictive powers and a posterior distribution of treatment effect does not affect the overall type I error rate. It allows to account in a statistical manner, for the uncertainty of interim data and external information into the adaptation decision making.
Simulations are necessary for the design of such a complex study, to determine sample size and to assess its operating characteristics as a function of the Bayesian decision rules, and of the unknown prevalence of the sub-population. Properties of treatment effect estimates and the preservation of trial integrity after its adaptation are also studied by simulations, compared to more conventional group sequential designs.

Wam1C2T4

Ranks of True Positives in Large Scale Genetics Experiments

Russell D. Wolfinger (Invited Speaker), Dmitri Zaykin, Lev Zhivotovsky, Wendy Czika, Susan Shao

SAS, U.S.A.

In the context of a large collection of statistical genetics tests in which the number of true positives (TPs) is small, we study the distribution of the ranks of TPs among the false positives (FPs). We investigate the relative efficiency of ranking measures and how many "best" results need to be screened to cover TPs with high probability, using a few different ways of assessing significance and adjusting for multiple testing. This way of looking at the problem can aid in optimally following up on initial significant findings and in planning of future large scale experiments.

A whole-genome association scan is a prominent example, where the number of tests, L, is now commonly in the hundreds of thousands. With modern high-throughput genotyping capabilities, L can be large simply from the number of measured genetics markers, which are usually single nucleotide polymorphisms (SNPs). L can then grow exponentially by considering tests of haplotypes constructed from all possible pairs of SNPs, all triplets, etc. We simulate markers that are in linkage disequilibrium, that is, have some correlation structure, typically blocked. The measure of association of the genetic markers with a binary or quantitative trait of interest is usually some kind of p-value, perhaps weighted towards effect size. Multiple testing methods investigated include Sidak and no adjustment at all.

Wam2C1T1

Multi-Stage Designs Controlling the False Discovery or the Family Wise Error Rate

Sonja Zehetmayer, Peter Bauer, Martin Posch

Section of Medical Statistics, Medical University of Vienna, Austria

When a large number of hypotheses are investigated, conventional single-stage designs may lack power due to low sample sizes for the individual hypotheses. We propose multi-stage designs where in each interim analysis 'promising' hypotheses are screened which are investigated in further stages. Given a fixed overall number of observations, this allows to spend more observations for promising hypotheses than with single-stage designs, where the observations are equally distributed among all considered hypotheses. We propose multi-stage procedures controlling either the Family Wise Error Rate (FWE) or the False Discovery Rate (FDR) and derive optimal stopping boundaries and sample size allocations (across stages) to maximize the power of the procedure.

Optimized two-stage designs lead to a considerable increase in power compared to the classical single-stage design. We show that going from two to three stages additionally leads to a distinctive increase in power. Adding a fourth stage leads to a further improvement, which is, however, less pronounced. Surprisingly, we found only small differences in power between optimized integrated designs, where the data of all stages is used in the final test statistics, and optimized pilot designs where only the data from the final stage is used for testing. However, the integrated design controlling the FDR appeared to be more robust against misspecifications in the planning phase. Additionally, we found that with increasing number of stages the drop in power when controlling the FWE instead of the more liberal FDR becomes negligible.

Our investigations show that the crucial point is not the choice of the error rate or the type of design (integrated or pilot), but the sequential nature of the trial where non-promising hypotheses are dropped in early phases of the experiment so that test decisions among the selected hypotheses can be based on considerably larger sample sizes compared to the classical single-stage design.

Wam2C1T2

Two-Stage Designs for Proteomic and Gene Expression Studies Applying Methods Differing in Costs

Alexandra Goll, Bauer Peter

Section of Medical Statistics, Medical University of Vienna, Austria

In gene expression and proteomic studies we generally deal with large numbers of hypotheses, where only for a small fraction of the hypotheses noticeable effects exist. Due to limited resources, the number of observations per hypotheses in a conventional single-stage design is low which limits the power. It has been shown that two-stage pilot and integrated designs are a good option to improve the power. In these sequential designs, the first stage is used to screen for the promising hypotheses, which are further investigated in the second stage. In the following we more thoroughly investigate this type of two-stage designs where the costs per measurement and effect sizes differ between the first and second stage. To compare different designs we assume that the total costs of the experiment are fixed. Both integrated and pilot designs are based on procedures either controlling the Family Wise Type I Error Rate (FWE) or the False Discovery Rate (FDR). Two scenarios are considered: In the first scenario the experimenter from the beginning may have the choice between two methods that differ in costs and effect sizes (a low-cost standard method or a high-cost improved method). In the second scenario different costs per measurement may arise if the same method is applied at both stages but specific experimental devices have to be produced at higher costs per measurement for the selected markers at the second stage. For the first scenario we show that depending on the cost and the effect size ratios between the methods it is preferable either to apply the low-cost or the high-cost method at both stages. For the second scenario we will show for which cost ratios between stages it is worthwhile to use (optimal) two-stage designs as compared to the single stage design. Finally we also look how design misspecifications in the planning phase would change the power of two-stage designs as compared to the single-stage design.

Wam2C1T3

Some Insights into FDR and k-FWER in Terms of Average Power and Overall Rejection Rate

Meng Du

Department of Statistics, University of Toronto, Canada

This paper provides some insights into the false discovery rate (FDR) and the *k*-familywise error rate (*k*-FWER), through comparing, in terms of the average power, an FDR controlling procedure by Benjamini and Hochberg (1995) and a *k*-FWER controlling procedure by Lehmann and Romano (2005). A further look at the overall rejection rate, the probability of obtaining at least one single discovery, explains the behavior patterns of the average powers of these two procedures that control different types of error rates.
Keywords: average power, false discovery rate, *k*-familywise error rate, large-scale multiple testing, overall rejection rate.

Wam2C1T4

A Weighted Hochberg Procedure

Ajit Tamhane, Lingyun Liu

Northwestern University, U.S.A.

It is often of interest to differentially weight the hypotheses in terms of their importance. Let $H_1,..,H_n$ be $n \geq 2$ null hypotheses with prespecified positive weights $w_1,..,w_n$ which add up to 1, and with p-values, $p_1,..,p_n$ respectively. It is desired to test them, taking into account their weights, while controlling the type I familywise error rate (FWER) at a designated level α. The well-known weighted Bonferroni (WBF) test rejects any H_i with $p_i \leq w_i\alpha$. Weighted Holm (WHM) and weighted Simes (WSM) procedures for this problem were proposed by Holm (1979), Hochberg and Liberman (1994) and Benjamini and Hochberg (1997); however, a weighted Hochberg (WHC) procedure is lacking. Benjamini and Hochberg proposed the following step-down WHM procedure: Let $p_{(1)} \leq \ldots \leq p_{(n)}$ be the ordered p-values, and let $H_{(1)},..,H_{(n)}$ and $w_{(1)},..,w_{(n)}$ be the corresponding hypotheses and weights, respectively. Then reject $H_{(i)}$ if $p_{(j)} \leq [w_{(j)}/\sum_{k=j}^{n} w_{(k)}]\alpha$ for $j = 1,..,i$; ; otherwise accept all remaining hypotheses. They also proposed the following WSM test: Reject $H_0 = \bigcap_{i=1}^{n} H_i$ if $p_{(i)} \leq \frac{\sum_{k=1}^{i} w_{(k)}}{\sum_{k=1}^{n} w_{(k)}}\alpha$ for some $i = 1,..,n$. We consider the following WHC procedure that uses the same critical constants as WHM given above, but operates in the step-up manner: Accept $H_{(i)}$ if $p_{(j)} > [w_{(j)}/\sum_{k=j}^{n} w_{(k)}]\alpha$ for $j = n,..,i$; otherwise reject all remaining hypotheses. We show that this procedure is not closed in general in the sense of Marcus, Peritz and Gabriel (1976) under the WSM test for subset intersection hypotheses except when the weights are equal. In the course of this demonstration we fill the gap in the incomplete closure proof given by Hochberg (1988) for the equal weights case. Also, a direct proof based on finding a lower bound on the probability of accepting all true hypotheses (see, e.g., Liu 1996) fails for unequal weights. However, simulation studies indicate that WHC does control FWER in the

limited number of cases that we have studied. We propose a conservative version of WHC using the critical matrix approach of Liu (1996) and compare its conservatism with WHC in the simulation study.

