The 8th International Conference on

Multiple Comparison Procedures

July 8 – 11, 2013

Highfield campus
University of Southampton
Southampton SO171BJ, UK

Organizing Committee

•Wei Liu (Co-Chair, University of Southampton, UK)

•Frank Bretz (Co-Chair, Novartis, Switzerland)

•Martin Posch (Co-Chair, Medical University of Vienna)

•Jason C. Hsu (Co-Chair, The Ohio State University, USA)

•Yoav Benjamini (Tel Aviv University, Israel)

•Chihiro Hirotsu (Meisei University, Japan)

•Vishwanath Iyer (Novartis Healthcare Pvt. Ltd., India)

•Ajit Tamhane (Northwestern University, USA)

•Bushi Wang (Boehringer-Ingelheim, USA)

•Sue-Jane Wang (FDA, USA)

•Daniel Yekutieli (Tel Aviv University, Israel)

Table of Contents

Sponsors of the MCP 2013 and the "Society for the Support of the International MCP Conference"

The Society for the Support of the International MCP Conference wishes to acknowledge the contribution of, and to express their warm appreciation to

S3RI and School of Maths, Southampton University
Novartis
Biometrical Journal
Wiley publisher
Millennium
Cytel

Keynote Speaker

- Gerhard Hommel (Universitats Mainz, Germany)

Southampton Statistical Sciences Research Institute (S3RI) Special Public Lecture

- Yoav Benjamini (Tel Aviv University)

Invited Speakers

- Carl-Fredrik Burman (Astra Zeneca)
- Thorsten Dickhaus (Humboldt-University Berlin)
- Anthony Hayter (University of Denver)
- Satoshi Kuriki (The Institute of Statistical Mathematics, Japan)
- Willi Maurer (Novartis Pharma AG)
- Cyrus Mehta (Cytel)

Invited Sessions

- Panel session on Open Access to Clinical Trial Data on Patient Level: opportunities and challenges
- Panel session on Errors in Multiple Testing: Big and Small, More or Less
- Key subgroup analysis issues in clinical trials
- Multiple testing in group sequential trials
- Adaptive designs with multiple objectives
- Multi-arm multi-stage clinical trials
- Selective inference
- Bayesian FDR

Information for Speakers

Each contribute talk has 20 minutes and each invited talk has 30 minutes (including question time).

Please bring your talk on an usb memory stick and load it on the computer in the Lecture Room before the start of your session. The chair of the session will be there to help you with this. Please could the chair of a session go to the session about 10 minutes before it starts.

Important Dates

July 08 Short courses (9:00 am – 12:30 pm / 1:30 – 5:00 pm)
July 08 Social Mixer: Staff Social Centre Hartley Suite (5:30 – 7:30)
July 09 Start main conference
July 09 Conference Dinner at Garden Court; group picture (6 – 9pm)
July 11 Conference end
July 12 Conference excursion: visit Stonehenge and Winchester (9:00am-5:00pm)

Important Location

Nightingale Lecture Theatre (NLT)	Room 1027, Building 67
Arts Lecture Theatre H (ALT H)	Room 2065, Building 02a
Arts Lecture Theatre J (ALT J)	Room 2077, Building 02a
EEE Lecture Theatre	Room 1015, Building 32
Staff Social Centre Hartley Suite	Building 38 and 40
Garden Court	Building 38 and 40

Campus map
http://www.southampton.ac.uk/visitus/campuses/highfield.html

Scientific Program

Short Courses

Monday, 8 July: 9:00 am – 12:30
ALT H (02a/2065)

Fundamentals of Multiple Testing and Biotechnology with Applications to Clinical Trials and Personalized Medicine
Jason C. Hsu and Xinping Cui

Monday, 8 July: 9:00 am – 12:30
ALT J (02a/2077)

Adaptive Designs
Martin Posch and Franz Koenig

Monday, 8 July: 1:30 – 5:00pm
ALT H (02a/2065)

Graphical approaches to multiple test problems
Frank Bretz, Ekkehard Glimm, and Willi Maurer

Monday, 8 July: 1:30 – 5:00pm
ALT J (02a/2077)

Gatekeeping procedures in clinical trials
Alex Dmitrienko

Sessions

Tuesday, 9 July, 9:00 – 10:30 am
Plenary Session
NLT (67/1027)

Keynote: Is the Use of p-values Adequate for the Presentation of Multiple Comparison Procedures?

by Gerhard Hommel

Eulogy to Yosi Hochberg

by Ajit Tamhane and Yoav Benjamini

Tuesday, 9 July, 11:00 am – 12:30 pm

Estimation in adaptive designs Chair: Franz Koenig	Multiple comparison under various models Chair: Peter Westfall
ALT H (02a/2065)	ALT J (02a/ 2077)
Exact Inference for Adaptive Group Sequential Designs Cyrus Mehta	**Multiple Wald Tests with Applications to Dynamic Factor Models** Thorsten Dickhaus
Maximum Likelihood Estimation following Interim Adaptations Alexandra Graf	A General Framework for Multiple Comparisons of Treatments with Ordinal Responses Tong-Yu Lu
Point and Interval Estimation following Adaptive Seamless Designs Peter K Kimani	Interval-Wise Control of the Family Wise Error Rate: a New Inferential Procedure in Functional Data Analysis Alessia Pini
The Sensitivity of the Triple Sampling Sequential Procedure to Departures from Normality Alan Kimber	MCPs for Non-Gaussian Distributed Endpoints- Using R Ludwig Hothorn

Tuesday, 9 July, 1:30 – 3:00 pm

Invited session: Multiple testing in group sequential trials Chair: Ajit Tamhane	Approaches to multiplicity (I) Chair: Ruth Heller	Miscellaneous topics (I) Chair: Siu Hung Cheung
ALT H (02a/2065)	ALT J (02a/2077)	(02/1039)
Group Sequential Procedures for Multiple Endpoints with Adaptive Allocation of Recycled Significance Levels to Stages Ajit Tamhane	Using Scan Statistics on Multiple Processes with Dependent Variables, with Application to Genomic Sequence Search Anat Reiner-Benaim	Improving Oncology Clinical Program by Use of Innovative Designs and Comparing Them via Simulations Olga Marchenko
Adaptive Group Sequential Design with Treatment Selection and Sample Size Reestimation Lingyun Liu	Smoothing of Stepwise Procedures Alexander Y Gordon	~~A Selection Procedure for Selecting the Least Increasing Failure Rate Average Distribution~~ ~~Suresh Kumar Sharma~~
Intelligent Interim Analysis Strategy David Li	Confidence Set for a Maximum Point of a Univariate Polynomial Regression Function in a Given Interval, with Extensions to Other Models Fang Wan	Controlling the Triple Sampling Power While Making Inferences for the Mean Ali Yousef
		Statistical Estimation for Fixation Points of Eye Movements Toshinari Kamakura

Tuesday, 9 July, 3:30 – 5:00 pm

Panel Session (I) Chair: Martin Posch and Franz Koenig (3:30-5:30pm)	Computation of multivariate probabilities Chair: Jianan Peng	Miscellaneous topics (II) Chair: David Li
ALT H (02a/2065)	ALT J (02a/2077)	(02/1039)
Open access to clinical trial data on patient level: opportunities and challenges Peter Bauer Yoav Benjamini Simon Day Trish Groves Franz Koenig Thomas Lang Martin Posch Sue-Jane Wang	**Recursive Integration Methodologies with Applications to Multiple Comparisons** Anthony Hayter Efficient Evaluation of Polyhedral Probabilities by Perturbing and Decomposing Polyhedra Tetsuhisa Miwa ~~Multiple Comparisons with a Control in Direction-Mixed Families of Hypothesis under Heteroscedasticity~~ ~~Parminder Singh~~ Multiple Comparison Procedure for Identifying The Minimum Effective Dose with More Than One Control Rajvir Singh Chauhan	Analysing Anti-malarial Trials: Using a False-claim Error Rate Alice Parry A Note on Comparing Several Variances with a Control Variance Anju Goyal The Influence of Shift-Variant data on Factor Analysis Fumihiko Hashimoto Stochastic Modeling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa District Mikiyas Gebresamuel Gebru

Wednesday, 10 July, 9:00 – 10:30 am

Invited session: Adaptive designs with multiple objectives Chair: Olga Marchenko	Multiple testing (I) Chair: Geraldine Rauch	Simultaneous confidence bands Chair: Wei Liu
ALT H (02a/2065)	ALT J (02a/2077)	(02/1039)
Using Early Outcome Data for Decision Making in Seamless Phase II/III Clinical Trials Nigel Stallard	**Geometrical Representation and Classification of Closed Consonant Weighting Schemes and Associated Multiple Tests** Willi Maurer	**Simultaneous Confidence Bands for Polynomial Regression Curves with the Volume-of-Tube Formula** Satoshi Kuriki
Complex Multiplicity Problems in Clinical Trials with Adaptive Sample Size Adjustment Jeffrey Maca	Weighting and Ordering Considerations for Multiple Testing Procedures in Clinical Trials Brian L. Wiens	Simultaneous Inference for Low Dose Risk Estimation with Quantal Data in Benchmark Analysis Jianan Peng
Enrichment Designs for the Development of Personalized Medicines Martin Posch	Mixed Directional False Discovery Rate Control in Multiple Pairwise Comparisons Using Weighted P-values Xinping Cui	Simultaneous Confidence Bands for a Percentile Line in Linear Regression with Application to Drug Stability Studies Yang Han
	Multiple Testing Method for The Directed Acyclic Graph, Using Shaffer Combinations Rosa Meijer	Comparisons of Simultaneous Confidence Bands for Linear Regression Shan Lin

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

Invited session: Multi-arm multi-stage clinical trials Chair: Chris Jennison	Multiple testing (II) Chair: Toshimitsu Hamasaki	Invited session: Selective inference Chair: Daniel Yekutieli
ALT H (02a/2065)	ALT J (02a/2077)	(02/1039)
Group Sequential Designs: Theory, Computation and Optimisation Chris Jennison	Calibration of P-values via the Dirichlet Process Mikelis Guntars Bickis	**Valid Post-Selection Inference** Andreas Buja
Optimal Design for Multi-Arm Multi-Stage Clinical Trials James Wason	Adjusted p-values for SGoF Multitesting Procedure. Definition and Properties Irene Castro Conde	**Selection Adjusted Confidence Intervals with More Power to Determine the Sign** Asaf Weinstein
Designing Multi-Arm Multi-Stage Clinical Trials with a Safety and an Efficacy Endpoint Thomas Jaki	New Multiple Testing Method under no Dependency Assumption, with Application to Multiple Comparisons Problem Li Wang A Sufficient Criterion for Control of Generalised Error Rates in Multiple Testing Sebastian Doehler	**Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression** Ulrike Schneider

Wednesday, 10 July, 1:30 – 3:00 pm

Invited session: Key subgroup analysis issues in clinical trials Chair: Alex Dmitrienko	Multiple testing (III) Chair: Helmut Finner	Gate keeping Chair: Ajit Tamhane
ALT H (02a/2065)	ALT J (02a/2077)	(02/1039)
Bayesian Subgroup Analysis James Berger	Directional Error Rates of Closed Testing Procedures Peter Westfall	Powerful Mixture-Based Gatekeeping Procedures in Clinical Trials Alex Dmitrienko
Overview of Subgroup Identification Approaches in Clinical Research Ilya Lipkovich	A Unifying Approach to the Shape and Change-Point Hypotheses in the Univariate Exponential Family Chihiro Hirotsu	Cyclic Stack Procedures with Parallel Gatekeeping George Kordzakhia
Region, Biomarker Subset or Patient Subpopulation: Are They Multiplicity Problem and When? Sue-Jane Wang	On the Moderated t-test and its Moderated p-values Jelle Goeman	Gatekeeping Procedures for Multiple Correlated Endpoints Including Responder Endpoints Yu-Ping Li
	Pairwise Comparisons of Treatments with Ordered Categorical Responses Yueqiong Lin	

Wednesday, 10 July, 3:30 – 5:00 pm

Multiplicity issues in complex clinical trials Chair: Bushi Wang	Higher criticism and goodness of fit tests Chair: Anthony J. Hayter
ALT H (02a/2065)	ALT J (02a/2077)
Complex Multiple Comparison Problems When Multiple Trials are Evaluated H.M. James Hung	Higher Criticism Test Statistics: Why Is The Asymptotics So Poor? Veronika Gontscharuk
Consistency-Adjusted Alpha Allocation Methods for Composite Endpoints Geraldine Rauch	Some New Results on Goodness of Fit Tests in Terms of Local Levels Sandra Landwehr
Sample Size Considerations in Complex Clinical Trials Toshimitsu Hamasaki	Normal Probability Plots with Confidence Wanpen Chantarangsi
Thresholding of a Companion Diagnostic Test Confident of Efficacy in Targeted Population Jason C. Hsu	

Wednesday, 10 July, 5:30 – 6:30 pm

EEE Lecture Theatre (32/1015)
S3RI Special Public Lecture: Are most research findings really false? Yoav Benjamini

Thursday, 11 July, 9:00 am – 10:30 am

Multiple testing in sequential designs Chair: Lingyun Liu	Using dependency in multiple testing Chair: Jelle Goeman
ALT H (02a/2065)	ALT J (02a/2077)
Multiple Testing in Group Sequential Trials using Graphical Approaches Frank Bretz	Asymptotic FDR Control under Weak Dependence and the Null Problem: A Counterexample Helmut Finner
Fixed Sequence Testing in Adaptive Designs with Sample Size Reassessment Franz Koenig	How to take into account dependency into multiple testing procedures? Etienne Roquain
Sequentially Rejective Graphical Procedures in Adaptive Treatment Selection Designs Toshifumi Sugitani	P-value Evaluation for Multiple Testing of Means under the Existence of Positive Correlations Yoshiyuki Ninomiya
Flexible Sequential Designs for Multi-Arm Clinical Trials Dominic Magirr	Permutation-Based Confidence Bounds for the False Discovery Proportion Aldo Solari

Thursday, 11 July, 11:00 am – 12:30 pm

Approaches to multiplicity (II) Chair: Veronika Gontscharuk	Approaches to multiplicity (III) Chair: Xinping Cui
ALT H (02a/2065)	ALT J (02a/2077)
A Decision-Theoretic Approach to Multiple Inference Carl-Fredrik Burman	Family-Wise Control of Both Type I and Type II Errors in Clinical Trials Bushi Wang
Adaptive Statistical Significance Threshold for Inference Guided Discovery Studies Cheng Cheng	Type II Generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination Jérémie Riou
Testing and Multiple-testing using Neutral-data Comparisons Dan J. Spitzner	Number of False Rejections and Differential Equations Marsel Scheer
Implementing False Discovery Rate Procedures for Simulation-Based Tests With Bounded Risk Georg Hahn	Scaled Fasle Discovery Proportion and Related Error Metrics Djalel Eddine Meskaldji

Thursday, 11 July, 1:20 – 2:50 pm

Panel session (II) Chair: Jason Hsu and Frank Bretz	Invited session: Bayesian FDR Chair: Daniel Yekutieli
ALT H (02a/2065)	ALT J (02a/2077)
Errors in Multiple Testing: Big and Small, More or Less James Berger Frank Bretz Jason Hsu H.M. James Hung Thomas Lang Willi Maurer Martin Posch Sue-Jane Wang David Wright	**Replicability Analysis for Genome-Wide Association Studies** Ruth Heller **Bayesian Variable Selection and Multiplicity Adjustment** Ziv Shkedy **FDR and FNR: Comparison of Loss Functions in Epidemiologic Surveillance** Annibale Biggeri

Talks

Abstracts are sorted by session.

Keynote

Is the Use of p-values Adequate for the Presentation of Multiple Comparison Procedures?

Gerhard Hommel

Institute of Medical Biostatistics, Epidemiology and Informatics
Universitätsmedizin Mainz, Germany

P-values are used as a tool for the construction of many multiple comparison procedures as well as of adaptive designs. First, I will describe several situations of this type and different possibilities how p-values are used and whether their use might be necessary.

However, p-values were critically considered in the past, in particular in mathematical statistics. Therefore, I will make some general remarks on the nomenclature about p-values, also with respect to the use in multiple testing and adaptive designs.

The most critical point is whether p-values can be considered as random variables. This point is discussed intensively. It is demonstrated that for the most practical situations p-values are random variables, in fact; however, one can construct counterexamples though these appear to be not quite realistic.

Exact Inference for Adaptive Group Sequential Designs

Cyrus Mehta, Ping Gao, Lingyun Liu

Cytel Inc.,USA

Methods for controlling the type-1 error of an adaptive group sequential trial were developed in seminal papers by Cui, Hung and Wang (Biometrics, 1999), Lehmacher and Wassmer (Biometrics, 1999), and \ms (Biometrics, 2001). However, corresponding solutions for the equally important and related problem of parameter estimation at the end of the adaptive trial have not been completely satisfactory. In this paper a method is provided for computing a two sided confidence interval having exact coverage, along with a point estimate that is median unbiased, for the primary efficacy parameter in a two arm adaptive group sequential design. The possible adaptations are not confined to sample size alterations but also include data dependent changes in the number and spacing of interim looks and changes in the error spending function. The procedure is based on mapping the final test statistic obtained in the modified trial into a corresponding backward image in the original trial. This is an advance on previously available methods, which either produced conservative coverage and no point estimates, or else provided exact coverage for one-sided intervals only.

Maximum Likelihood Estimation following Interim Adaptations

Alexandra Graf, Georg Gutjahr, Werner Brannath

Section for Medical Statistics, Medical University of Viennab.
Competence Center for Clinical Trials, University of Bremen, Austria

There has been increasing interest over the last years in adaptive two-stage clinical trials where more than one treatment groups are compared to a common control. These trials allow for design adaptations as e.g. sample size reassessment or treatment selection at an interim analysis. Hypothesis testing methods have been developed that allow for such adaptations without compromising the overall type one error rate. Whereas the methodological issues concerning hypothesis testing are well understood, up to now, it is not clear how to deal with parameter estimation after interim adaptations. It is well known, that for such designs, the maximum likelihood estimator (MLE) may be biased. Several methods have been proposed to reduce the bias. However, these methods do only apply to specific adaptation rules and hence are not generally applicable. In particular, in designs were adaptation rules are not fixed in advance, estimation is still an unsolved issue, so that in practice the MLE is still used. Therefore, we investigate the behavior of the MLE: We investigate the sample size reassessment and selection rule leading to the maximal bias or maximal MSE respectively when using the MLE at the end of the adaptive trial. We thereby consider scenarios where more than one treatment groups are compared to a common control, with and without interim treatment selection. We consider the case of unlimited sample size as well as scenarios with restrictions on the sample size reassessment rules.

Point and Interval Estimation following Adaptive Seamless Designs

Peter K Kimani, Susan Todd, Nigel Stallard

University of Warwick, UK

In order to accelerate drug development, adaptive seamless designs (ASDs) have been proposed. We consider two-stage ASDs where in stage 1, data are collected to compare several experimental treatments to a control. At stage 2, further data for the control and the experimental treatment that shows highest benefit over the control based on stage 1 data, are collected. The final confirmatory analysis that is performed at the end of stage 2 includes stages 1 and 2 data. Although such designs are efficient because unlike traditional designs, data used for treatment selection are also used in the confirmatory analysis, they pose statistical challenges in making inference. We will focus on point and interval estimation at the end of the trial for the treatment difference of the selected treatment and the control. Estimation is challenging because the control is compared to multiple experimental treatments at stage 1, and the experimental treatment that appears to be the most effective is selected which may lead to bias. Estimators derived need to account for this fact. In this talk, we will describe the characteristics for a new bias adjusted point estimate and a new unbiased point estimate that we have developed. For interval estimation, we will describe the general methodology used to construct confidence intervals and give results of a comprehensive comparison of the various methods for constructing confidence intervals following an ASD.

The Sensitivity of the Triple Sampling Sequential Procedure to Departures from Normality

Alan Kimber, Ali Yousef

Southampton Statistical Sciences Research Institute, University of Southampton, UK

In 1981 Hall introduced the sequential triple sampling procedure for point and interval estimation of the mean from a normal population with unknown variance. It involves up to three stages: a pilot stage, a main study stage and a fine tuning stage and it uses a stopping rule that mimics the fixed sample size rule, had the variance been known. The properties of this procedure are well understood in the case of exact normality of the underlying distribution. However, in practice the exact form of the underlying distribution is rarely known. In this paper we discuss asymptotic results for the properties of the procedure when the underlying distribution is unknown. More specifically, we consider underlying distributions that are in a neighbourhood of the normal, as are commonly used in classical robustness studies, to investigate the robustness of point and interval estimates derived from the sequential triple sampling procedure. Provided the underlying distribution is not too far from normal the asymptotic results are shown to agree well with estimated finite sample results. Some practical implications are also discussed.

Multiple Wald Tests with Applications to Dynamic Factor Models

Thorsten Dickhaus

Humboldt-University Berlin, Department of Mathematics

We are concerned with families of linear hypotheses in parametric statistical models. Exploiting structural properties of multivariate chi-squared distributions, we construct critical regions for vectors of Wald statistics in such models, controlling the family-wise error rate or the false discovery rate, respectively. In this, we make use of the asymptotic distribution of these vectors for large sample sizes, assuming that the model is identified and model restrictions are testable. Under these and further regularity assumptions, Wald statistics are asymptotically equivalent to likelihood ratio statistics, but often easier to compute in practice.

As a specific (non-standard) application, we elaborate simultaneous statistical inference methods in dynamic factor models (DFMs). DFMs are popular tools in econometrics to describe the dynamics of a multi-dimensional time series by a lower-dimensional set of (possibly latent)common factors. The resulting error terms are referred to as specific factors. Based on central limit theorems for time series regression developed by Hannan (1973),several problems of interest in practice can be formalized as linear hypotheses regarding parameters of the frequency-domain representation of the model. This extends the work of Geweke and Singleton (1981) on likelihood-based inference in DFMs. For instance, we address the questions "Which of the specific factors have a non-trivial autocorrelation structure?" and "Which of the common factors have a lagged influence on the observable process?", demonstrating the relevance of the proposed methods for practical applications.

Geweke, J. F., Singleton, K. J. (1981) Maximum likelihood "confirmatory" factor analysis of economic time series. Int. Econ. Rev. 22, 37-54.
Hannan, E. (1973) Central limit theorems for time series regression. Z. Wahrscheinlichkeits-theor. Verw. Geb. 26, 157-170.

A General Framework for Multiple Comparisons of Treatments with Ordinal Responses

Tong-Yu Lu, Wai-Yin Poon, Siu Hung Cheung

China Jiliang University

Several latent variable models have been employed to analyze ordinal categorical data which can be conceptualized as manifestations of an unobserved continuous variable. In this project, we suggest to use a more general framework to develop appropriate latent variable models for the comparison of treatments with ordinal responses. The proposed class of models is based on the location-scale family and is rich enough to include many important existing models for analyzing ordinal categorical variables, including the proportional odds model, the ordered probit-type model, and the proportional hazards model. An estimation procedure is provided for the identification and estimation of the general latent variable model, which allows for the location and scale parameters to be freely estimated. The flexibility of the proposed framework enables us to generate useful testing procedures that facilitate important statistical inferences such as location and/or dispersion comparisons among treatments. Examples are given to illustrate the proposed methods.

Interval-Wise Control of the Family Wise Error Rate: a New Inferential Procedure in Functional Data Analysis

Alessia Pini, Simone Vantini

MOX, Department of Mathematics "F. Brioschi", Politecnico di Milano, Italy

We present a novel inferential technique for functional data (i.e., Interval Testing Procedure, or ITP) that is based on the interval-wise control of the Family Wise Error Rate (FWER). The procedure starts with the representation of data on a suitable ordered functional basis. Then, the significance of each coefficient of the expansion is tested by means of joint permutation tests. Finally, the univariate results are combined by means of the implementation of multivariate NPC tests on intervals. This latter combination of p-values enables the construction of a p-values heatmap and the multiplicity correction of the marginal results.

In addition to showing that the ITP have an interval-wise control of the FWER, we prove the two following theorems: (i) the global level of the ITP is bounded above by the global level of the Global Testing Procedure (which however provides only a weak control of the Family Wise Error Rate and does not provide any guide to the interpretation of the test result) and from below by the global level of the Closed Testing Procedure (which provides a strong control of the Family Wise Error Rate but it is computationally unfeasible in the functional framework). The same result holds for the global power of the procedure; (ii) marginally for each component, the level of the ITP is bounded above by the level of the global test and from below by the level of the Closed Testing Procedure. The same result holds for the marginal power.

By means of a simulation study, we evaluate the tightness of such inequalities in different scenarios of increasing aggregated false hypotheses, showing that the ITP behave more similarly to the CTP when the number of false hypotheses is low, and more similarly to the GTP when the number of false hypotheses is high. In the same scenarios, we compare the ITP with both the Bonferroni-Holm and the Benjamini-Hochberg procedures. In this latter case, we verify empirically that the ITP seems to outperform the Benjamini-Hochberg

correction procedure and the Bonferroni-Holm correction procedure on all false hypotheses not occurring at the boundaries between ``true'' and ``false'' regions. This latter finding may suggest the use of the ITP when the false and true hypotheses are expected to be aggregated in intervals of basis components.

Finally, we apply the ITP to two different case studies: inference for the mean function of the daily temperature profiles in Milan (Italy) using a Fourier basis expansion based on the amplitude-phase decomposition; and inference for the difference between vascular geometry and hemodynamic features of the Internal Carotid Artery of two groups of subjects with a different severity of aneurysm pathology, using a B-spline basis expansion and selecting the zones of the carotid artery that are significantly different between the two groups.

MCPs for Non-Gaussian Distributed Endpoints- Using R

Ludwig Hothorn

Leibniz University Hannover, Germany

There is a considerable discrepancy between the MCP-methods assuming Gaussian distribution and homogeneous variances in the literature and the common software, and the practically occurring different types of endpoints, namely: i) proportions (e.g. tumor rates), ii) skewed distributed endpoints (e.g. the ASAT enzyme in toxicology), iii) survival functions, iv) mortality-adjusted tumor rates (poly-3 estimates without cause-of-death information), v) counts (and proportions) with between-subject-variability (overdispersion) (e.g. number of micronuclei), vi) ordered categorical data (e.g. graded histopathological findings). Based on the asymptotic approach in general parametric models (Hothorn et al. 2008) and the R packages multcomp, MCPAN and SimpComp, by means of case studies the estimation of related simultaneous endpoints for different contrast matrices are demonstrated, such as Dunnett-type, Williams-type and Grand-Mean-type.

Moreover, the usefulness of a non-parametric version for relative effects (Konietschke et al. 2012) is demonstrated using the R package nparcomp and ratio-to-control tests are explained using the R package, particularly in the case of variance heterogeneity.

Hothorn,T; Bretz,F. and Peter Westfall. Simultaneous inference in general parametric models. Biometrical Journal, 50(3):346{363, (2008).

Konietschke, F; L.A. Hothorn, Brunner, E. Rank-based multiple test procedures and simultaneous confidence intervals Electronic Journal of Statistics Vol. 6 (2012) 737–758.