Wam2C2T1

Multiple Testing in Change-Point Problem with Application to Safety Signal Detection

Jie Chen

Merck Research Laboratories, Inc., U.S.A.

Detection of a change point usually requires testing multiple null hypotheses. In this talk we focus on the inference of a change in the ratio of two time-ordered Poisson stochastic processes, by developing multiple testing procedures which offer the control of some error rates. Possible extensions of the procedures to multiple change-points are explored. The procedures are illustrated using a real data example for drug safety signal detection and a simulation study.

Wam2C2T2

Sequentially Rejective Test Procedures for Partially Ordered Sets of Hypotheses

David Edwards, Jesper Madsen

Novo Nordisk A/S, Denmark

A popular method to control multiplicity in confirmatory clinical trials is to use a hierarchical (sequentially rejective) test procedure, based on an apriori ordering of the hypotheses. The talk describes a simple generalization of this approach in which the hypotheses are partially ordered. It is convenient to display the partial ordering as a directed acylic graph (DAG). To obtain strong FWE control, certain intersection hypotheses must be inserted into the DAG. The resulting DAG is called partially closed. The purpose of the approach is to enable the construction of inference strategies for confirmatory clinical trials that more closely reflect the trial objectives.

Wam2C2T3

Simultaneous Confidence Intervals by Iteratively Adjusted Alpha for Relative Effects in the One-Way Layout

Thomas Jaki, Martin J. Wolfsegger

Lancaster University, United Kingdom

A bootstrap based method to construct 1−alpha simultaneous confidence intervals for relative effects in the one-way layout is presented. This procedure takes the stochastic correlation between the test statistics into account and results in narrower simultaneous confidence intervals than the application of the Bonferroni correction. Instead of using the bootstrap distribution of a maximum statistic, the coverage of the confidence intervals for the individual comparisons are adjusted iteratively until the overall confidence level is reached. Empirical coverage and power estimates of the introduced procedure for many-to-one comparisons are presented and compared with asymptotic procedures based on the multivariate normal distribution.

Wam2C2T4

Stepwise Testing of Multiple Dose Groups Against a Control With Ordered Endpoints

James Francis Troendle (Invited Speaker)

NIH, U.S.A.

Hierarchical gatekeeper methods exist for testing multiple dose clinical trials with multiple endpoints. This paper considers the case of ordered endpoints where an endpoint will only be tested at a given dose if all higher endpoints were found significant at that dose. Existing stepwise procedures based on the Bonferroni procedure are compared to new methods that incorporate correlation either through an assumption of Gaussian distribution or through resampling. The methods are compared by simulation for power and control of the familywise error.

Wpm1C1T1

A New Method to Identify Significant Endpoints in a Closed Test Setting

Carlos Vallarino, Joe Romano, Michael Wolf, Dick Bittman

Takeda Pharmaceuticals NA, U.S.A.

We present a new multiple testing procedure that has a maximin property under the normal assumption. The new method alters the rejection region of the simple sum test to make it consonant, i.e. to guarantee that rejection of the intersection hypothesis, in a closed test setting, implies the significance of at least one endpoint. Consonance is a desirable property which increases the ability to reject false individual null hypotheses. Designed to perform well when testing related endpoints, the new procedure is applied to PROactive, a cardiovascular (CV)-outcome trial of patients with type 2 diabetes and CV-disease history. Had the PROactive trial considered its two main endpoints as co-primary, the new method shows how efficacy for one key endpoint could have been established.

Wpm1C1T2

Proportion of True Null Hypotheses in Non High-Dimensional Multiple Testing Problems: Procedures and Comparison

Mario Walther, Claudia Hemmelmann; Rüdiger Vollandt

Institute of Medical Statistics, Computer Science and Documentation, Friedrisch-Schiller-Universität Jena, Germany

When testing multiple hypotheses simultaneously, a quantity of interest is the proportion of true null hypotheses. Knowledge about this proportion can improve the power of different multiple test procedures, which control the generalized family-wise error rate, the false discovery rate or the false discovery proportion. For instance in stepwise procedures the critical values, with which the p-values have to be compared, can be increased, if an upper bound of the proportion of true null hypotheses is known.
There are a lot of authors who concerned with establishing methods of estimating the proportion of true null hypotheses. Most of the introduced procedures are based on several thousands p-values,
which are often assumed to be independent. These procedures work very well, however, problems arise when the dimension of the multiple testing problem is only in the few hundreds and the data are correlated. There the latter one is for example the case in EEG, proteomic or fMRI data. Within this framework we pose the question, what is a “good” estimation of the proportion of true null hypotheses. We therefore introduce several criteria to evaluate the efficiency of the estimations. One criterion will be the probability that a certain estimation method overestimates the proportion of true null hypotheses. Another criterion will be whether the confidence interval of the proportion of true null hypotheses is contained in a range of a pre-specified accuracy.
In this talk, we will explain methods for estimating the proportion of true null hypotheses, which are also suitable for non high-dimensional multiple testing problems with correlated p-values. Furthermore we will evaluate and compare the quality of the estimators regarding the introduced criteria in a simulation study.

Wpm1C1T3

An Exact Test for Umbrella Ordered Alternatives of Location Parameters: the Exponential Distribution Case

Parminder Singh

Guru Nanak Dev University, Amritsar, India

A new procedure for testing the null hypothesis against umbrella ordered alternative with at least one strict inequality, where is the location parameter of the ith two-parameter exponential distribution, is proposed. Exact critical constants are computed using recursive integration algorithm. Tables containing these critical constants are provided to facilitate the implementation of the proposed test procedure. Simultaneous confidence intervals for certain contrasts of the location parameters are derived by inverting the proposed test statistic. In comparison to existing tests, it is shown, by a simulation study, that the new test statistic is more powerful in detecting umbrella type alternatives when the samples are derived from exponential distributions. As an extension, the use of the critical constants for comparing Pareto distribution parameters is discussed.

Wpm1C1T4

Procedures Controlling Generalized False Discovery Rate

Sanat Sarkar, Wenge Guo

Department of Statistics, Temple University, U.S.A.

Procedures controlling error rates measuring at least *k* false rejections, instead of at least one, can potentially increase the ability of a procedure to detect false null hypotheses in situations where one seeks to control *k* or more false rejections having tolerated a few of them. The *k*-FWER, which is the probability of at least *k* false rejections and generalizes the usual familywise error rate (FWER), is such an error rate that is recently introduced in the literature and procedures controlling it have been proposed. An alternative and less conservative notion of error rate, the *k*-FDR, which is the expected proportion of *k* or more false rejections among all rejections and generalizes the usual notion of false discovery rate (FDR) will be introduced in this talk. Procedures with the *k*-FDR control dominating the Benjamini-Hochberg stepup FDR procedure and its stepdown analog under independence or positive dependence and the Benjamini-Yekutieli stepup FDR procedure under any form of dependence will be presented.

Wpm1C2T1

Effects of Dependence in High-Dimensional Multiple Testing Problems

Kyung In Kim, Mark A. van de Wiel

Eindhoven University of Technolog, The Netherlands

We consider effects of dependence among variables of high-dimensional data in multiple hypothesis testing problems. Recent simulation studies considered only simple correlation structure among variables, which was hardly inspired by real data features. Our aim is to describe dependence as a network and systematically study effects of several network features like sparsity and correlation strength. We discuss a new method for efficient guided simulation of dependent data, which satisfy the imposed network constraints. We use constrained random correlation matrices and perform extensive simulations under nested conditional independence structures. We check the robustness against dependence of several popular FDR procedures such as Benjamini-Hochberg FDR, Storey's q-value, SAM and other resampling based FDR procedures. False Non-discovery Rates and estimates of the number of null hypotheses are computed from those methods and compared. Our simulations studies show that popular methods such as SAM and the q-value seem to overestimate nominal FDR significance level under dependence conditions. On the other hand, the adaptive Benjamini-Hochberg procedure seems to be most robust and remain conservative. Finally, the estimates of the number of true null hypotheses under various dependence conditions are variable.

Wpm1C2T2

A Semi-Parametric Approach for Mixture Models: Application to Local FDR Estimation

Jean-Jacques Daudin, A. Bar-Hen, L. Pierre, S. Robin

INRA AgroParisTech, France

In the context of multiple testing, the estimation of false discovery rate (FDR) or local FDR can be stated in the mixture model context. We propose a procedure to estimate a two-components mixture model where one component is known. The unknown part is estimated with a weighted kernel function, which weights are defined in an adaptative way. We prove the convergence and unicity of our estimation procedure. We use this procedure to estimate the posterior population probabilities and the local FDR.

Key words: FDR, Mixture model, Multiple testing procedure, Semi-parametric density estimation.

Wpm1C2T3

Two New Adaptive Multiple Testing Procedures.

Etienne Roquain, Gilles Blanchard

MIG- INRA Jouy-en-Josas, France

The proportion π_0 of true null hypotheses is a quantity that often appears explicitly in the FDR control bounds. Recent research effort has focussed on finding ways to estimate this quantity and incorporate it in a meaningful way in a multiple testing procedure, leading to so-called “adaptive” procedures.
We present here two new adaptive step-up multiple testing procedures:
- The first procedure that we present is a one-stage step-up procedure. We prove that it has a correct (and strong) FDR control given that the test statistics are independent. If there the set of rejection is not too large (typically less than 50%), this procedure is less conservative than the so-called “two-stage linear step-up procedure” of Benjamini, Krieger and Yekutieli (2006). Moreover, preliminary simulations show that this new procedure seems to still have a correct FDR control when the test statistics are positively correlated.
- The second procedure that we present is a two-stage step-up procedure. We prove that it has a correct (and strong) FDR control in the “distribution free” context. Because the techniques used in the distribution free context are inevitably less precise, this new adaptive procedure is more conservative than thoses built under independence. However, it will be relevant if we expect a “large” proportion of rejected hypotheses (typically more than 50%).

Wpm1C2T4

Multi-Stage Gatekeeping Procedures with Clinical Trial Applications

Alex Dmitrienko (Invited Speaker), Ajit Tamhane

Eli Lilly and Company, U.S.A.

This talk introduces a general approach to constructing gatekeeping procedures for multiple testing problems arising in clinical trials with hierarchically ordered objectives (primary/secondary endpoints, dose-control comparisons, etc). The approach is applied to set up gatekeeping procedures based on popular multiple tests (Holm, fallback and Hochberg tests), resampling and parametric tests. The resulting procedures have a straightforward multi-stage structure that facilitates the implementation of gatekeeping procedures and communication of the results to non-statisticians. One can also account for logical restrictions among multiple analyses and improve the power of individual tests by eliminating comparisons that are no longer clinically meaningful. The general approach is illustrated using clinical trial examples.

Wpm2C1T1

A Unifying Approach to Non-Inferiority, Equivalence and Superiority Tests

Chihiro Hirotsu

Meisei University, Japan

Two approaches of multiple decision processes are proposed for unifying the non-inferiority, equivalence and superiority tests in a comparative clinical trial for a new drug against an active control. One is a method of confidence set with confidence coefficient 0.95 improving the consumer's and producer's risks of the usual approach of the naïve confidence interval. It requires to include 0 within the region as well as to clear the non-inferiority margin so that a trial with somewhat large number of subjects for proving non-inferiority of a drug which is actually inferior should be unsuccessful.
The other is the closed testing procedure combining the one- and two-sided tests by applying the partitioning principle and justifies the switching procedure unifying the non-inferiority, equivalence and superiority tests. In particular regarding the non-inferiority the proposed method justifies simultaneously the old Japanese Statistical Guideline (one-sided 0.05 test) and the International Guideline (two-sided 0.05 test). The method is particularly attractive changing the strength of the evidence of relative efficacy of the test drug against a control at five levels according to the achievement of the clinical trial.
Key words: Bio-equivalence, closed testing procedure, confidence set, non-inferiority, partitioning principle, superiority.

Wpm2C1T2

Multiplicity-Corrected, Nonparametric Tolerance Regions for Cardiac ECG Features

Gheorghe Luta, S. Stanley Young, Alex Dmitrienko

National Institute of Statistical Sciences, U.S.A.

Electrocardiograms are used to evaluate possible effects on the heart induced by drug candidates. These waveforms are quite complex and many numerical features of these waveforms are extracted for statistical evaluation. In addition, various covariates, heart rate, gender, age, etc., also need to be taken into account. There is a need to consider the multiple questions under consideration. Our idea is to combine two statistical methodologies, nonparametric tolerance regions and resampling-based multiple testing correction. We will review electrocardiograms and their standard numerical characteristics, and place this work into the framework of drug evaluation clinical trials. Using real data, we will show how nonparametric tolerance regions can be used with resampling multiplicity adjustments. The product of this strategy will be tolerance regions that adapt to the shape of the observed distributions and control over the family-wise error rate over the clinical trial.

Wpm2C1T3

Comparing Treatment Combinations with the Corresponding Monotherapies in Clinical Trials

Ekkehard Glimm, Norbert Benda

Novartis Pharma AG, Switzerland

The intention of many clinical trials is to show superiority of a treatment over two others. E.g. a combination therapy may be compared to the corresponding monotherapies. In such a trial two drugs are administered simultaneously. A beneficial effect might arise from a synergistic effect of the monotherapies. Even in presence of an antagonistic effect, however, a simple superiority of the combination drug might be sufficient, e.g. as a way to overcome dose limitations of the monotherapies.

The standard confirmatory statistical test consists of two tests at level α and rejection if both of them are significant. This approach was called min test by Laska and Meissner (1989) who showed that it is uniformly most powerful in a certain class of monotone tests. However, while it exhausts the α-level if the difference between monotherapy effects approaches infinity, it is very conservative in the practically more relevant situation of similar monotherapy effects. Sarkar et al. (1995) have shown that it is possible to construct tests that are uniformly more powerful than this approach, if the notion of monotonicity is abandoned.

In this talk, we will present alternatives to the tests suggested by Sarkar et al., some of which are also uniformly more powerful than the min test, and others which simply have a different power profile (e.g. are advantageous for small or large effect differences).

Simulations and asymptotic considerations will be used to investigate where and how much power is gained depending on the constellation of the therapeutic effects. Finally, the concept of monotonicity and its practical implications will be discussed.

Wpm2C1T4

Poster Session 1

Simultaneous Confidence Intervals for Overdispersed Count Data

Daniel Gerhard, Frank Schaarschmidt, Ludwig A. Hothorn

Institute of Biostatistics, Leibniz University of Hannover, Germany

The application of simultaneous confidence intervals for count data can be beneficial for various research objectives, like observing tumor counts in clinical trials and non-clinical studies or for the comparison of insect abundance in agricultural field trials. The confidence intervals considered here are constructed with parameter estimates from a generalized linear model, assuming the counts to be Poisson or in case of overdispersion negative-binomial distributed. Multiplicity is taken into account by a corresponding quantile of the multivariate t-distribution with certain correlation. In a simulation study we investigated the coverage probability of confidence intervals for different distributional assumptions in various factorial designs for several sample sizes. It is shown, that nominal level alpha is reached only at a number of observations larger than 20 and sufficient large sample means. Simulation studies and evaluation of examples were performed in the free software environment R using the packages gamlss and multcomp.

Bretz, F., Genz, A., Hothorn, L.A. (2001): On the numerical availability of multiple comparison procedures. Biometrical Journal 43: 645-656.

McCulloch, C.E. and Searle, S.R. (2001): Generalized, linear and mixed models. John Wiley & Sons, Inc.

Rigby, R.A. and Stasinopoulos D.M. (2004): Generalized additive models for location, scale and shape. Applied Statistics, 54: 1-38.

Approximative Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions and Poly-3-Adjusted Tumour Rates

Frank Schaarschmidt, Martin Sill, Ludwig A. Hothorn

Institute of Biostatistics, Leibniz University of Hannover, Germany

Simultaneous confidence intervals for contrasts of means in a one-way layout with k independent samples are well established for Gaussian distributed data. Procedures approaching different practical questions are available, as all-pairs or many-to-one comparisons, comparison with average, or different tests for order-restricted alternatives. However, if the distribution of the response is not Gaussian, corresponding methods are usually not available or not implemented. For the two cases: i) *k* binomial proportions (Price and Bonett, 2004), and ii) *k* Poly-3-adjusted tumour rates (Bailer and Portier, 1988) we extended recently proposed confidence interval methods for the difference of two proportions or single contrasts to multiple contrasts by using quantiles of the multivariate normal distribution. The small sample performance of the proposed methods was investigated in simulation studies. For binomial proportions and poly-3-adjusted tumour rates, the simple adjustment of adding 2 pseudo-observations to each sample estimate leads to reasonable coverage probabilities. The methods are illustrated by evaluation of real data examples of a clinical trial and a long-term carcinogenicity study. The proposed methods and examples are available in the R package MCPAN.