Group Sequential Procedures for Multiple Endpoints with Adaptive Allocation of Recycled Significance Levels to Stages

Dong Xi, Ajit C. Tamhane

Department of IEMS, Northwestern University

Bretz et al. (2009) and Burman et al. (2009) proposed a graphical approach for testing hypotheses on multiple endpoints when the significance levels from rejected hypotheses are recycled to unrejected hypotheses. Maurer and Bretz (2013) extended this graphical approach to constructing a group sequential procedure (GSP). The extension requires a scheme to allocate the recycled significance level to the stages of the GSPs for unrejected hypotheses. Maurer and Bretz's scheme allocates the recycled significance level to all stages. Ye et al. (2013) proposed a similar scheme. However, this scheme is not efficient since it allocates some of this significance level to the stages previous to the one when recycling takes place, and these previous stages cannot be revisited. Ye et al. (2013) proposed another scheme which allocates the recycled significance level only to the final stage, but that scheme also has drawbacks. We refer to these as non-adaptive allocation schemes. We propose an adaptive allocation scheme which recycles the significance level only to the current and future stages. It enhances power and reduces the expected sample size while strongly controlling the familywise error rate. Both efficacy and futility error spending functions are considered. Simulation studies to compare powers of alternative schemes are carried out and illustrative examples are given.

Adaptive Group Sequential Design with Treatment Selection and Sample Size Reestimation

Lingyun Liu

Cytel Inc., USA

Multi-arm multi-stage designs are very efficient to evaluate several treatments against a common control. The sample size and study duration of such design tends to be smaller since it allows early stopping for overwhelming efficacy or early dropping of ineffective treatments. However the adoption of such design in practice is rather limited due to the fact that there are few available methods. The other major hurdle for such design to be widely used in practice is the lack of powerful software package for design and simulation. Magirr, Jaki and Whitehead (2012) generalize the Dunnett test to derive boundaries for monitoring such trials with normal endpoint and known variance. Their method monitors such trials based on the maximum of the Wald statistics. We propose a general framework of monitoring such trials using the score statistics. As it is for two arm design, the score statistics have independent increments which facilitate the computation of the stopping boundary. In addition, treatment selection and sample size reestimation at interim are allowed. Valid inference will also be provided.

Intelligent Interim Analysis Strategy

David Li

Pfizer

Spending functions have become a very popular approach in clinical trials where several interim analyses are planned. This presentation will introduce an adaptive alpha spending approach when the interim analysis timings are pre-determined. The advantage of the proposed approach is that one can adjust the alpha to be spent at the next interim analysis based on the observed results at the current interim analysis: one can spend more next time if observed results are promising, or less otherwise, or even can skip the next interim analysis by spending zero if the next interim analysis is deemed to be unrewarding.

Using Scan Statistics on Multiple Processes with Dependent Variables, with Application to Genomic Sequence Search

Anat Reiner-Benaim

University of Haifa, Israel

The problem of locating sequences of interest along the genome is frequently confronted by genome researchers. The challenge here is to identify short intervals within noisy and much longer sequences, which exhibit certain behaviour. One example is the search for introns, which are DNA intervals that are spliced out on the path to synthesize proteins. Inference on the presence of intronic intervals can be made using genome-wide expression data produced by the tiling array technology. A scan statistic is suggested to test whether an interval, within a specified search region, is exhibiting the behaviour expected to occur in an intronic interval. The statistic integrates several important considerations related to the dependence between adjacent measures of expression along the genomic sequence. An analytical assessment of the scan statistics distribution considering this dependence is presented, along with its effect on FDR and power when testing simultaneously many random processes (genes). The performance of the suggested analysis is compared to methods that search continuously along the whole-genome sequence, and consider alternative false discovery criteria.

Smoothing of Stepwise Procedures

Alexander Y Gordon

University of North Carolina at Charlotte, USA

Cohen, Kolassa and Sackrowitz (2007) introduced the concept of a smooth multiple testing procedure. They found that the standard stepwise procedures are not smooth and came up with a method of approximating a stepwise procedure by a smooth one.

However, such smoothing may destroy certain "good" properties of a procedure, including monotonicity.

In the talk, an alternative method of smoothing of stepwise procedures will be discussed, which leaves intact their important desirable properties.

Confidence Set for a Maximum Point of a Univariate Polynomial Regression Function in a Given Interval with Extensions to Other Models

Fang Wan, Wei Liu, Frank Bretz, Yang Han

S3RI and School of Mathematics, University of Southampton, UK

The determination of a maximum point of the regression function in a constrained covariate region is often of great interest in regression analysis. Since the regression function needs to be estimated and its maximum point can only be estimated based on the random observations, the focus of the research is therefore to construct a confidence set for a maximum point of the regression function in the given covariate region.

In this paper, we consider the construction of a confidence set for a maximum point of a general univariate polynomial regression function in a given interval, with an extension to generalized linear models. Our method is based on inverting a family of acceptance sets of corresponding hypothesis tests (Neyman, 1937). Examples are given to illustrate our method and compare the confidence sets constructed using our method with those using other methods.

Improving Oncology Clinical Program by Use of Innovative Designs and Comparing Them via Simulations

Olga V Marchenko, Joel Miller, Tom Parke, Inna Perevozskaya, Jiang Qian, Yanping Wang

VP, Innovation, Quintiles, USA

The design of an oncology clinical program is much more challenging than the design of a separate study. The standard approach has been proven to be not very successful during the last decade; the failure rate of Phase 3 studies in oncology is about 66%. Improving the development strategy by applying innovative statistical methods is one of the major objectives for biostatisticians designing and supporting oncology clinical programs. Modeling and simulation approaches can help to optimize an individual trial, to see the benefit of novel designs, and to increase success of a clinical program by making better decisions while developing a drug. This presentation is built on the work of the DIA ADSWG oncology sub-tem on an Adaptive Program. With representatives from a number of institutions, this group compared four hypothetical oncology development programs using probability of the clinical program success and expected net present value (eNPV). Simulated scenarios were used to motivate and illustrate the key ideas.

A Selection Procedure for Selecting the Least Increasing Failure Rate Average Distribution

Suresh Kumar Sharma

Panjab University, Chandigarh, India

For k independent absolutely continuous increasing failure rate average (IFRA) life distributions , Deshpande (1983) proposed a measure of departure of from exponentiality. Later, Link (1989) considered another measure of this departure against monotone failure rate average alternatives. In this paper, we use the measure defined by Link for detection of IFRA-ness of life distribution . A two stage selection procedure is proposed to select the least IFRA distribution, that is, the distribution associated with the smallest value of the measure. This selection procedure is based on a U-statistic, an estimator of the measure and can be implemented even when the IFRA life distributions belong to different families. The applications of this procedure are discussed for some well-known distributions viz Gamma, Generalized Exponential and Lehmann Type. Simulation study is conducted for finding the probability of correct selection.

Controlling the Triple Sampling Power While Making Inferences for the Mean

Ali Yousef, Alan Kimber, Mun S. Son, Hosny Hamdy

Kuwait University, KW

This paper gives a rigorous account of the sensitivity of the triple sampling sequential confidence interval for the mean of an unknown continuous underlying distribution, where the first six moments are assumed exist but are unknown. This problem arises, for example, in quality control, where careful attention should be given to controlling the Type II error probability while monitoring the quality mean. First, a triple sampling sequential confidence intervals for the mean is constructed using methodology developed by Hall (1981) and then asymptotic characteristics of the constructed interval are discussed and justified under Hall's conditions. Moreover, an asymptotic second order approximation of a differentiable and bounded function of the stopping variable is given and investigated in calculating Type II error probability. The impact of several parameters on Type II error is explored under some underlying distributions; normal, uniform and exponential.

Statistical Estimation for Fixation Points of Eye Movements

Toshinari Kamakura, Kosuke Okusa

Chuo University Tokyo, Japan

The problem of estimation of fixation points of human beings is very important in the field of computer visions and marketing science. The centers of eyeballs are tracked by the eye cameras, and two-dimensional coordinates of the spatial locations are recorded with time stamps. The eye fixation points can be defined by the locations where the velocity of the eye movement is kept below the comparatively slow velocity. The problem of estimating the points can be formulated by two dimensional change point problems. In this article we propose new method of estimating spatial fixation points considering spatial and time correlation. For controlling the error rate for detecting simultaneous spatial change points we can use FDR (Bnajamini and Hochberg, 1995) and Bayesian estimation (Barry and Hartigan, 1993) techniques. We can illustrate the stable estimated regions for fixation points based on real eye movement data.

Recursive Integration Methodologies with Applications to Multiple Comparisons

Anthony Hayter

University of Denver

This talk discusses how recursive integration methodologies can be used to evaluate high-dimensional integral expressions.

This has applications to many multiple comparisons problems where critical point evaluations often require such high-dimensional integral evaluations.

Recursive integration can allow an integral expression of a given dimension to be evaluated by a series of calculations of a smaller dimension.

This significantly reduces the computation time.

Applications of the recursive integration methodology are illustrated with several examples.

Efficient Evaluation of Polyhedral Probabilities by Perturbing and Decomposing Polyhedra

Tetsuhisa Miwa, Satoshi Kuriki, Anthony J. Hayter

National Institute for Agro-Environmental Sciences, Japan

In many applications we need to evaluate probabilities inside polyhedra bounded by linear hyper planes. In this paper we show an efficient method to evaluate probabilities inside polyhedra by expressing them by polyhedral cones. Varchenko (1987) and Lawrence (1991) gave a procedure to express a polyhedron in general position by cones. First we shall show another simple proof of their procedure, which leads to an easy algorithm for implementing the procedure on the computer. In the second part of the paper, we provide a perturbation method which enables us to apply the procedure to polyhedra which are not in general position. If normal probability is concerned, we can use the recursive integration by Miwa, Hayter and Kuriki (2003) to evaluate cone probabilities. Then we can evaluate any polyhedral probabilities.

Multiple Comparisons with a Control in Direction-Mixed Families of Hypothesis under Heteroscedasticity

Rajvir Singh chaau, Parminder Singh, Narinder Kumar

Guru Nanak Dev University, Amritsar, India

This article deals with the problem of simultaneously comparing several treatments with a control treatment under the assumption of unknown and unequal variances. The existing procedures for this problem are for the families of inferences in which all hypotheses are either of one-sided or of two-sided. For more general inferential families that contained a mixture of one- and two-sided inferences, a procedures is proposed to compare several treatments with a control treatment in one-way layout when variances are unknown and unequal.

Computation of the critical constants such that the proposed inference procedure meet optimal criteria are discussed and selected critical points and p-values are tabulated to facilitate the implementation of the proposed procedure. Power of the proposed procedure with a existing competitor procedure for the problem is discussed. Finally, an illustration of the procedure is made with a numerical data.

Multiple Comparison Procedure for Identifying The Minimum Effective Dose with More Than One Control

Rajvir Singh Chauhan, Narinder Kumar, Parminder Singh

Panjab University, Chandigarh-160014, India

In this paper, we develop a step-down procedure to find the Minimum Effective Dose (MINED) of a new drug when there are more than one control drugs. The computation of critical points required to implement the proposed procedure, is discussed by taking the normal probability model under equal sample size allocation. Power of the test is computed, and some power comparisons are made under different sample sizes.

Analysing Anti-malarial Trials: Using a False-claim Error Rate

Alice Parry, Thomas Jaki, Ian Hastings, Katherine Kay

Lancaster University, UK

The aims of a successful malaria treatment are primarily to cure the original infection and then secondly to prevent new infections. Although prevention of new infections is difficult/impossible it is nevertheless beneficial if the length of time between infections could be measured and increased. Consequently, it would therefore be advantageous to include the time that a patient is free from the parasites into the assessment of the efficacy of a treatment.

We have devised a multiple endpoint approach which incorporates the proportion cured along with the time to a new infection in the same analysis (i.e. a binary endpoint jointly with a survival endpoint), using score statistics, to give an overall assessment of the efficacy of the treatment.

The talk will discuss the initial set up of the conditional hypotheses and then explore an approach for finding the critical values and sample size using a 'false-claim' error rate as opposed to the traditional family-wise error rate. In particular, the 'false-claim' error rate looks at the probability of making claims rather than rejecting individual hypotheses. The principal consequence of using a 'claim-wise' error is that making two wrong rejections is penalized heavier than with the family-wise error rate. The motivation for using a 'false-claim' error rate will be discussed in detail.

A Note on Comparing Several Variances with a Control Variance

Anju Goyal

Panjab University, Chandigarh, India

Consider the problem of comparing variances of populations with the variance of a control population. When the experimenter has a prior expectation that the variances of populations are not less than the variance of a control population, one-sided simultaneous confidence intervals provide more inferential sensitivity than two-sided simultaneous confidence intervals. But the two-sided simultaneous confidence intervals have advantages over the one-sided simultaneous confidence intervals as they provide both lower and upper bounds for the parameters of interest. In this article, a new multiple comparison procedure is developed for comparing several variances with a control variance, which provides two-sided simultaneous confidence intervals for the ratios of variances with the control variance and maintains the inferential sensitivity of one-sided simultaneous confidence intervals. Computation of the critical constants for the implementation of the proposed procedure is discussed and the selected critical constants are tabulated.

The Influence of Shift-Variant data on Factor Analysis

Fumihiko Hashimoto

Graduate School of Economics, Osaka City University, Japan

The term "shift-variant" means that degradation and a noise change with positions on a X-Y plane.

This term is used in an optical image system.

Although the "factor analysis" often used in the multivariate analysis of social science assumes the quantity of the noise or other features in the position on each measurement axis to be "shift-invariant", but this assumption is not true.

This research investigate on that if we remove the assumption of "shift-invariant" assumption, (compare with ordinaly factor analysis), what difference occur between these two models.

However, introducing "shift-variant" model into the conventional factor analysis, actually it causes many mathematical restrictions.