Bailer, J.A. and Portier, C.J. (1988): Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics 44, 417-431.
Price, R.M. and Bonett, D.G. (2004): An improved confidence interval for a linear function of binomial proportions: Computational Statistics & Data Analysis 45: 449-456.

Lets ROC on Microarrays

Carina Silva-Fortes, Maria Antónia Amaral Turkman, Lisete Sousa

Escola Superior de Tecnologia da Saúde de Lisboa-Insituto Politécnico de Lisboa, Portugal

There are new statistical challenges posed by data from microarray experiments, due to the exploratory nature of experiments and the huge number of genes under investigation. There are many statistical techniques to analyze those data, but sometimes they are too difficult to implement. We present the advantages of the application of receiver operating characteristic (ROC) analysis in microarray data, in particular on selection of genes that are differentially expressed in different known classes of tissue. We also present one example of application of ROC analysis to select the optimal cut-off value for gene classification.
Key-words: ROC curves, microarrays, optimal cut-off, differential expression

Performance of Multiple Testing Procedures for Genomic Differences in Groups of Papillary Thyroid Carcinoma Analysed by Array CGH

Herbert Braselmann, Eva Malisch, Kristian Unger, Gerry Thomas, Horst Zitzelsberger

GSF National Research Center for Environment and Health, Institute of Molecular GSF - National Research Center, Institute for Molecular Radiation Biology, Germany

Microarray-based Comparative Genomic Hybridization (array CGH) allows to detect DNA copy number differences between a reference

genome and a tumour genome at thousands of chromosomal sites simultanously. Papillary thyroid carcinomas (PTC) are often carrying RET/PTC rearrangements which have been shown to be heterogeneously distributed within tumour tissues. Thus, it is likely that additional gene alterations are present in these tumours. Moreover, RET/PTC negative tumours should exhibit alternative changes. Therefore we have investigated 33 PTC (20 adult tumours, 13 infantile, post-Chernobyl tumours) with known RET/PTC status (RET/PTC positive: 11 adult, 10 infantile cases, RET/PTC negative: 9 adult, 3 infantile cases) by array CGH to uncover such unknown gene alterations in PTC. Endpoints are given as log2-transformed intensity ratios (log2-ratios), which are further simplified to gain or loss status variables. For the analysis of group differences for approximately 1000 preselected genomic sites between adult versus infantile or between RET/PTC positive versus RET/PTC negative tumours, results of a multiple t-test for smoothed log2-ratios and of multiple Fisher's exact tests for gains or losses are presented. When testing for age dependence, Benjamini-Hochberg's FDR procedure resulted in about 50-100 significant differences, similar as a maximum permutation procedure (maxT) for the t-tests. For comparison of the RET/PTC status groups, FDR procedures resulted in hundreds of significant differences, whilst the maxT procedure yielded 46 significances. Fisher's exact test for gains or losses yielded throughout a smaller number of significant differences. Typically for array CGH, a large part of the intensity ratios are positively correlated among the samples within chromosomal segments of variable length.
The results demonstrate exemplarily the performance of false discovery rate (FDR) and familywise error rate (FWER) p-value adjustments for a type of high-dimensional data. Results are also dependent on the data preprocessing methods and the chosen endpoint.

Multiplicity Adjusted Location Quotients

Gemechis Dilba, Frank Schaarschmidt, Bichaka Fayissa

Institute of Biostatistics, Leibniz University of Hannover, Germany

Location quotient is an index frequently used in geography and economics to measure the relative concentration of activities. For binomial data, the problem consists of simultaneously comparing the ratios of the individual proportions to the overall proportion. Apparently, this is a multiple comparison problem and up to now multiplicity adjusted location quotients have not been addressed. In fact there is a negative correlation between the comparisons when proportions of the subgroups are compared with the proportion of all the subgroups combined. Here, we propose adjusted location quotients based on existing probability inequalities and by directly using the asymptotic joint distribution of the associated z-statistics. A simulation study is carried out to investigate the performance of the various methods in terms of achieving a nominal simultaneous coverage probability. A simple adjustment of Fieller confidence intervals is observed to work quite well. The proposed methods will be illustrated on a health utilization data.

Forecasting Monthly Temperature and Relative Humidity Using Time Series Analysis

Inderjeet Kaushik, PR Maiti

Institute of Technology, Banaras Hindu University, India

Prediction of climatic factors like temperature and relative humidity is a stochastic process. In this paper an effort is made to model these parameters by using time series analysis for forecasting monthly temperature and relative humidity. For the analysis and forecasting purpose last 12 years monthly data of Mirzapur district is used. Time series ARIMA models provide quite satifactory results than other time series models.

Application of Multiple Comparison Procedures for Analysis of Naltrexone and Fluoxetine Effects for Treatment of Heroin Dependence

Elena V. Verbitskaya, Evgeny M. Krupitsky, Edwin E. Zvartau, Marina V. Tsoi-Podosenin, MD, Valentina Y

Laboratory of Biomedical Statistics, St.-Petersburg Pavlov State Medical University, Russia

A previous study of 52 patients randomized to naltrexone or naltrexone placebo demonstrated that naltrexone was clinically effective for preventing relapse to heroin addiction in Russia. This study was done to replicate these early results in a larger sample, and see if the combination of an SSRI antidepressant with naltrexone might alleviate the depression, anxiety and anhedonia typically associated with opioid detoxification and improve the results of

naltrexone treatment. 280 heroin addicts who completed detoxification at addiction treatment units in St. Petersburg and provided informed consent were randomized to a 6 month course of biweekly drug counseling and one of four groups of 70 subjects/group: Naltrexone 50 mg/day (N) + Fluoxetine 20 mg/day (F); N + Fluoxetine placebo (FP); Naltrexone placebo (NP) + F; or NP + FP. Medications were administered under double-dummy/double-blind conditions. Primary endpoint was relapse rate and the main analysis for it was survival analysis. There were several secondary endpoints that assessed changes of such psychometric characteristics as craving for heroin (VASC), Global Assessment of Functioning (GAF; DSM-IV, 1994), Beck Depression Inventory (BDI; Beck et al, 1961), Brief Psychiatric Rating Scale (BPRS; Overall, Gorham, 1962), Spielberg State-Trait Anxiety Test (SSTAT) (Spielberg et al, 1976), Scale of Anhedonia Syndrome (SAS; Krupitsky et al, 1998). We met several problems in regard to statistical analysis of such data: 1) multiple secondary endpoints,. 2) multiple timpoints, all secondary endpoints were tested 3-13 times during trail; 3) the big change of number of patients relapsed or lost for follow up that causes a big disbalance at the end of the trail that limited the usage of repeated measures MANOVA. At the end of six months, 43% of subjects in the N+F group remained in the study and had not relapsed as compared to 36% in the N+FP group, 21% in the NP+F group, and 10% in the NP+FP group. Combination of two samples (pilot and main) increases the sample at the end of the 6 month period. There was prominent effect of drugs on characteristics of addiction (MANOVA, Games-Howel Post hock test), Repeated measures ANOVA showed that were no effect of treatments on psychometric characteristics, only effect of time: it includes only those patients, who stay in th program til the end of the study. But MANCOVA tests with on the 3 month data and 6 month data demonstrated effect of Naltrexone.

Optimal Allocation of Sample Size in Two-Stage Association Studies

Shu-Hui Wen, CK Hsiao

Department of Public Health, Tzu-Chi University, Taiwan

Multiple testing occurs commonly in genome-wide association studies with dense SNPs map. With numerous SNPs, not only the genotyping cost and time increase dramatically, most traditional family-wise error rate (FWER) controlling methods may fail for being too conservative and lose power when detecting SNPs associated with disease. Lately, more powerful two-stage strategies for multiple testing have received great attention. In this paper, we propose a grid-search algorithm for an optimal design for sample size allocation under these two-stage procedures. Two types of constraints are considered, one is on the overall cost and the other on sample size. With the proposed optimal allocation of sample size, bearable false positive results and larger power can be achieved to meet the limitation on study design. As a general rule, the simulations indicate that allocating at least 80% of the total cost in stage one provides maximum power, as opposed to other methods. If per-genotyping cost in stage two differs from that in stage one, downward proportion of the total cost in earlier stage maintains good power. For limited total sample size, evaluating all the markers on 55% of the subjects in the first stage provides maximum power while the cost reduction is approximately 43%.