Then, we propose that we would like to adopt genetic algorithm to solve "shift-variant" data on factor analysis.

Stochastic Modeling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa District

Mikiyas Gebresamuel Gebru

Ethiopia

The objectives of this thesis was to model seasonal variations in claim intensities and to evaluate the dependency of covariates on claim rates. The data for this thesis were obtained from claimants registered during September 2009 to August 2011, both inclusive at the Ethiopian Insurance Corporation in Hawassa. We present a procedure for consistent estimation of the claim frequency for motor vehicles in the Ethiopian Insurance Corporation, Hawassa District. The seasonal variation is modeled with a non-homogeneous Poisson process with a time varying intensity function. Covariates of the policy holders, like gender and age, is corrected for in the average claim rate by Poisson regression in a GLM setting. An approximate maximum likelihood criterion is used when estimating the model parameters. The seasonal parameters are found to be $\pm25\%$ and to be statistically significant. February has highest while August has lowest claim rate. Only age group 36-45 has significantly lower claim rate than age group 20-25. The rate is about one third. Lastly female is not found to have significantly lower claim rates than males, however, there are indications that maight be slightly so.

Using Early Outcome Data for Decision Making in Seamless Phase II/III Clinical Trials

Nigel Stallard

Warwick Medical School, University of Warwick, Coventry, CV4 7AL

Adaptive seamless phase II/III designs enable a clinical trial to be conducted in stages with the most promising of a number of experimental treatments selected on the basis of data observed in the first stage to continue along with the control treatment to the second and any subsequent stages. A number of methods have been proposed for the analysis of such trials to address the statistical challenge of ensuring control of the type I error rate. These methods rely on the independence the test statistics used in the different stages of the trial.

In some settings the primary endpoint can be observed only after long-term follow-up. In this case if short-term endpoint data are available, these could be used along with any long-term data at the interim analysis to inform treatment selection. The use of such data breaks the assumption of independence underlying existing analysis methods. This talk presents a new method that allows for the use of short-term data whilst controlling the overall type I error rate. Simulation study results will be presented showing that the use of the short-term endpoint data can lead to an increase in power when the short and long-term endpoints are correlated.

Complex Multiplicity Problems in Clinical Trials with Adaptive Sample Size Adjustment

Jeffrey Maca

Center for Statistics in Drug Development, Quintiles, Morrisville, NC 27560.

Multiplicity problems with several “sources” of multiplicity become increasingly more common in confirmatory Phase III clinical trials. Multiplicity may be induced by the analysis of multiple endpoints evaluated in several patient populations or at several dose levels. Additional sources of multiplicity may include simultaneous comparisons with a placebo and an active control as well as multiple types of treatment comparisons (non-inferiority and superiority). In this talk we will discuss multiplicity problems arising in confirmatory trials with multiple endpoints/dose-placebo comparisons and mid-course design changes. While standard methods are available for addressing multiplicity in a fixed-design setting (known as gatekeeping methods), very little is known about the performance of gatekeeping procedures in trial designs with adaptive elements. We will discuss methods that enable clinical trial sponsors to integrate gatekeeping procedures into adaptive sample size adjustment designs and ensure strong control of the overall Type I error rate in this complex multiplicity problems. A case study will be introduced to illustrate the statistical methodology and assessment of sample size adjustment designs with multiple objectives.

Enrichment Designs for the Development of Personalized Medicines

Martin Posch

Section for Medical Statistics, Center, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

In situations where the response to a treatment may depend on genetic biomarkers, it is important to identify biomarker based (sub-)populations, where the treatment has a positive benefit risk balance. One approach to identify relevant subpopulations is subgroup analyses where the treatment effect is estimated in biomarker positive and biomarker negative groups. Subgroup analysis are challenging because different types of risks are associated with inference on subgroups: On the one hand, ignoring a relevant subpopulation one could miss a treatment option due to a dilution of the treatment effect in the full population. Even, if the diluted treatment effect can be demonstrated in an overall population, it is not ethical to treat patients that do not benefit from the treatment, if they can be identified in advance. On the other hand selecting a spurious sub-population is not without risk either: it might increase the risk to approve a inefficient treatment (inflating the type I error rate), or may wrongly lead to restricting an efficient treatment to a too narrow fraction of a potential benefiting population. The latter can not only lead to reduced revenue from the drug, but is also unfavorable from a public health perspective.

We investigate these risks for non-adaptive study designs that allow for inference on subgroups using multiple testing procedures as well as adaptive designs, where subgroups may be selected in an interim analysis. Quantifying the risks with utility functions the characteristics of such adaptive and non-adaptive designs are compared for a range of scenarios.

Geometrical Representation and Classification of Closed Consonant Weighting Schemes and Associated Multiple Tests

Willi Maurer

Novartis Pharma AG, Basel

Hommel et al. (2007) showed that sequentially rejective weighted Bonferroni tests for m logically independent hypotheses form a class B of closed consonant procedures defined by weighted Bonferroni tests on the resulting 2m-1 intersection hypotheses. The underlying weighting schemes then obey a monotonicity criterion. "Classical" procedures such as the weighted Bonferroni-Holm test, fixed sequence and fallback tests, as well as more recent proposals are subsets of this class. Among them are the truncated Bonferroni-Holm tests (Dmitrienko et al., 2008), the graphical procedures of Bretz et al. (2009) and of Burman et al. (2009).

Weighting schemes on the intersection hypotheses can be represented as vectors in m2(m-1) - dimensional Euclidian space. The set of vectors representing procedures from class B is finite and closed under convex combinations and spans a lower dimensional polytope embedded in this space. A point of this polytope B represents a class of equivalent multiple test procedures. Results on the relationship between different classes of procedures are presented. For non-convex classes - like the graphical procedure of Bretz et al.- the convex combination of its members to "entangled" graphs allows to generate sequentially rejective procedures that have properties the single members do not necessarily have (Maurer and Bretz, 2012). Results from polytope theory can be used to gain new insights into the properties of and connections between the respective test procedures. For example one can show that strictly hierarchical (fixed sequence) test procedures are vertices (extreme points) of the polytope B and that the default graphs defined by Burman et al. and the entangled graphs are points in the sub-polytope spanned by these vertices. Other results are, e.g., that the Bonferroni-Holm Procedure is the center of polytope B and any point of the sub-polytope can be represented as convex combination of at most d+1 hierarchical procedures, where $d \leq m!$ is the dimension of the sub-polytope. Recently described connections to statistical

ranking (Sturmfels and Welker, 2012) will be shortly touched. In general, the geometrical representation allows better understanding and comparing different ways to describe and represent equivalent procedures, their advantages and disadvantages in terms of the number of free parameters needed to describe them and the ease of understanding its “practical” properties.

Bretz, F, Maurer, W, Brannath, W, and Posch, M. (2009) A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28: 586-604.

Burman C.-F., Sonesson C., and Guilbaud O. (2009) A recycling framework for the construction of Bonferroni-based multiple tests. Statistics in Medicine, 28: 739-761.

Dmitrienko A, Tamhane A and Wiens B. (2008) General multi-stage gatekeeping procedures. Biometrical Journal; 50:667–677.

Hommel, G, Bretz, F and Maurer, W. (2007) Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. Statistics in Medicine, 26: 4063–4073.

Maurer, W and Bretz, F. (2012) Memory and other properties of multiple test procedures generated by entangled graphs. Statistics in Medicine, Publ. online DOI: 10.1002/sim.5711.

Sturmfels, B and Welker, V (2012) Commutative algebra of statistical ranking, Journal of Algebra 361: 264–286.

Weighting and Ordering Considerations for Multiple Testing Procedures in Clinical Trials

Brian L. Wiens, Alex Dmitrienko, Olga Marchenko

Alcon Research, Ltd., USA

We discuss the problem of selecting parameters for multiple testing procedures in confirmatory Phase III clinical trials , especially hypothesis weights and hypothesis ordering. We identify classes of multiple testing procedures that provide different interpretations of these parameters. The classes include basic single-step procedures that employ fixed hypothesis weights as well as more powerful multistep procedures that adaptively re-weight the hypotheses during the testing process. We examine the behavior of different classes of multiple testing procedures in problems with weighted hypotheses and a priori ordered hypotheses and provide practical guidelines for the choice of hypothesis weights and hypothesis ordering. Notably, a stepwise procedure with symmetric re-weighting is invariant to ordering of the tests. The concepts discussed in the paper are illustrated using case studies based on clinical trials with multiple endpoints.

Mixed Directional False Discovery Rate Control in Multiple Pairwise Comparisons Using Weighted P-values

Haibing Zhao, Xinping Cui, Shyamal Peddadac

University of California, Riverside, USA

In many applications researchers are interested in making pairwise comparisons among k test groups on the basis of m outcome variables. Often m is very large. For example, such situations arise in gene expression microarray studies involving several experimental groups. Researchers are often not only interested in identifying differentially expressed genes between a given pair of experimental groups but are also interested in making directional inferences such as whether a gene is up or down regulated in one treatment group relative to another. In such situations, in addition to the usual errors such as false positive (Type I error) and false negative (Type II error), one may commit directional error (Type III error). For example, in a dose response microarray study, a gene may be declared to be up-regulated in the high dose group compared to the low dose group when it is not. In this paper we introduce a mixed directional false discovery rate (mdFDR) controlling procedure by weighting the raw p-values to reflect the proportion of up or down regulated genes in each pairwise comparison. Performance of the proposed methodology is evaluated theoretically. Empirical comparisons reveal that the proposed methodology is at least as powerful as the mdFDR controlling method of Guo et al. (2010).

Multiple Testing Method for The Directed Acyclic Graph, Using Shaffer Combinations

Rosa Meijer, Jelle Goeman

Leiden University Medical Center, Netherlands

We present a novel multiple testing method for testing null-hypotheses that correspond to nodes in a directed acyclic graph (DAG). Such DAG-structured multiple testing problems can be encountered in various settings. One well-known example is the problem of performing a gene-set analysis, in which multiple gene-sets and individual genes are tested on their association with a clinical outcome. Each hypothesis about a gene or gene-set can be considered a node in a DAG in which the edges correspond to subset-relationships between the nodes. The gene ontology (GO) graph is a specific example of such a DAG.

Although several multiple testing methods have been developed for gene-set analysis, the novelty of our method is that it uses the specific DAG structure to make statements on the possibility of certain configurations of true and false null hypotheses. By constructing/extending the DAG in such a way that every node is the intersection of its child-nodes, it will often happen that the logical relationships between the hypotheses will create restricted combinations, which means that not all remaining hypotheses can simultaneously be true. Using this information can reduce the multiple testing burden. Our method can be seen as an extension of Meinshausen's familywise error rate controlling procedure for tree-structured hypotheses.

Implementing our method requires repeated solutions of instances of the minimum hitting set problem, which is known to be an NP-hard problem. Depending on the size of the DAG, we either calculate these solutions exactly by using an ILP (integer linear programming) solver or we use approximations based on a greedy algorithm.

The method will be illustrated by testing Gene Ontology terms for evidence of differential expression in a survival setting.

Simultaneous Confidence Bands for Polynomial Regression Curves with the Volume-of-Tube Formula

Satoshi Kuriki

The Institute of Statistical Mathematics, Japan

A polynomial that is nonnegative over a given interval is called a positive polynomial. In Kato and Kuriki (2013), we considered the likelihood ratio test for the hypothesis of positivity that the estimand polynomial regression curve is a positive polynomial. The null distribution is obtained as a mixture of chi-square distributions via the one-sided volume-of-tube formula. In this talk, we propose associated simultaneous confidence bands for polynomial regression curves, and examine the one-sided and two-sided improved confidence bands by comparing with the conventional ones. A growth curve rate data is examined as a numerical example. We also mention the Fourier regression. Its confidence bands can be constructed in a similar way, but geometric invariants needed in the volume-of-tube formula are obtained more explicitly.

Simultaneous Inference for Low Dose Risk Estimation with Quantal Data in Benchmark Analysis

Jianan Peng

Acadia University, Canada

Risk assessment studies where human, animal or ecological data are used to set safe low dose levels of a toxic agent is challenging as study information is limited to high dose levels of the agent. Al-Saidy et al.(Biometrics 59:1056-1062, 2003) developed the hyperbolic band for low dose inference with quantal response data. However, the shape of the constant width band may be more desirable than that of the hyperbolic band in some applications in risk analysis as a confidence band should be narrower especially near the lower end of low doses. In this talk, we will show that a constant width band should do better than the hyperbolic band considered by Al-Saidy et al. (2003).

Simultaneous Confidence Bands for a Percentile Line in Linear Regression with Application to Drug Stability Studies

Yang Han, Wei Liu, Frank Bretz, Fang Wan

S3RI and School of Mathematics, University of Southampton, UK

Simultaneous confidence bands have been used to quantify unknown functions in various statistical problems. A common statistical problem is to make inference about a percentile line in linear regression. Construction of simultaneous confidence bands for a percentile line has been considered by several authors, e.g., Steinhorst and Bowden (1971), Turner and Bowden (1977, 1979) and Thomas and Thomas (1986). But only conservative symmetric bands, which use critical constants over the whole covariate range (-∞, ∞), are available in the literature. Methods given in this paper allow the construction of exact simultaneous confidence bands for a percentile line over a finite interval. Furthermore, we propose a method of constructing an asymmetric confidence band corresponding to each given symmetric confidence band. Comparison under the average band width criterion shows that the exact symmetric bands can be substantially narrower than the corresponding conservative symmetric bands available in the literature so far. Moreover, we find that asymmetric confidence bands are uniformly and can be very substantially narrower than the corresponding exact symmetric bands. Therefore, asymmetric bands should always be used under the average band width criterion. We illustrate the proposed methods with a real example on drug stability study.