Nonparametric Tolerance Bounds for Gene Selection

S. Stanley Young, Gheorghe Luta

National Institute of Statistical Sciences, USA NISS, U.S.A.

A tolerance bound “covers” a specific proportion, P, of the distribution with a fixed level of confidence. The usual interest is to cover the central part of the distribution. For certain problems, e.g. selection of genes from a microarray experiment for further characterization, there is a need to select a set of genes expected to contain the most extreme P proportion of the genes tested. So rather than statistically testing each gene and selecting the gene if some multiple testing threshold has been obtained, our idea is to select a set of genes that contain, with specified confidence, the most extreme genes in the set of genes tested.

Adjusting for Multiple Testing

Mohamed Moussa, Nil

Faculty of Medicine, Kuwait University, Kuwait

Multiple hypotheses testing is a common problem in medical research. Multiple hypotheses testing theory provides a framework for defining and controlling appropriate error rates in order to protect against wrong conclusions. A one-way analysis of variance (ANOVA) is used when the effect of an explanatory factor with more than two groups on continuous outcome variable is explored. If the ANOVA statistics show significant difference in means between factor groups, then multiple pairwise comparisons are performed to find which groups are significantly different from another. This is done either by specific group differences using planned (apriori) comparisons which are decided before the ANOVA is run, or using post-hoc (aposteriori) tests which involve all possible comparisons between groups. Post-hoc tests are data-driven and hence are inferior to the thoughtful planned tests. Type 1 error increases with the number of comparisons, hence some adjustments are made to preserve it. If *n* comparisons are made, the probability that at least one of them will be significant is 1- (1-α)*n, if all *n* individual null hypotheses (H0) are true, α is the probability of falsely rejecting H0. It is preferable to run a small number of planned comparisons rather than a large number of unplanned post-hoc tests. Post-hoc tests vary from being conservative to liberal with no adjustment for multiple comparisons. A conservative test is one in which the actual significance is smaller than the stated critical significance level. Thus conservative tests may incorrectly fail to reject H0. The choice of post-hoc test is mainly determined by equality of the variance. Equal variance post-hoc tests are either conservative tests (Scheffe, Tukey's honestly significant difference, HSD, Bonferroni, and Sidak) or liberal tests (Fisher's least significant difference, LSD, Duncan's new multiple range test, Student – Newman – Keuls, SNK). Equal variance not assumed post-hoc tests include Games Howell and Dunnett's C tests.

The aim of this paper was to apply the existing multiple comparison procedures in exploring the effect of physical activity level (Very light, Light, Moderate, and Heavy) on the continuous cardiovascular risk marker 'total sialic acid'.

Our results showed that the LSD test is the most liberal post-hoc test showing three significant comparisons (Very light/Light, p=0.008; Very light/Moderate, p=0.005; Very light/Heavy, p=0.012). The Tukey's HSD, Bonferroni, and Sidak showed the same significant comparisons (Very light/Light; Very light/Moderate) with p-values 0.039, 0.023 in Tukey's HSD, p= 0.047, 0.027 in Bonferroni, and p=0.046, 0.027 in Sidak tests respectively. The Scheffe's test was the most conservative showing only one significant comparison (Very light/Moderate, p=0.044). Epidemiologists show less enthusiasm about formal adjustment procedures since they increase type 2 error and hence decrease statistical power to detect significance. There is an extreme view that denies the need for adjustments for multiple comparisons. They argue that multiple comparisons are appropriate if the universal H0 and omnibus HA are of interest, but in most studies they are of no interest. Studies with a single key interest apriori planned may often generate stronger evidence on a specific hypothesis than studies with aposteriori multiple interests. It is emphasized that adjustments for multiple testing are required in confirmatory studies whenever results from multiple tests have to be combined in one final conclusion and decision. It is suggested that multiple-comparison procedures are frequently adopted unnecessarily. Provided that a selected number of well-defined individual null hypotheses are specified apriori at the design, there are situations in which multiple tests of significance can be performed without adjustment of type 1 error rate.

Biotechnology as Chance for Food Safety

Kakha Nadiradze

Biotechnology Center of Georgia, Georgia

At the beginning of the 21thcentury, the Modern Bio and Microbiology Techniques plays a very Important role in the world agriculture, environment and ecology and scientific and research activities. Due to new approaches of the researches realities, the Bio and Microbiology can be considered as a one of the important field of the agriculture with a number of problems that have to be resolved in the interest of all countries jointly. Since humans began to live in settled agricultural communities they have been involved in a constant battle to reduce the impact of pests-insects, mites, mollusk, pathogens, weeds, mammals and bird-on their crops. They have to control these problems through manual methods, intercropping, tillage and composting, as well as more innovative methods like the use of predatory vertebrates. Mankind has always exploited the potential of beneficial organisms to control pests, in what we now call biological control. At its simplest, biological control or bio-control is the deliberate use of one or more organisms to control another organism that has became a pest. Within bio-control three different approaches:

Classical bio-control:

traditionally used for permanent suppression of an alien pest through the introduction and release of co-evolved or highly specific natural enemies from the pests are of origin

Augmentation:

the Release or application of (usually indigenous) natural enemies in large numbers to control pest outbreaks.

Conservation:

The promotion of practices favoring the activity of indigenous natural enemies against either native or non-native pests.

There are four mail categories of biological agents:

Insect parasitoids that are parasitic on other insects in early stage of development but eventually kill their hosts; most are Hymenoptera (wasps) or Diptera (files). Predatory invertebrates and vertebrates that eat prey species Phytophagous or plant-eating invertebrates associates with weeds. Microbial agent including bacteria, viruses,

fungi and nematodes. Some microbial or their byproducts (toxins) are formulates into bio-pesticide preparations, which are used in a similar way to chemical pesticides. The use of biological control is not without risk; many people are aware of the disastrous impacts of the cane food introduces in a non-scientific attempt to control.
Bio-control should underpin most pest management programmers to establish a sustainable balance in the environment:
It replaces reduces the need for chemical control
It readily integrates with little or no negative impact on the ecosystem
It is a long –term means of control
It is more cost –effective

Statistical Method for Finding Protein-Binding Sites from ChIP-Chip Tiling Arrays

Taesung Park, Haseong Kim, Jae K. Lee

Deptartment of Statistics, Seoul National University, South Korea

Recently, high-resolution tiling chromatin-immunoprecipitation chips (ChIP-chip) have being increasingly used to find the protein-binding sites, replication origins of chromosomes, and DNase hypersensitive sites. However, due to the non-ignorable noises and high-resolution of tiling arrays, it is very difficult to obtain a sufficient number of biological replicates of ChIP-chip tiling arrays with a high reproducibility. Further, not many solid statistical methods are currently available to analyze ChIP-chip tiling arrays. We propose a new statistical method to map the transcription factor IID (TFIID) binding sites using the ChIP-chip tiling arrays without any replicate. The proposed method adopts a local error pooling method to control the high noise levels of tiling arrays caused by the correlations between the adjacent probes. Our real data application of 38 NimbleGene ChIP-chip tiling arrays containing a total of 14,535,659 50-mer oligonucleotides, positioned at every 100 basepairs(bp) throughout the human genome, successfully identified the 6,411 active promoters in human cells which are bound by the general transcription factor IID (TFIID).

Maximum Contrast Tests and Model Selection under Order Restriction

Xuefei Mi, L.A. Hothorn

Biostatistics Unit, Leibniz University of Hannover, Germany

The use of order-restricted hypotheses is a common approach to increase power. Hereby, simple-order and tree-order are three common types of order restrictions. In this talk we focused on how to selection a suitable pattern of them. Several approaches are available for these problems, such as max-t statistics according to Hirotsu and Srivastava (2000) which can be formulated as maximum contrast approach belonging to the broader class of multiple contrast tests (MCT). The disadvantage of MCT is that it can only reject the global null hypotheses. The local finding rate of the true pattern of the alternative is low. Recently, Anraku (1999), Zhao et. (2002) and Ninomiya (2005) developed information-criterion based log-likelihood method for model selection approaches under certain types of order restriction. These methods have better finding rate of the true pattern, but do not control the alpha rate. They treat the null model as one of the possible pattern among all others and are not constructed as hypothesis test to reject the alternatives. In this talk we will compare these two methods for simple-order and tree-order. Also we will present a modification which can control the alpha rate under simple-order restriction.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). Order restricted statistical inference. Wiley, New York.