1. Steinhorst, R.K., Bowden, D.C. (1971). Discrimination and confidence bands on percentiles. Journal of the American Statistical Association, 66, 851-854.
2. Thomas, D.L. and Thomas, D.R. (1986). Confidence bands for percentiles in the linear regression model. Journal of the American Statistical Association, 81, 705-708.
3. Turner, D.L. and Bowden, D.C. (1977). Simultaneous confidence bands for percentile lines in the general linear model. Journal of the American Statistical Association, 72, 886-889.
4. Turner, D.L. and Bowden, D.C. (1979). Sharp confidence bands for percentile lines and tolerance bands for the simple linear model. Journal of the American Statistical Association, 74, 885-888.

Comparisons of Simultaneous Confidence Bands for Linear Regression with Interval Constraint on Predictor Variables

Shan Lin

School of Mathematics and Statistics Northeast Normal University, China

This work compares several key methods of constructing simultaneous confidence bands for a linear regression model with each predictor variable constrained in an interval. These methods include the conservative method of Naiman (1986), the approximate method of Sun and Loader (1994) and the simulation-based method of Liu et al. (2005). They are compared in terms of the critical values under various designs. It is found that the conservative band becomes exact in simple linear regression case, the simulation-based band is clearly narrower than the approximate band though not uniformly, but the better of the latter two has critical values almost as good as the exact band.

Group Sequential Designs: Theory, Computation and Optimisation

Chris Jennison

University of Bath, UK

It is standard practice to monitor clinical trials with a view to stopping early if results are sufficiently positive, or negative, at an interim stage. We shall explain how properties of stopping boundaries can be calculated and how boundaries can be optimised to minimise expected sample size while controlling type I and II error probabilities.

Although constraints on error probabilities complicate this optimisation problem, a solution is possible through unconstrained Bayes decision problems which are conveniently solved by dynamic programming. We shall present details of numerical computation for group sequential tests and their optimisation for particular criteria. We shall discuss applications in a variety of settings, including the derivation of optimal adaptive designs in which future group sizes are allowed to depend on previously observed responses; designs which test both for superiority and non-inferiority; and group sequential tests which allow for a delay between treatment and response.

Optimal Design for Multi-Arm Multi-Stage Clinical Trials

James Wason, Thomas Jaki

MRC Biostatistics Unit, Cambridge, CB2 0SR, UK

In early stages of drug development there is often uncertainty about the most promising among a set of different treatments. In order to ensure the best use of resources it is important to decide which, if any, of the treatments should be taken forward for further testing. Multi-arm multi-stage (MAMS) trials provide gains in efficiency over separate randomised trials of each treatment. They allow a shared control group, dropping of ineffective treatments before the end of the trial and stopping the trial early if sufficient evidence of a treatment being superior to control is found.

In this talk I discuss optimal MAMS designs for normally distributed endpoints. An optimal design has the required type-I error rate and power, but minimises the expected sample size (ESS) at some combination of treatment effects. Finding an optimal design requires searching over the stopping boundaries and sample size per stage, potentially a large number of parameters. We propose a method which combines quick evaluation of specific designs and an efficient stochastic search for the optimal design parameters. The search can also take the allocation ratio between controls and active treatments into account, allowing further efficiency gains.

In the two-arm case, the triangular design has good expected sample size properties, and is immediate to find. Here we find that there is potential for greater improvements over the triangular design, especially as the number of stages or active treatments increases. The triangular design still serves as a quick-to-find and near-optimal design, so may still be useful for design of MAMS trials.

Designing Multi-Arm Multi-Stage Clinical Trials with a Safety and an Efficacy Endpoint

Thomas Jaki

Medical and Pharmaceutical Statistics Research Unit Department of Mathematics and Statistics Lancaster University, UK

Multi-arm clinical trials that compare several active treatments to a common control have been proposed as an efficient means of making an informed decision about which of several treatments should be evaluated further in a confirmatory study. Additional efficiency is gained by including interim analyses and in particular seamless Phase II/III designs have been the focus of recent research. Common to recent work is the constraint that selection and formal testing should be based on a single efficacy endpoint, despite the fact that in practice, safety considerations will often play a central role in determining selection decisions. In this talk we develop a multi-arm, multi-stage design for a trial with an efficacy and safety endpoint. The design extends group-sequential ideas and considers the situation where a minimal safety requirement is to be fulfilled and the treatment yielding the best combined safety and efficacy trade-off satisfying this constraint is selected for further testing. The treatment with the best trade-off is selected at the first interim analysis while the whole trial is allowed to comprise of J analyses. We show that the design controls the family-wise error rate in the strong sense and illustrate the method through an example and simulation. We find that the design is robust to miss-specification of the correlation between the endpoints and requires a similar number of subjects to a trial based on efficacy alone.

Calibration of P-values via the Dirichlet Process

Mikelis Guntars Bickis

Department of Mathematics and Statistics, University of Saskatchewan, Canada

In testing a simple statistical hypothesis, the P-value is defined as the probability, assuming the null hypothesis, of the most extreme event that actually happened. It is commonly described as quantifying the amount of evidence against the null hypothesis, although this interpretation has been disputed by Berger and Sellke. Non-statisticians frequently misinterpret the P-value as the (posterior) probability of the null hypothesis, and even those with statistical training sometimes confuse it with probability of type I error. Indeed, although P-values are ubiquitous in applied statistics, they play a role in neither Bayesian nor Neymanian theories of inference.

In 2001, Sellke, Bayarri, and Berger proposed a calibration of P-values whereby they could be given a Bayesian interpretation. In the case of multiple P-values, Efron proposed in 2005 the local false discovery rate as the (estimated) posterior probability of the null hypothesis given the P-value.

When one has a large number of P-values from related hypotheses, their empirical distribution can be used to make inferences about the proportion of true null hypotheses. Under certain regularity conditions, the posterior probability of the null hypotheses can be calculated as a ratio of slopes of the actual distribution. To obtain a smooth estimate of this distribution, the P-values can be modelled as arising from normally-distributed test statistics in which the location parameter itself has an underlying distribution, consisting of an atom at zero mixed with a distribution of alternatives. The prior of this distribution of alternatives is modelled as a Dirichlet process. The posterior mean of the distribution of alternatives is then used to calibrate the P-values as posterior probabilities.

Adjusted p-values for SGoF Multitesting Procedure. Definition and Properties

Irene Castro Conde, Jacobo De Uña-Álvarez

University of Vigo, Spain

In the paper Carvajal-Rodríguez et al. (2009) a new multitest correction named SGoF (from Sequential Goodness-of-Fit) was introduced; this method was extended to possibly correlated tests in de Uña-Álvarez (2012), who introduced the Beta-Binomial SGoF (BB-SGoF) procedure. Both SGoF and BB-SGoF have the property of increasing their statistical power when increasing the number of tests, which is very useful in omic sciences: genomics, proteomics, etc. because they typically involve the simultaneous testing of hundreds or thousands of hypotheses. Statistical properties and false discovery rate and power levels in practical settings for SGoF-type strategies were further investigated in de Uña-Álvarez (2011, 2012) and Castro-Conde and de Uña-Álvarez (2013). In this talk we introduce adjusted p-values for SGoF method and we investigate their properties. Time permitting, adjusted p-values for BB-SGoF will be presented and discussed too.

Carvajal-Rodríguez A, de Uña- Álvarez J and Rolán- Álvarez E (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. BMC Bioinformatics 10:209.

Castro-Conde I y de Uña-Álvarez J (2013). Performance of Beta-Binomial SGoF multitesting method for dependent gene expression levels: a simulation study. Proceedings of BIOINFORMATICS 2013 International Conference on Bioinformatics Models, Methods and Algorithms (Pedro Fernandes, Jordi Sole-Casals, Ana Fred and Hugo Gamboa Eds.), SciTePress.

de Uña-Álvarez J (2011). On the statistical properties of SGoF multitesting method. Statistical Applications in Genetics and Molecular Biology Vol. 10, Iss. 1, Article 18.

de Uña-Álvarez J (2012). The Beta-Binomial SGoF method for multiple dependent tests. Statistical Applications in Genetics and Molecular Biology Vol. 11, Iss. 3, Article 14.

New Multiple Testing Method under no Dependency Assumption, with Application to Multiple Comparisons Problem

Li Wang, Xingzhong Xu, Yong A

Beijing Institute of Technology, China

Nowadays, multiple hypotheses testing mainly focuses on constructing stepwise procedures under some error rate control, such as Familywise Error Rate (FWER), False Discovery Rate, and so forth. However, most of these procedures are obtained in independent case, and when there is correlation across tests, the dependency may increase or decrease the chance of false rejections. In this talk, a totally different testing method is proposed, which doesn't focus on specific error control, but pays attention to the overall performance of the collection of hypotheses and the structure utilization among hypotheses. Since the main purpose of multiple testing is to pick out the false ones from the whole hypotheses, and present a rejection set, motivated by the principle of simple hypothesis testing, we give the final testing result based on the estimation of the set of all the true null hypotheses I_0. We intend to find the "largest" set J_0, with the p-value $p(J_0)$ for the intersection hypothesis H_{J_0} larger than α, as an estimator of I_0. Our method can be applied in any dependent case provided that a reasonable p-value can be obtained for each intersection hypothesis. We illustrate the new procedures with application to multiple comparisons problems. Theoretical results show the consistency of our method, and investigate their FWER behavior. Simulations suggest that our procedures have a better overall performance compared with some existing procedures in dependent case, especially in the total number of type I and type II errors. What's more, the rank of the rejected (or non-rejected) hypotheses sequence is different from that based on marginal-distributed p-values.

A Sufficient Criterion for Control of Generalised Error Rates in Multiple Testing

Sebastian Doehler

Darmstadt University of Applied Sciences

Based on the work of Romano and Shaikh (2006 a, b) and Lehmann and Romano (2005} we give a sufficient criterion for controlling generalised error rates for arbitrarily dependent p-values. This criterion is formulated in terms of matrices associated with the corresponding error rates and thus it is possible to view the corresponding critical constants as solutions of sets of certain linear inequalities. This property can in some cases be used to improve the power of existing procedures by finding optimal solutions to an associated linear programming problem.

Romano and Shaikh (2006a). On stepdown control of the false discovery proportion. In Optimality. The second Erich L. Lehmann symposium.
Romano and Shaikh (2006b). Stepup procedures for control of generalizations of the familywise error rate.Ann. Stat. 34(4), 1850-1873.
Lehmann and Romano (2005). Generalizations of the familywise error rate. Ann. Stat. 33(3), 1138-1154.

Valid Post-Selection Inference

Andreas Buja, Richard Berk, Larry Brown, Kai Zhang, Linda Zhao

The Wharton School, University of Pennsylvania, USA

It is common practice in statistical data analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid ``post-selection inference'' by reducing the problem to one of simultaneous inference and hence suitably widening conventional confidence and retention intervals. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing ``simultaneity insurance'' for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffe protection. Importantly it does NOT depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

Selection Adjusted Confidence Intervals with More Power to Determine the Sign

Asaf Weinstein, William Fithian, Yoav Benjamini

University of Pennsylvania, USA

In problems involving multiple parameters, it is common to construct confidence intervals for only those parameters highlighted as interesting by a selection procedure that depends on the observed data. As such, using a marginal CI for each of the selected parameters no longer offers the right coverage level, not even on the average over the constructed intervals, thus calling for some selection adjustment.

We address this problem by developing a CI for the location parameter of a symmetric, unimodal distribution conditional on the estimator being bigger than some constant threshold. Our CI offers early sign determination, that is, it avoids including parameters of both signs for relatively small values of the estimator, and is obtained through inverting a family of tests designed to endow the interval with the property above. The CI is not of a constant shape as a function of the observation, but it converges to the usual, symmetric CI when the observation is big, and the computation is implemented in available software.

In a multiple parameter setting where the FCR of Benjamini and Yekutieli is the error rate to be controlled, constructing the conditional CI for each selected parameter is trivially valid when selection is based on passing a constant threshold; We show that for a wide class of selection procedures that involve a data-dependent (random) threshold, including the procedure of Benjamini and Hochberg, the FCR is still controlled when using the conditional CI, if the estimators are independent. Furthermore, in this case the FCR using our method is not only bounded from above, but in fact is essentially equal to, the nominal level q, provided that with high probability at least one parameter is selected. Using the conditional intervals is therefore an appealing alternative to the FCR-adjusted intervals of Benjamini and Yekutieli, that only guarantee $FCR \geq q/2$.

Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression

Ulrike Schneider

Vienna University of Technology

We study the distribution of hard-, soft-, and adaptive soft-thresholding estimators within a linear regression model where the number of parameters k can depend on sample size n and may diverge with n. In addition to the case of known error-variance, we define and study versions of the estimators when the error-variance is unknown. We derive the finite-sample distribution of each estimator and study its behavior in the large-sample limit, also investigating the effects of having to estimate the variance when the degrees of freedom n-k does not tend to infinity or tends to infinity very slowly. Our analysis encompasses both the case where the estimators are tuned to perform consistent model selection and the case where the estimators are tuned to perform conservative model selection. We discuss consistency, uniform consistency and derive the uniform convergence rate under either type of tuning.

(This is joint work with Benedikt Pötscher).

Bayesian Subgroup Analysis

James Berger

Department of Statistical Science, Duke University, Durham, NC 27708

Subgroup analysis raises difficult issues of multiplicity adjustment, since the number of subgroups that can be separately tested can be quite large and the test statistics for different subgroups are typically highly dependent. The Bayesian approach to multiplicity adjustment in general, and hence to subgroup analysis, is to correct for the multiple testing through appropriate choice of the prior probabilities of the models that are implied by the subgroup analysis. Choice of these prior probabilities and methods of implementation of the ensuing Bayesian analysis will be discussed in this presentation with applications to vaccine trials and biomarker identification.