Bretz, F and Hothorn, L.A. (2002). Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. Statistics Medicine 21, 3325

Ninomiya, Y. (2005). Information criterion for Gaussian change-point model. Statistics & Probability Letters 72, (3): 237-247

Zheng, L. and Peng, L. (2002). Model selection under order restriction. Statistics & Probability Letters 57:44, 301-306

Hirotsu, C. and Srivastava, M. S. (2000). Simultaneous confidence intervals based on one-sided max t test. Statistics & Probability Letters 49, 25-37.

Poster Session 2

Bayesian Classification and Label Estimation Via Em Algorithm: a Comparative Study

Marilia Antunes, Lisete Sousa

Faculty of Sciences, University of Lisbon, Portugal

Gene classification problem is studied considering the ratio of gene expression levels, X, in two-channel microarrays and a non-observed categorical variable indicating how differentially expressed the gene is: not differentially expressed, down-regulated or up-regulated. Supposing X from a mixture of Gamma distributions, two methods are proposed and results are compared. The first method is based on a hierarchical Bayesian model. The conditional probability of a gene to belong to each group is calculated and the gene is assigned to the group for which this conditional probability is higher. The second method uses EM algorithm to estimate the most likely group label for each gene, that is, to assign the gene to the group which contains it with the higher estimated probability.

Estimation of Parameters in Unconditional Categorical Regression

Kamal Azam, Grami A., Ph.D et al

Tehran University of Medical Sciences, Iran

In large–scale sampling, we are always facing non-responses item(s) non-response or unit(s) or both. In fitting a model to the data we have two groups of variables, namely dependent and independent variables. Non-response may occur for any of these groups of variables. In this paper we assume that Y as a categorical dependent variable, Z and X as independent variables. The first two variables are fully observed and we assume that the mechanism of missing-ness is random (MAR). In order to estimate parameters a model is devised based on likelihood function for the whole data set including missing data and the estimation of parameters are compared with those obtained by statistical software such as S-Plus which are only based on complete observed data and ignore missing units .

Our results show that the estimations obtained using maximum likelihood based model is superior to the standard estimations for the approach utilized by the soft wares. The comparison is made on a set of health survey data on goiter disease carried out in Qazvin province.

Key words: Missing At Random, Logistic Regression, Goiter Disease, Maximum Likelihood

Adjustment Method to Address Type I Error and Power Issues with Outcome Multiplicity and Correlation

Richard Blakesley, Sati Mazumdar, Patricia Houck

University of Pittsburgh, U.S.A.

Multiple comparisons call into question the validity of individual hypothesis testing due to type I error inflation. Several adjustment methods exist in statistical literature to protect type I error. However, their type I error and power performance suffer with increasing outcome multiplicity and correlation. Single-step approaches (Bonferroni, Sidak) protect type I error for independent outcomes, but become conservative with increasing correlation and suffer from lack of power. Stepwise approaches (Holm, Hochberg, Hommel) demonstrate improved power over single-step methods. Methods which use correlation components in the adjustment formulae (Dubey/Armitage-Parmar and R-Squared Adjustment) address overcorrection of type I error to a limited extent. Resampling methods (Bootstrap MinP and Step-Down MinP) incorporate correlation structure, but there exist caveats and implementation limits. We propose combining a stepwise approach with a new correlation component to stabilize type I error protection and maintain high power to reject false null hypotheses regardless of outcome multiplicity and correlation levels.

Methods:

Simulations were conducted in the R statistical package. Multivariate normal datasets were simulated in each experiment under varying conditions of effect sizes (uniform, split), number of outcomes (4, 8, 12, 24), correlation structure (compound symmetry, block symmetry, decreasing dependence) and strength of outcome correlation. For each simulated dataset, two-sample t-tests were performed for each outcome, adjustment methods were applied, and type I error and three power formulations (minimal, maximal, average) were estimated. The proposed method used the Sidak form, a step-up approach, and a measure of hypothesis independence. Previously mentioned methods were included for comparison.

Results:

The proposed method demonstrated stable type I error protection across the explored correlation structures. It also showed similar or greater power (all formulations) than examined methods with conservative type I error protection. These results held for increased outcome multiplicity.
Conclusion:
The new method holds promise to allow high power to make inferences without concern for type I error issues regarding multiple correlated outcomes.
Funding Source:
NIMH T32 MH073451, NIMH P30 MH071944

Quantile Curve Estimation and Visualization for Non-Stationary Time Series

Dana Draghicescu, Serge Guillas, Wei Biao Wu

Hunter College, City University of New York, U.S.A.

This talk addresses the problem of quantile curve estimation for a wide class of non-stationary and/or non-Gaussian processes. We discuss several smoothed quantile curve estimates, give asymptotic results, and introduce a data-driven procedure for the selection of the optimal smoothing parameter. This methodology provides a statistically accurate and computationally efficient graphical tool, that can be used for the exploration and visualization of the behavior of time-varying quantiles for time series with complex structures. A Monte Carlo simulation study and two applications to ozone time series illustrate the findings.

Scale and Suitable Analysis

Fumihiko Hashimoto

Osaka City University, JAPAN

It seems that they are fully conscious of the rank of scales, such as an “ordinal scale” and an “interval scale”, in the treatise of a medical field. They use non-parametric analysis for low-level scale (e.g. nominal scale), and use parametric or non-parametric analysis for higher-level scale (e.g. interval scales) by their prudence. However, this processing may sometimes bring about a “false” statistical result.

Author of this paper was engaged in research of the medical field as a statistic professional, there are many papers treats higher level scale with lower level analysis according to text of statistics. On this paper, we clarify that lower level analysis for higher level scale is not “prudent” but rather bring into mistaken by showing simulation data and our realistic data. The date measured with a certain scale have to be analyzed with suitable statistics.

Parametric Multiple Contrast Tests in the Presence of Heteroscedasticity

Mario Hasler, Ludwig A. Hothorn

Leibniz University of Hannover, Germany

We describe a new method to facilitate multiple contrast tests for normally distributed data in the presence of heteroscedasticity. It keeps the α-level best whilst readily available methodology tends to yield conservative or liberal test decisions, respectively. Both differences in and ratios of means are addressed. We compare the new method with former ones by α-simulations.

• G Dilba, E Bretz, V Guiard, and L. A. Hothorn. Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. Methods Of Information In Medicine, 43(5):465–469, 2004.
• G Dilba, F Bretz, and V Guiard. Simultaneous confidence sets and confidence intervals for multiple ratios. Journal Of Statistical Planning And Inference, 136(8):2640–2658, August 2006.
• G Dilba and F Schaarschmidt. mratios: Inferences for ratios of coefficients in the general linear model, 2006. R package version 1.2.1.
• PA Games and JF Howell. Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. Journal of Educational Statistics, 1(2):113–125, 1976.
• Y Hochberg and AC Tamhane. Multiple comparisons procedures. John Wiley and Sons, Inc., 1987.
• T Hothorn, F Bretz, and P Westfall. multcomp: Simultaneous Inference for General Linear Hypotheses, 2006. R package version 0.991-5.
• R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
• FE Satterthwaite. An approximate distribution of estimates of variance components. Biometrics, 2:110–114, 1946.
• AC Tamhane and BR Logan. Finding the maximum safe dose level for heteroscedastic data. Journal of Biopharmaceutical Statistics, 14(4):843–856, 2004.
• BL Welch. The significance of the difference between two means when the population variances are unequal. Biometrika, 29:350–362, 1938.