Overview of Subgroup Identification Approaches in Clinical Research

Ilya Lipkovich

Center for Statistics in Drug Development, Quintiles, Morrisville, NC 27560

Vast literature has been generated in medical and statistical journals over the last 15 years concerning the subgroup analysis methodology and assessment of validity/credibility of subgroup analysis methods for clinical trial data. We see a shift of emphasis from presenting "good practices" of subgroup analysis to developing aggressive subgroup identification strategies under the umbrella of "individualized medicine/tailored therapeutics". This presentation will provide an overview of key subgroup identification methods that often originated in data mining and machine learning fields and present a novel subgroup identification procedure (SIDES procedure). This procedure is an ensemble method based on recursive partitioning. The SIDES procedure will be illustrated using several clinical trial examples.

Region, Biomarker Subset or Patient Subpopulation: Are They Multiplicity Problem and When?

Sue-Jane Wang

Office of Biostatistics, FDA, Silver Spring, MD 20993

The intended patient population needs to be defined in the to-be-marketed medical product label once substantial and persuasive evidence of a safe and effective new treatment can be established statistically and clinically. In order to achieve this integrated study aim, this presentation attempts to define a statistical framework for the design and analysis of subgroups defined by biomarker subset(s) or specific patient subpopulation(s), recognizing the multiplicity for a clinical development program. Often such trials enroll patients from multi-regions geographically, known as global trials. Aside from methodological challenges in multi-regional clinical trials, newer concepts and methods for design and analysis approaches to subgroup evaluation will be presented. In addition, regulatory science considerations will also be articulated via case studies to illustrate the complex regulatory decision making potentially beyond the well-known multiplicity issues.

Directional Error Rates of Closed Testing Procedures

Peter Westfall

Texas Tech University, USA

Closed multiple testing procedures are common in biopharmaceutical protocols. Whether their directional error rates are controlled is largely an open problem. In this article, directional error rates are investigated using analytical, numerical, and Monte Carlo methods. A Monte Carlo variance reduction method amenable to this purpose is presented. A factorial design is used to identify possible problem areas, and directional error rates are simulated. No cases of excess directional error are found for typical applications involving noncentral multivariateT distributions. However, directional error rates in excess of the nominal are found when using regression function tests with nearly collinear linear combinations, both for one-sided and two-sided tests.

A Unifying Approach to the Shape and Change-Point Hypotheses in the Univariate Exponential Family

Chihiro Hirotsu

Collaborative Research Center, Meisei University, Japan

Usually a change-point model assumes a step-type change at some point of time series. Then as an interesting thing it has been shown that the max acc. t test for the isotonic hypothesis is appropriate also for detecting a change-point. It comes from a relationship that each corner vector of the polyhedral cone defined by the isotonic hypothesis corresponds to a component of the change- point model. Actually max acc. t is the maximal component of the projections of the observation vector on to the corner vectors of the polyhedral cone. The relationship has been extended in the case of normal model to the convexity and the slope change hypotheses. In this talk we extend the idea further to a univariate exponential family. So unifying implies not only to unify the shape and change-point hypotheses but also to unify several distributions in the univariate exponential family. There are a lot of work on the isotonic and step-type change-point hypotheses. So in this talk we mainly deal with the convexity and the slope change hypotheses. We propose a maximal contrast statistic based on the doubly accumulated statistics, whereas the isotonic inference is based on the usual cumulative sum statistics. Use of the doubly accumulated statistics is rather novel, whereas the cumulative sum is very popular. A second order Markov property of the serial component statistics is shown and based on it a very efficient and exact algorithm for the p-value calculation is obtained.

On the Moderated t-test and its Moderated p-values

Jelle Goeman

Leiden University Medical Center, Netherlands

Moderated t-tests such as limma are very popular for the analysis of high-dimensional genomics data. Such tests improve the variance estimate of each individual probe by using additional information from the other probes using empirical Bayes arguments. By “borrowing strength” in this way the moderated t-test is better able to separate true from false hypotheses than the classical t-test, especially in data in which the number of probes is very large and the sample size is very small. But there is a price to pay for using the moderated t-test. We argue that p-values resulting from a moderated t-test lose an important property that classical p-values do have. Whereas p-values from a classical t-test are uniformly distributed if the null hypothesis is true, the p-values from a moderated t-test are uniform only in some average sense. The lack of this uniformity property is a serious problem if we want to adjust these p-values for multiple testing. We give practical examples of problematic situations arising when multiple testing methods are combined with the moderated t-test. Most notably, we show that the frequently-used combination of false discovery rate control and a lower bound on the estimated effect size (“fold change”) can result in excessive error rates.

Pairwise Comparisons of Treatments with Ordered Categorical Responses

Yueqiong Lin, Siu Hung Cheung, Wai-Yin Poon, Tong-Yu Lu

School of Management, Fuzhou University, Fuzhou, China

Many clinical trials involve the analysis of ordered categorical responses. The Wilcoxon-Mann-Whitney test (WMW) has been a popular choice for the comparison of two treatments with ordered categorical data. For pairwise comparisons with more than two treatments, the modified WMW test based on a logistic regression model with the proportional odds assumption can be applied. In this project, we consider another approach that employs a normal latent variable model. The major benefit of using our proposed method is that it is more robust with respect to type I error control in the presence of heterogeneous treatment variances. Examples will be given to illustrate the implementation of our procedure.

Powerful Mixture-Based Gatekeeping Procedures in Clinical Trials

Alex Dmitrienko, George Kordzakhia

Quintiles, USA

A general mixture method for constructing gatekeeping procedures for complex multiplicity problems in clinical trials was developed in Dmitrienko and Tamhane (2011, 2013). The general method accommodates a very broad class of logical relationships among the hypotheses of interest, including combinations of serial and parallel logical restrictions. We show in this talk that the decision rules used in mixture-based gatekeeping procedures can be streamlined in certain special cases that are common in confirmatory Phase III clinical trials. Examples include "two-dimensional" multiplicity problems with serial restrictions frequently arising in Phase III clinical trials designed to evaluate the efficacy profile of new treatments using ordered multiple endpoints at several dose levels. The modified approach to defining gatekeeping procedures streamlines their implementation and leads to a power gain compared to the standard mixture-based approach. The new approach will be illustrated using a Phase III clinical trial.

Cyclic Stack Procedures with Parallel Gatekeeping

George Kordzakhia, Alex Dmitrienko

Food and Drug Administration, USA

This talk introduces a method for constructing multistage parallel gatekeeping procedures with retesting options. The cyclic stack procedures are designed for testing families of hypotheses that are linearly ordered so that the higher ranked families serve as a parallel gatekeeper to the lower ranked families.

The testing proceeds through the ordered list of families from the top of the list to the bottom and then returns directly to the top. If a family is fully rejected, it is removed from the list. The cyclic stack procedures allow multiple repeated retesting of the families with increasingly more powerful tests. This approach serves as an extension of multistage parallel gatekeeping procedures (Dmitrienko, Tamhane, and Wiens (2008)).

Gatekeeping Procedures for Multiple Correlated Endpoints Including Responder Endpoints

Yu-Ping Li, Alan Hopkins, Jin-Sying Lin

Theravance, Inc., USA

Most human diseases are characterized by multiple measures, including quantitative measurements, signs, symptoms, and patient-reported outcomes (PROs); hence, the study endpoints selected for evaluation of treatment effectiveness can be highly correlated with each other. Commonly, the more rigorous responder endpoints can be derived from the continuous endpoints.

Gatekeeping strategies for hierarchically ordered hypotheses provide clinical trial researchers with useful tools for managing multiplicity in clinical trials that have multiple correlated endpoints. A simulation study was conducted to investigate the overall power of a prospective phase III trial with continuous and responder endpoints using parallel gatekeeping procedures to control the overall Type I error rate. 1000 simulated datasets, each with 12 weeks of data for 300 subjects per group (active vs. placebo), were generated based on effect sizes observed from a phase II study. Treatment comparisons were conducted for a total of 9 correlated endpoints. The null hypotheses of these endpoints were grouped into multiple families for gatekeeping multiplicity adjustment. The hierarchical order of hypothesis families was initially chosen based on the clinical importance of the endpoints and the need to meet regulatory requirement for approval and the results observed from prior phase 2 studies. Various re-structuring of hypotheses and different gamma values for the truncated Holm and truncated Hochberg multiple test procedures were investigated. The power for each hypothesis family and for each individual endpoint was estimated for each scenario simulated.

The simulation study demonstrated that combining the primary efficacy endpoint with one key secondary endpoint into the first hypothesis family provided a power advantage for the subsequent families over the scenario of forming the above two endpoints into two separate hypothesis families. The truncated Holm and truncated Hochberg multiple test procedures yielded similar results. The gatekeeping procedures described here provided clinical trial researchers with useful tools for managing multiplicity in clinical trials

that have correlated endpoints including responder endpoints, especially in the case when the responder endpoints were derived from the continuous, longitudinal type endpoints.

Complex Multiple Comparison Problems When Multiple Trials are Evaluated

H.M. James Hung

US Food and Drug Administration, USA

In the setting that multiple clinical trials are considered for decision-making on a drug product, multiple comparison problems can be very challenging. The current paradigm that stipulates control of the overall type I error for each individual trial may not be sufficient to handle such challenging problems. One problem concerns assessment of benefit versus risk, where an endpoint such as mortality can serve as an efficacy endpoint and a safety endpoint. Another problem concerns use of intermediate endpoints as a basis of approval for pre-marketing and then followed by confirmatory evidence for assessing a clinical endpoint. In this talk, I shall share the challenging problems and shed some light on possible solutions.

Consistency-Adjusted Alpha Allocation Methods for Composite Endpoints

Geraldine Rauch, Meinhard Kieser

University of Heidelberg, Heidelberg, Germany

Composite endpoints are often used as primary efficacy endpoints, particularly in the field of oncology and cardiology. These endpoints combine several events of interest within a single variable. Thereby, it is intended to enlarge the expected effect size and thus to increase the power of the clinical trial. However, the interpretation of composite endpoints can be difficult, as the observed effect for the composite does not necessarily reflect the effects of the single components. Therefore, it might not be adequate to judge the efficacy of the new intervention exclusively on the composite effect. Including the most relevant component effects in an efficacy claim assessed by a confirmatory test strategy could overcome this problem, however imposes the problem of multiplicity. Moreover, to show superiority of the new intervention with respect to single components is usually not realistic in these settings as the expected individual effects are small.

Alosh and Huque (2009,2012) and Li et al. (2012) recently proposed consistency-adjusted alpha allocation methods which can be used and extended to address this particular problem. We discuss several alpha allocation methods for composite endpoints, compare their power properties and apply the methods to several clinical trial examples. Moreover, we face the general problem of correlation-adjusted adjustment procedures taking into account the special correlation structure between composite endpoints and their components.

Alosh M, Huque MF (2009): A flexible strategy for testing subgroups and overall population. Stat Med 28:3-23.

Huque MF, Alosh M (2012): A consistency-adjusted strategy for accomodating and underpowered primary endpoint. J Biopharm Stat 22: 160-179.

Li H, Sankoh AJ, D'Agostino Sr RB (2012): Extension of adaptive alpha allocation methods for strong control of the family-wise error rate. Stat Med, Epub ahead of print.

Sample Size Considerations in Complex Clinical Trials

Toshimitsu Hamasaki, Tomoyuki Sugimoto, Takashi Sozu, Scott R Evans

Osaka University Graduate School of Medicine, Japan

The determination of sample size and the evaluation of power are critical elements in the design of a clinical trial. If a sample size is too small then important effects may not be detected, while a sample size that is too large is wasteful of resources and unethically puts more participants at risk than necessary.

In last decade, confirmatory clinical trials with multiple objectives have become increasingly common in many disease areas. Such trials include, for example, the hierarchical investigation of multiple doses or regimens of a new treatment, evaluation of two or more clinical primary and secondary endpoints, evaluation of several populations, the switching of assessment from non-inferiority to superiority, or any combination thereof.

Having such multiple objectives provides the opportunity to characterize an intervention's multidimensional effects, but also creates challenges in handling multiplicity particularly with respect to the evaluation of power and the calculation of sample size during the design of clinical trials.

The focus of this presentation is sample size issues in clinical trials with multiple primary endpoints. When more than one endpoint is viewed as important, trials can be designed with the aim of either: T1) evaluation of the significance on all of the endpoints, or T2) evaluation of the significance on at least one endpoint with a prespecified ordering of objectives. When the aim is T1, no adjustment is needed to control the type I error rate. The hypothesis associated with each endpoint should be evaluated at the same significance level as is required for all of the objectives. However, the type II error rate increases as the number of endpoints being evaluated increases. In contrast, when the aim is T2, an adjustment is needed to control the Type I error rate. We first discuss the simplest case of a superiority trial comparing two interventions with respect to continuous endpoints, without an interim analyses. We then discuss the other endpoint scales, i.e., binary and time-to-events. We briefly mention extensions to include interim analyses.

Thresholding of a Companion Diagnostic Test Confident of Efficacy in Targeted Population

Jason C. Hsu

The Ohio State University, USA

In personalized medicine, continuous biomarker values are often dichotomized to classify patients into target and non-target populations. Cast in the setting of normally distributed responses that are modeled linearly (such as diabetes and psychiatry), we provide a method of inferring which thresholds correspond to target populations that benefit from the treatment. By providing simultaneous confidence intervals for efficacy corresponding to all candidate thresholds, our method allows for flexible decision-making, taking into consideration marketing potential based on both the size of the target population and efficacy in the target population. Under the assumption of the general linear model (GLM), advantages of our approach over the Jiang, Freidlin, Simon (2007) approach are that (1) formulation is clinically meaningful, (2) imbalance in the data would not lead to misleading inference, (3) simultaneous confidence intervals are provided, (4) error rate and confidence level computations are precise, (5) simple SAS codes are available.

Higher Criticism Test Statistics: Why Is The Asymptotics So Poor?

Veronika Gontscharuk, Sandra Landwehr, Helmut Finner

Department of Statistics in Medicine, Faculty of Medicine, Heinrich-Heine-University Duesseldorf, Germany

This presentation focuses on the so-called higher criticism (HC) test statistics, which can be seen as normalized Kolmogorov-Smirnov statistics. The limiting distribution of HC statistics was already investigated in the late 1970s, cf. Eicker (1979) and Jaeschke (1979). Unfortunately, the convergence of the distribution of the (standardized) HC statistic to the limiting (Gumbel) distribution is extremely slow so that nice asymptotic results related to HC statistics can hardly be used for a nite sample. Perhaps this is why HC statistics seem to have fallen into oblivion for a long time. Nevertheless, about ten years ago Donoho and Jin (2004) brought HC test procedures back to life. They showed that HC tests are "capable of optimally detecting the presence of signals that are so weak and so sparse that the signal cannot be consistently estimated". A series of further results on the HC concept were published in the following years, e.g., cf. Jager and Wellner (2007), Hall and Jin (2008, 2010). However, a number of open questions arising in the HC context remain, some of which will be addressed in this talk. First, it is known that HC statistics are asymptotically sensitive for some special kind of alternatives that dier from the null distribution in the moderate tails. Based on the theory of stochastic processes we will give an explanation why a specic intermediate range is crucial for HC statistics. Furthermore, the calculation of suitable critical values for HC tests remains a challenging problem for nite samples. For this case we discuss some alternative approximations for the distribution of the HC statistic.

Some New Results on Goodness of Fit Tests in Terms of Local Levels

Sandra Landwehr, Veronika Gontscharuk, Helmut Finner

Department of Statistics in Medicine, Faculty of Medicine, Heinrich-Heine-University Duesseldorf, Germany

Goodness of fit tests, which can be viewed as simultaneous multiple test procedures, are typically sensitive for specific alternatives. For example, it is well-known that the Kolmogorov-Smirnov test has high power against alternatives that differ from the null hypothesis in the central range and rather low power against alternatives that deviate from the null in the tails. Unlike the Kolmogorov-Smirnov case, tests based on the Higher Criticism statistic are known to be sensitive in the moderate tails, see Eicker (1979) or Jaeschke (1979).
With regards to the works of, e.g., Donoho & Jin (2004, 2008) and Hall & Jin (2008, 2010), we observe a renewed interest in the Higher Criticism statistic in the context of classification problems in high-dimensional settings with rather sparse signals. In general, an interesting aspect for a multiple test procedure is to consider for each order statistic of the p-values its chance to exceed the corresponding critical value. We will call these probabilities local levels. For goodness of fit tests we can understand those as an indicator as to where one would expect high/low local power of the test, and thus, local levels
provide a method to compare tests with respect to areas of sensitivity. In this talk, we provide a methodological basis for obtaining local levels for a wide range of goodness of fit tests. We present new results on the properties of finite and asymptotic local levels of the Kolmogorov-Smirnov and Higher Criticism tests, which help to gain a deeper understanding of their limiting behaviour. Moreover, in view of these results, we address the question of how to construct goodness of fit tests with equal local levels.

[1] Eicker, F. (1979) The asymptotic distribution of the suprema of the standardized empirical processes. Ann. Stat. 7, 116-138.
[2] Jaeschke, D. (1979) The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. Ann. Stat. 7, 108-115.
[3] Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures.Ann. Stat. 32, 962-994.
[4] Donoho, D. and Jin, J. (2008) Higher criticism thresholding: Optimal feature

selection when useful features are rare and weak. PNAS 105, 14790-14795.
[5] Hall, P. and Jin, J. (2008) Properties of Higher Criticism under strong dependence. Ann. Stat. 36, 381-402.
[6] Hall, P. and Jin, J. (2010) Innovated Higher Criticism for detecting sparse signals in correlated noise. Ann. Stat. 38, 1686-1732.

Normal Probability Plots with Confidence

Wanpen Chantarangsi, W. Liu, F. Bretz, A.J. Hayter, S. Kiatsupaibul

S3RI and School of Mathematics, University of Southampton, UK

The importance of normal distribution is undeniable since it is an underlying assumption of many statistical tools. This is the reason why checking the assumption of normality is required prior to applying the normal model to data in hand. There are two types of procedures in assessing whether a population has a normal distribution based on a random sample: graphical methods (e.g., Q-Q plots, histograms and boxplots) and non-graphical methods (e.g., Anderson-Darling test and Shapiro-Wilk test). The normal quantile-quantile plot (Q-Q plot), also called normal probability plot, is the most commonly used diagnostic tool for assessing whether a random sample is drawn from a normally distributed population. In addition, Normal probability plots are also routinely used to check whether the residuals are normally distributed.

In this study we provide, on a normal probability plot, exact simultaneous intervals into which the points fall with probability $1-\alpha$ if the sample is taken from a population with normal distribution. These simultaneous intervals provide therefore an objective way to judge whether the plotted points fall close to a straight line. Several different sets of simultaneous intervals associated with Kolmogorov-Smirnov test (D test), Michael test (D_m test), D_n test, D_{beta} test and D_{new} test are investigated, including the power comparison among these graphical methods and with the non-graphical Anderson-Darling test and Shapiro-Wilk test.

Multiple Testing in Group Sequential Trials using Graphical Approaches

Frank Bretz, Willi Maurer

Novartis, Switzerland

We consider the situation of testing multiple hypotheses repeatedly in time using recently developed graphical approaches. We focus on closed testing procedures using weighted group sequential Bonferroni tests for the intersection hypotheses. Under mild monotonicity conditions on the error spending functions, this allows the use of sequentially rejective graphical procedures in group sequential trials. The methodology is illustrated with a numerical example from a real clinical trial.

Fixed Sequence Testing in Adaptive Designs with Sample Size Reassessment

Franz Koenig, Florian Klinglmüller, Cyrus Mehta, Lingyun Liu

Medical University of Vienna Center for Medical Statistics, Informatics, and Intelligent Systems Section for Medical Statistics, Austria

Recently adaptive designs attracted much interest where the sample size is increased for promising interim results of the primary endpoint still using the conventional test statistic [1]. Complex multiple testing strategies for testing primary and secondary endpoints have been extensively discussed for fixed sample designs, only few publications deal with this issue in the group sequential setup [2, 3]. We investigate the impact on the multiple type I error rate when hierarchically testing primary and secondary endpoints using adaptive designs with sample size reassessment using conditional and predictive power arguments [4] for the primary endpoint. Different testing strategies will be investigated. E.g., we elaborate whether in the promising zone approach by [1] after the rejection of primary endpoint, the secondary endpoint can be also tested using the conventional pooled test statistic (ignoring the adaptive nature of the trial). This will be compared to other adaptive methods. Extending the work of [5] and [1] we define promising zones for both primary and secondary endpoints, where the type I error rate will be strictly controlled if a certain type of sample size increase is performed. References: [1] Mehta, C. R. and Pocock, S. J. (2010). Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statistics in Medicine, 30:3267-3284. [2] Glimm, E., Maurer, W. and Bretz, F. (2010). Hierarchical testing of multiple endpoints in group-sequential trials. Statistics in Medicine, 29:219-228. [3] Tamhane, A. C., Mehta, C. R. and Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. Biometrics, 66:1174-1184. [4] Bauer P, Koenig F. (2006). The reassessment of trial perspectives from interim data--a critical view. Statistics in Medicine; 25(1):23-36. [5] Brannath, W. and Koenig, F. (2007). Multiplicity and flexibility in clinical trials. Pharmaceutical Statistics, 6:205-216.

Sequentially Rejective Graphical Procedures in Adaptive Treatment Selection Designs

Toshifumi Sugitani, Frank Bretz, Willi Maurer, Toshimitsu Hamasaki

Department of Biomedical Statistics Osaka University Graduate School of Medicine, Japan

In recent clinical trials, adaptive designs incorporating treatment selection at pre-specified interim analyses have become more popular in practice as they can integrate two clinical studies into a single study. In such designs, the final result of continued treatments should be statistically independent of the interim result of discontinued treatments. However, it seems that this independence is not so much considered in the literature: The general methodological principle for adaptive treatment selection designs relies on the spirit described by, for example, Bretz et al.(Biom. J; 48, 623-634, 2006). However, this general principle does not meet the aforementioned independence condition (i.e., conditional error rate of the continued treatments is not free from the result of the discontinued treatments). In this presentation, by extending the previous work of Sugitani et al. (Biom. J; to appear), we describe a new methodological principle for adaptive treatment selection designs in terms of closure principle. Following the principle, we then consider the extension of the Bonferroni-based graphical approach by Bretz etal. (Statist. Med; 28, 586-604, 2009) to adaptive treatment selection design settings. The extended graphical approach still preserves sequential rejectiveness (i.e., consonance) as well as the independence between the final result of continued treatments and the interim result of discontinued treatments. Furthermore,we discuss the application of the extended graphical approach to more complex clinical trials (e.g., combination of adaptive treatment selection design and group-sequential design).

Flexible Sequential Designs for Multi-Arm Clinical Trials

Dominic Magirr, Nigel Stallard, Thomas Jaki

Medical and Pharmaceutical Statistics Research Unit,
Lancaster University, UK

Adaptive designs that are based on group-sequential approaches have the benefit of being efficient as stopping boundaries can be found that lead to good operating characteristics with test decisions based solely on sufficient statistics. The drawback of these so called "pre-planned adaptive" designs is that unexpected design changes are not possible without impacting the error rates. "Flexible adaptive designs" on the other hand can cope with a large number of contingencies at the cost of reduced efficiency. In this work we focus on two different approaches for multi-arm multi-stage trials which are based on group-sequential ideas and discuss how these pre-planned adaptive designs can be modified to allow for flexibility. We demonstrate how an impressive overall procedure can be found by combining a well chosen pre-planned design with an application of the conditional error principle to allow flexible treatment selection.

Asymptotic FDR Control under Weak Dependence and the Null Problem: A Counterexample

Helmut Finner, Veronika Gontscharuk

Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Duesseldorf, Germany

A question of general interest is whether the false discovery rate (FDR) is asymptotically controlled under weak dependence for classical multiple test procedures which control the FDR under certain independence assumptions, e.g. linear step-up tests and their plug-in variants, procedures based on the asymptotically optimal rejection curve. In multiple hypotheses testing, weak dependence typically appears if the empirical cumulative distribution function of all test statistics with respect to true null hypotheses converges in some probabilistic sense as the number of true null hypotheses tends to infinity. In this talk we will restrict attention to the case where p-values are at hand and where a hypothesis is rejected if the corresponding p-value is less than or equal to some random data dependent threshold τ. If τ is asymptotically bounded away from zero, that is, if the proportion of rejected hypotheses is asymptotically positive, the assumption of weak dependence typically suffices to guarantee asymptotic FDR control. More difficult is the case where τ is close to zero and the proportion of rejected hypotheses tends to zero (paraphrased as the null-problem). The null-problem can often be observed in applications with small sample sizes but a large number of hypotheses, e. g., in SNP and gene expression analyses. It will be shown by means of an example that the FDR may become arbitrarily large under weak dependence if the null problem appears.

How to take into account dependency into multiple testing procedures?

Gilles Blanchard, Sylvain Delattre, Pierre Neuvial, Etienne Roquain

UPMC, University Pierre et Marie Curie, Laboratoire de Probabilités et Modèles Aléatoires, France

One of the most challenging issue in contemporary multiple testing is to correctly incorporate the (potentially strong) dependencies between the individual tests. In this talk, we first show that the false discovery proportion (FDP) of Benjamini-Hochberg (BH) procedure can have a distribution not concentrated around its expectation. Hence, the FDP can be well above the nominal level, which implies a false interpretation of the proportion of errors among the rejections of the BH procedure.

We then propose to circumvent this problem by following two distinct approaches, depending on whether or not it is desirable to keep the original p-values to make the individual decisions:
1) keeping the original p-values: we show that the false discovery proportion can be correctly controlled with other types of procedures, based on a "joint family-wise error rate" control;
2) transforming the original p-values: by assuming that the dependency structure is known, we describe a PCA-based procedure to effectively remove dependency from the original p-values, resulting in an improved power.

The new procedures will be presented together with theoretical results and illustrative numerical experiments.

P-value Evaluation for Multiple Testing of Means under the Existence of Positive Correlations

Yoshiyuki Ninomiya

Institute of Mathematics for Industry, Kyushu University, Japan

We consider a multiple testing problem in which the mean vectors for several groups are compared. We assume that data are distributed according to a multivariate Gaussian distribution with unknown positive correlations, some of which are large. In this case, if we use conventional t statistics for each testing, some of them have large positive correlations, and then evaluating an upper bound of the p-value by Bonferroni's method leads to a too conservative testing. After estimating the unknown correlations, we apply a new evalution method for the upper bound, which is made by combining Taylor et al. (2007, Biometrika) and Ninomiya & Fujisawa (2008, Biometrics).

Permutation-Based Confidence Bounds for the False Discovery Proportion

Aldo Solari

University of Milano-Bicocca, Italy

One of the most challenging issues in multiple testing concerns the dependence structure of the p-values. Permutation-based methods, when applicable, can be a powerful tool as they adapt to the unknown joint distribution of the p-values. Examples are the method of Westfall and Young (1993), which controls the familywise error, and the method of Meinshausen (2006), which controls the false discovery proportion.