A Simulation Study On the Gain in Power of Multiple Test Procedures by Using Information on the Number of True Hypotheses

Claudia Hemmelmann, Andreas Ziegler, Rüdiger Vollandt

Institute of Medical Stastistics, Computer Sciences and Documentation, University Klinikum Jena, Germany

It is known that the knowledge of the number of true hypotheses leads to increased power of some multiple test procedures. However, the number of true hypotheses is unknown in general and must be estimated. We aim at showing how the gain in power is dependent upon the accuracy of the estimation of the number of true hypotheses.
We simulate m-dimensional random vectors and employ different multiple test procedures by utilizing several upper bounds of the number of true hypotheses. We consider multiple test procedures which control the family-wise error rate (Holm method), the generalized family-wise error rate (Hommel and Hoffmann; procedure A of Korn and colleagues), the false discovery rate (Benjamini and Hochberg) and the false discovery proportion (Lehmann and Romano; procedure B of Korn and colleagues) and apply the average power and the all pairs power for the evaluation. Clearly, the more accurate the estimate of the number of true hypotheses is, the larger the gain in power. The power increases when the number of true hypotheses decreases. But this increase of power also depends upon the error rate and several distribution parameters. For example, the gain of power is independent from the correlation between the vector components for the procedure of Hommel and Hoffmann whereas the gain of power increases with increasing correlation for the procedure A of Korn and colleagues. We also compute the corresponding error rate by an underestimation of the number of true hypotheses. For some procedures and error rates, respectively, no underestimation is allowed, e.g. for the Holm method and procedure of Benjamini and Hochberg, whereas for others the number of true hypotheses can be underestimated up to 60-70 percent, e.g. for the procedure of Hommel and Hoffmann and procedure of Lehmann and Romano.

On Orthogonal Series Estimation Methods

Mei Ling Hunag, Percy Brill

Department of Mathematics, Brock University, Canada

This paper discusses nonparametric orthogonal series estimation methods. The main focus is on a Hermite series density estimator and a trigonometric series density estimator. The paper gives comparisons of the properties of these two estimators with other nonparametric density estimation methods, for example, kernel density estimation and other methods. Computational simulation results are obtained. The paper also discusses several examples of applications in medical research and other fields.

Study on Statistical Analysis for Adverse Drug Reaction in Korea

Hyeon Jeong Kim, Eunhee Kim, Mun Sin Kim, Junghoon Jang, Bong Hyun Nam

National Institute of Toxicological Research, Korea

In the case of developed countries, the spontaneous reporting systems for the adverse drug reactions and the management of their databases have been constructed systematically. However, the overall systems for the adverse drug reaction in Korea is insufficient compared to developed countries. In addition to the reporting cases of the adverse drug reaction have been recently increased due to the obligation of reporting them, but the statistical analysis methods for these data have been not studied sufficiently. So we investigated the spontaneous reporting systems for the adverse drug reactions and the statistical analysis methods in developed countries such as USA, UK, Australia, and WHO and applied the statistical methods in Korea data and compared the methods.

Testing Equality of Two Mean Vectors with Uniform Covariance Structure when Missing Observations Occur

Kazuyuki Koizumi, Toshiya Iwashita,Takashi Seo

Tokyo University of Science, Japan

We consider the test for equality of two mean vectors and the simultaneous confidence intervals when observations are missing at random in the intraclass correlation model. Hotelling's T^2 test for the equality of two mean vectors is given by an extension of Seo and Srivastava (2000) when the missing observations are of the monotone type. Finally, numerical example is presented.

Inequalities for Multivariate Normal Probabilities of Nonsymmetric Rectangles

Vered Madar

Tel-Aviv University, Israel

Šidák inequality (1967) provides a product bound to the joint normal probabilities of rectangles. It permits arbitrary correlation structure and also extend able to elliptically contoured distributions (Das Gupta et al 1971). As a such it has many useful applications in Multiple Comparisons Procedures. We extend the symmetric inequality by Šidák (1967) to a much stronger inequality on nonsymmetric rectangular regions, and show some applications.

Methodological Issues in the Design and Sample Size Estimation of a Cluster Randomized Trial to Evaluate the Effectiveness of Clinical Pathways

Sara Marchisio, Massimiliano Panella, Manzoli Lamberto, DiStanislao Francesco

University of Eastern Piedmont, Italy

Clinical pathways emerged as an important tool to reduce unnecessary variations and to improve the outcomes for patients. Despite enthusiasm and diffusion, the widespread acceptance of clinical pathways remain questionable because very little prospective controlled data demonstrated their effectiveness, mainly because of the complexity of the study design and management. We performed a cluster multi-centre randomized controlled clinical trial to evaluate the effect of applying clinical pathways to process and outcome indicators and to the costs sustained to assist the patients with heart failure. We compared the results obtained treating the patients with clinical pathways to the results obtained with the usual care. Since a clinical pathway is not a single intervention to be compared with a placebo but its eventual benefits come from a mix of complex actions that are implemented at the institutional level (appropriate use of practice guidelines and supplies of drugs and ancillary services, new organization and procedures, patient education, etc.), we randomly assigned hospitals, rather than individual patients, to either introduce the pathway or continue usual care. The primary outcome measure was in-hospital mortality. Since in Italy the in-hospital mortality rates range from 5% to 17%, we expected that clinical pathways succeeded to control mortality to 5% to be clinically relevant. Based on this goal a sample size of 424 patients (212 in each group) was required for the study to have 80% power at the 5% significance level (two-sided). We adjusted the sample size using an inflation factors of 2.015 to account for the cluster randomization (7 clusters per trial arm, cluster size of 30 patients, ICC of 0.035). In addition to common descriptive statistics (Fisher exact and Kruskal Wallis test for categorical and continuous variables, respectively), performed at the cluster level, the differences in the rate of in-hospital deaths and

unscheduled admissions across groups and according to each variable under study were evaluated using random-effects logistic regression, thus accounting for the clustering effect. Variables were included if significant at the 0.10 level (backward approach), with the exception of age which was forced to entry. The presence of multicollinearity, interaction and higher power terms was assessed to check final model validity. A cluster randomized trial have conceptual validity and relevant advantages in terms of patient's management and study expenditures. However, the conduction of a cluster randomized trial implies some specific ethical issues and, moreover, several methodological modifications in the statistical analysis and sample size estimation as shown in this paper. Therefore, the paper is intended as a methodological instrument to support the investigators in conducing a trial to evaluate complex interventions in healthcare.

Controlling the Number of False Positives Using the Benjamini-Hochberg Procedure

Paul Somerville (presented by title)

University of Central Florida, U.S.A.

In multiple hypotheses testing, it is challenging to adequately control the rejection of true hypotheses while still maintaining reasonable power to reject false hypotheses. For very large numbers of hypotheses, using the traditional family-wise error rate (FWER) can result in very low power for testing single hypotheses. Benjamini and Hochberg (1955) proposed a powerful multiple step procedure which controls FDR, the "False Discovery Rate". The procedure can result in a large number of false positives. Van der Laan, Dudoit and Pollard (2004) proposed controlling a generalized family-wise error rate *k*-FWER (also called gFWER(k)), defined as the probability of at least (*k*+1) Type I errors (*k*=0 for the usual FWER). Lehmann and Romano (2005) suggested new and simple methods of controlling *k*-FWER and the proportion of false positives (PFP) (also called False Discovery Proportion FDP). Somerville and Hemmelmann (2006) proposed controlling *k*-FWER by limiting the number of steps in step-up or step-down procedures. In this paper the procedure is applied to the Benjamini-Hochberg FDR procedure. Formulas are developed and Fortran 95 programs have been written. Tables are presented giving the maximum number of steps in the Benjamini-Hochberg procedure which will assure that $P(U \le k) \ge 1-\alpha$, for various values of k and a, where U is the number of false positives.

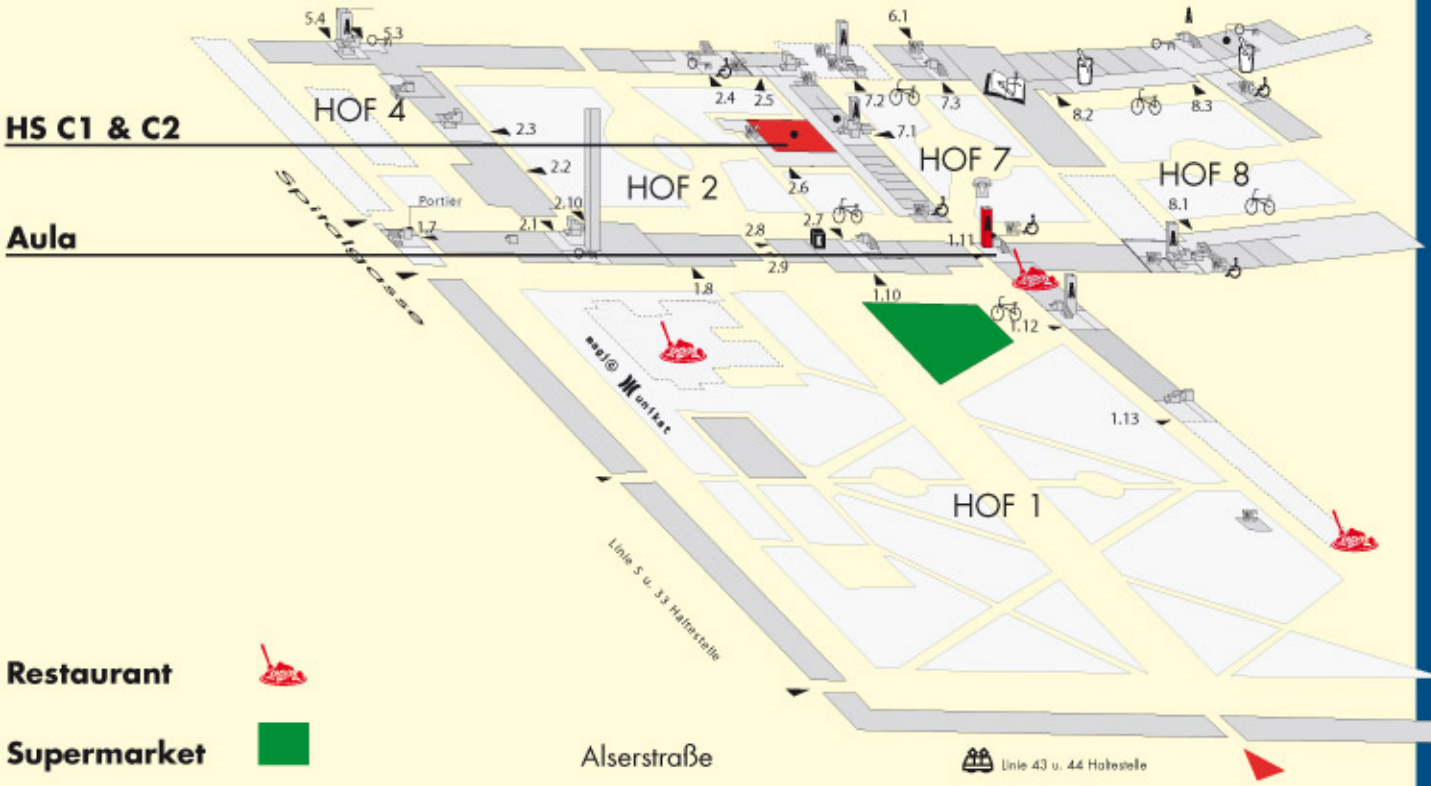
Presenters Index

Goll Alexandra	Wam2C1T3	115
Götte Heiko	Mpm1C2T2	41
Gould A. Lawrence	Wam1C1T3	106
Guilbaud Olivier	Tpm2C1T3	93
Hare David	Tpm1C2T1	82
Hashimoto Fumihiko	Poster Session 2	154
Hasler Mario	Poster Session 2	155
Heller Ruth	Tam2AT2	75
Hemmelmann Claudia	Poster Session 2	156
Hirotsu Chihiro	Wpm2C1T2	131
Holland Burt	Tpm1C2T2	83
Hommel Gerhard	Tam1C1T1	53
Hothorn Ludwig	Mpm2C1T2	46
Hsu Jason	Tam2C1T2	66
Hunag Mei Ling	Poster Session 2	157
Hyeon Jeong Kim	Poster Session 2	157
Inderjeet Kaushik	Poster Session 1	140
Jaki Thomas	Wam2C2T4	121
Joarder Anwar	Tam1AT4	64
Jowaheer Vandna	Tpm2AT2	101
Kelly Patrick	Tpm2C2T3	97
Kim Kyung In	Wpm1C2T2	127
Klingenberg Bernhard	Tam1C2T1	57
Koenig Franz	Mpm1C2T3	42
Koizumi Kazuyuk	Poster Session 2	158
Li David	Tpm1AT2	87
Lin Dan	Mam2C2T3	35
Lin Shan	Tam2AT3	76
Liu Jen-pei	Mam2C2T1	33
Liu Wei	Tpm1C2T3	84
Logan Brent	Wam1C1T2	105
Luta Gheorghe	Wpm2C1T3	132
Madar Vered	Poster Session 2	158
Marchenko Olga	Wam1C2T3	110
Marchisio Sara	Poster Session 2	159
Maurer Willi	Mpm1C2T1	40
Mehta Cyrus	Mam2C1T4	32
Mi Xuefei	Poster Session 1	148
Miller Frank	Wam1C2T1	108
Miwa Tetsuhisa	Tpm1AT1	86
Moussa Mohamed	Poster Session 1	144

Muino Jose M	Mpm2C2T3	51
Müller Hans-Helge	Tam2C2T1	70
Nadiradze Kakha	Poster Session 1	146
Nishiyama Takahiro	Tam1AT3	63
Ozturk Omer	Tpm1C2T4	85
Park Taesung	Poster Session 1	147
Peng Jianan	Tam2AT4	77
Ploner Alexander	Tpm2C2T1	95
Posch Martin	Tpm1C1T4	81
Reiner-Benaim Anat	Mpm1C1T3	39
Roehmel Joachim	Mpm2C1T1	44
Romano Joseph	Tpm1C1T1	78
Roquain Etienne	Wpm1C2T4	129
Sarkar Sanat	Wpm1C2T1	126
Schaarschmidt Frank	Poster Session 1	136
Scherag Andre	Tpm2C2T2	96
Schimek Michael G.	Tam1C2T4	60
Shaffer Juliet	Tpm1C1T2	79
Silva-Fortes Carina	Poster Session 1	137
Singh Parminder	Wpm1C1T4	125
Solari Aldo	Mpm2C1T4	48
Somerville Paul	Poster Session 2	161
Stallard Nigel	Mpm1C2T4	43
Strassburger Klaus	Tpm2C1T4	94
Sutradhar Brajendra	Tpm2AT1	100
Tamhane Ajit	Wam2C2T1	117
Timmesfeld Nina	Tam2C2T2	71
Troendle James	Wpm1C1T1	122
Ushijima Masaru	Mpm2C2T2	50
Vallarino Carlos	Wpm1C1T2	123
Vandemeulebroecke Marc	Tam2C2T4	73
Verbitskaya Elena V.	Poster Session 1	140
Victor Anja	Tpm2C2T4	98
Vock Michael	Tpm1AT3	88
Walther Mario	Wpm1C1T3	124
Wang Jixian	Wam1C2T2	109
Wang Sue Jane	Mam2C1T1	29
Wen Shu-Hui	Poster Session 1	142
Westfall Peter	Tpm2C1T1	90
Wolf Michael	Tam2C1T1	65
Wolfinger Russ	Wam2C1T1	113

Wu Samuel	Tpm2AT4	103
Yekutieli Daniel	Wam1C1T1	104
Young S. Stanley	Poster Session 1	143
Zagdanski Adam	Mpm2C2T4	52
Zehetmayer Sonja	Wam2C1T2	114
Zuber Emmanuel	Wam1C2T4	111

Conference Venue: Altes AKH (Campus) Spitalgasse 2, Hof 2, 1090 Wien



Corporate Donors and Sponsors

- ADDPLAN GmbH
- AstraZeneca PLC
- Baxter International Inc.
- Cytel Inc.
- Dr. Willmar Schwabe Arzneimittel
- F. Hoffmann-La Roche Ltd
- Johnson & Johnson Pharmaceutical Research & Development, L.L.C.
- Merck & Co, Inc.
- Merck KGaA Germany
- Novartis Pharma AG
- SAS Institute Inc.

Institutional Donors and Sponsors

- Austrian Statistical Society
- Section of Medical Statistics / Medical University of Vienna
- U.S. Food and Drug Administration
- Working Group "Multiple Methods" of the German Region / IBS

ISBN: 978-3-200-00977-6