Here we present an uniform improvement of the Meinshausen's method that is applicable to large-scale testing situations as they arise, for example, in genomics. This improvement is analogous to the improvement from the single-step procedure to the step-down procedure of Westfall and Young. Software to perform the proposed procedure is available in the cherry R package.

A Decision-Theoretic Approach to Multiple Inference

Carl-Fredrik Burman

AstraZeneca, Pharmaceuticals, Mölndal, Sweden

Multiple comparisons are difficult. Not only can the methodology be challenging to grasp for an investigator, the variety of testing procedure at hand can be daunting. In any given experimental situation, there are often a multitude of procedures that control the family-wise error rate (FWER) strongly. So which one should be used and how should different hypotheses be weighted? The statisticians' role is often to guide the investigator team in this decision process. We show how a graphic approach to multiple testing can be used pedagogically.

For confirmatory clinical trials, FWER control is generally a regulatory requirement. The investment in the phase III programme is large, and the outcome may depend on the choice of multiple testing strategy. As many individuals are involved, it is desirable to provide a structured support of the decision. We propose that a decision-theoretic framework should be used, to explicitly model the value of different sets of rejections as well as the prior knowledge and uncertainty regarding the parameters of interest. Based on such assumptions, the choice of multiplicity procedure can be optimised.

Adaptive Statistical Significance Threshold for Inference Guided Discovery Studies

Cheng Cheng, Jun W Hyun, Stanley Pounds

St. Jude Children's Research Hospital, USA

Recent advancement of biotechnology and data mining applications has vastly increased the demands for statistical inference guided discovery (screening and selection) of associations among a few hundred to a few million variables observed on a large number of subjects. For example, currently a human genome-wide association study (GWAS) typically screens 0.5 to 2 million variables with the same number of statistical tests. A popular choice for a statistical significance threshold for GWAS is the “genome wide significance” P<10-7. Apart from biological issues, notably for 0.5 million tests, this threshold coincides with the Bonferroni adjustment to control the family-wise type-I error rate (FWER) at 0.05; hence a very conservative choice that may not well suited for the exploratory purpose. This type of adjustment combined with the “winner’s curse” phenomenon ultimately reduces statistical efficiency, requiring larger and larger sample sizes in order to discover signals truly underlying the subject matter. Alternative approaches include less conservative variations of the Bonferroni procedure to control (generalized) FWER, and control and estimation of the false discovery rate (FDR). A frequent issue in practice is which FDR level is suitable for the study at hand; often multiple levels are examined (e.g., using the q-value) and one is chosen based largely on subjective, practical, and (but less often) subject-matter considerations. We have developed a novel approach to massive multiple hypotheses tests that determines the significance threshold for a given study adaptively (data-driven), by minimizing a criterion function. This procedure is adaptive in the sense that it relaxes the significance threshold as much as possible to reduce the number of false negative errors, under a penalty of incurring too many false positive errors; the resulting threshold is a point where further relaxation beyond it will no longer be beneficial, i.e., reduction of the false negative error level no longer outweighs the fast increase of the false negative error level. The criterion function is constructed by careful considerations of the shape of the empirical distribution function of all P values and the form of a penalty function. This function consists of two terms: the driving term is a function similar to the area under an ROC curve, and the penalty term is a

special function of the expected number of false positive errors. Performance comparisons to popular FDR-based methods are obtained on both simulated and real datasets. The results show that the adaptive significance threshold can provide a nice balance between the levels of false positive and false negative errors in certain rather difficult scenarios.

Testing and Multiple-testing using Neutral-data Comparisons

Dan J. Spitzner

University of Virginia, USA

This talk will describe a method of calibrating Bayes factors derived from a concept known as a neutral-data comparison. The result is a novel assessment of evidence that may be interpreted as a well-formulated alternative to a Bayes factor that is drastically less sensitive to the choice of prior. Neutral-data comparisons furthermore admit a novel approach with which to adjust for multiple testing, which the investigator has shown to exhibit remarkable performance in variable selection settings. The technique is also promising for clustering applications.

When choosing between two models, M0 and M1, say, neutral data are imaginary data that are thought subjectively to exhibit evidence for neither M0 nor M1--they are neutral between them. Consequently, because the data are neutral, one might expect that a Bayes factor calculated on such data would equal one; however, the Jeffreys-Lindley paradox implies that it will not be one if the scales of the priors differ drastically between the two models (and neutral data are not closely tied to scale). The theory of neutral-data comparisons interprets this phenomenon as an incoherency that results with the use of vague priors, and corrects for it by replacing the Bayes factor's evidence assessment as the ratio of posterior to prior odds with the ratio of posterior odds calculated on the observed data to that calculated on neutral data.

The talk will discuss approaches to selecting neutral data, including methods that consider rates of asymptotic testing consistency, and methods that connect neutral data to unit-information priors and BIC. The latter are particularly helpful for handling nuisance parameters and for working within model-choice contexts that involve multiple models.

An advantage of neutral-data comparisons is that they are easy to configure for the detection of non-signals, which requires delicate maneuvering to achieve using Bayes factors. When suitably configured within a multiple-testing context, this flexibility can yield powerful asymptotic-consistency properties, including the ability to select the correct model in ultra-high dimensions even when the correct model involves sparse signals.

These ideas are examined and illustrated on data exemplifying the Behrens-Fisher problem, and on a data set of adverse-event frequencies in a vaccine trial.

More discussion appears in the article Spitzner, D. J. (2011), Neutral-data comparisons for Bayesian testing, Bayesian Analysis, 6:603-638 (http://bayesian.org/publications), and the unpublished technical report accessed by the link,
http://people.virginia.edu/~djs4y/preprints/NDC2_v1.pdf.

Implementing False Discovery Rate Procedures for Simulation-Based Tests With Bounded Risk

Axel Gandy, Georg Hahn

Imperial College London, UK

Consider multiple hypotheses to be tested for statistical significance using a procedure which controls the False Discovery Rate (FDR), e.g. the method by Benjamini-Hochberg. Instead of observing all p-values directly, we consider the case where they can only be computed by simulation. This occurs e.g. for bootstrap or permutation tests.

Naively, one could use an equal number of samples for the estimation of the p-value of each hypothesis and then apply the original FDR procedure. This technique is certainly not the most efficient one, nor does it give any guarantees on how the results relate to the FDR procedure applied to the true p-values.

This talk presents MMCTest, a more sophisticated approach that uses fewer samples for all those hypotheses which can already be classified with sufficient confidence and more samples for all those which are still unidentified. The algorithm is designed to give, with high probability, the same classification as the one based on the exact p-values.

A simulation study on actual biological data, given by a microarray dataset of gene expressions, shows that for a realistic precision, MMCTest draws level with the performance of current methods which unlike MMCTest do not give a guarantee on its classification being correct. An ad-hoc variant of MMCTest which forces a complete classification outperforms established methods.

Family-Wise Control of Both Type I And Type II Errors in Clinical Trials

Bushi Wang, Naitee Ting

Boehringer Ingelheim Pharmaceuticals, Inc., USA

Controlling the family-wise error rate has been implemented as the strong control of family-wise type I error in the statistics community without any doubt. However, we discussed in this article that the family-wise control of type II error is sometimes as important as controlling the type I error, especially in clinical trials. One challenge to control the family-wise type II error is how to define such error rate when we usually have a composite alternative hypothesis. We propose a simple definition of family-wise type II error rate and several procedures to control family-wise type I and type II error rate simultaneously. Sample size determination is also discussed in detail in the article.

Type II generalized Family-Wise Error Rate Formulas with Application to Sample Size Determination

Philippe Delorme, Pierre Lafaye de Micheaux, Benoit Liquet, and Jérémie Riou

Danone Research, Bordeaux University, France

Co-primary endpoints are increasingly used in clinical trials. The significance of the study is concluded if and only if at least r null hypotheses are rejected among the m null hypotheses. In this context, statisticians need to take into account multiplicity problems. Nowadays, an extensive literature exists for data analysis and sample size computation when r=1 and r=m. Therefore, the aim of this work consists in the development of a methodology which permits to compute the sample size for all common procedures used in clinical trials, namely ``single-step'' and ``step-wise'' procedures, for any value of r. The results are then applied on two clinical trials ``Pre-RELAX-AHF'' and ``Pneumovacs''. This trials are respectively designed to investigate the effectiveness of a drugs against acute heart failure, and the immunogenicity of a vaccine against pneumococcus, for both on a set of co-primary endpoints.

Number of False Rejections and Differential Equations

Marsel Scheer

Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Duesseldorf, Germany

Controlling the k-FWER at level alpha in multiple testing problems with m null hypotheses means that the probability of rejecting at least k true null hypotheses is bounded by alpha, cf. e.g. [1,2,3]. Considering k as fixed may be viewed as unsatisfactory. In this talk we propose a new and more flexible approach, NFRX control for short, where k is allowed to depend on the unknown number m1 of false null hypotheses (NFRX = Number of False Rejections eXeedance). For example, it seems more appropriate to require k = k(m1) to be small (large) if the number of false null hypotheses is small (large). It will be shown, that the NFRX is connected to differential equations. Based on these equations, we construct suitable multiple tests and present sufficient conditions such that the NFRX is asymptotically (m tends to infinity) controlled, that is, the probability of rejecting at least k(m1) true null hypotheses is asymptotically bounded by alpha. The material presented here is part of [4].

[1] Victor, N. (1982). Exploratory data analysis and clinical research. Meth. Inform. Med., 21, 53-54.
[2] Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In: Bauer, P. et al. (Eds.): Multiple Hypothesenpruefung. Springer, Berlin, 154-161.
[3] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. Ann. Statist., 33, 1138-1154.
[4] Scheer, M. (2012). Controlling the Number of False Rejections in Multiple Hypotheses Testing. Dissertation, Heinrich-Heine Universitaet, Duesseldorf.

Scaled Fasle Discovery Proportion and Related Error Metrics

Djalel Eddine Meskaldji, Jean-Philippe Thiran, Stephan Morgenthaler

Ecole Polythecnique Fédérale de Lausanne, Switzerland

In multiple testing, a variety of control metrics of false positives have been introduced such as the Per Family Error Rate (PFER), Family-Wise Error Rate (FWER), the False Discovery Rate (FDR), the False Exceedence Rate (FER). In this talk, we present a comprehensive family of error rates together with a corresponding family of multiple testing procedures (MTP). Based on the needs of the problem at hand, the user can choose a particular member among these MTPs. The new error rate limits the number of false positives FP relative to an arbitrary non-decreasing function s of the number of rejections R. The quantity is called, the scaled false discovery proportion SFDP=FP/s(R). We present different procedures to control either the P(SFDP>q) or the E(SFDP) for any choice of the scaling function.

An obvious choice is s(R)=min(R;k). As does FDR, this particular error rate FP/s(R) accepts a fixed percentage of false rejections among all rejections, but only up to R =k, then a stricter control takes over and for R > k, the number of false positives is limited to a percentage of the fixed value k, similar to PFER. The corresponding family of multiple testing procedures bridges the gap between the PFER (k =1) and the FDR (k = number of tests). A similar such bridge is obtained when s(R)=Rg with 0<=g<=1, which for g=0.5 controls the percentage of false discoveries relative to the square root of R. In the talk, we discuss the choice of the parameters k and g based on the minimization of the expected loss t E(FP) - E(TP) = t E(FP) - E(R - FP) which is based on the idea that a false positive costs a penalty of 1<t units, while a true positive corresponds to a gain of 1 unit.

Replicability Analysis for Genome-Wide Association Studies

Ruth Heller, Daniel Yekutieli

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

The paramount importance of replicating associations is well recognized in the genome-wide associaton (GWA) research community, yet methods for assessing replicability of associations are scarce. We suggest an empirical Bayes method for discovering whether results have been replicated across studies, in which we estimate the optimal rejection region for discovering replicated associations. We show that this region can be very different than the rejection region of a typical meta-analysis, that is aimed at discovering associations. We also show that the difference between the optimal rejection region and a rejection region based on p-values may be large.

We apply our method to four type two diabetes (T2D) GWA studies. Out of 803 SNPs discovered to be associated with T2D using a typical meta-analysis, we discovered 219 SNPs with replicated associations with T2D. We recommend complementing a meta-analysis with a replicability analysis for GWA studies.

Bayesian Variable Selection and Multiplicity Adjustment

Ziv Shkedy

Hasselt University, Belgium

In a typical dose-response experiment the response variable is measured across increasing levels of an active compound. The aim of the analysis is to test the null hypothesis of no dose effect against an order alternative. Several tests, such as the likelihood ratio test and multiple contrast tests (MCT) can be used to test the null hypothesis within the frequentist framework.

In this paper we focus on hierarchical Bayesian modeling of dose response data. Within the hierarchical Bayesian framework, one of the major challenges is related to the question how to perform Bayesian inference and in particular how to adjust for multiplicity. We discuss an order restricted Bayesian Variable Selection (BVS) model which can be used in order to calculate the posterior probability of a specific model given the data and the model parameters. In particular, we use the posterior probability of the null model (of no dose effect) for inference and investigate the similarity of the BVS model to the MCT procedure.

Several real examples and a simulation study will be discussed in order to illustrate the similarity and difference between the MCT and BVS approaches.

FDR and FNR: Comparison of Loss Functions in Epidemiologic Surveillance

Biggeri A, Catelan D, Cecconi L

Department of Statistics, Informatics and Applications "G. Parenti" University of Florence (IT)

We present a Bayesian approach to multiple testing. It consists in a tri-level hierarchical model which assigns a positive probability mass to the null. Marginalizing the full joint posterior distribution, for each test hypothesis, we can obtain posterior classification probabilities – the posterior probability of the null. Posterior inference on classification probabilities is highly dependent on the choice of the hyper-prior and we suggest to rely on subject-specific information to derive a portfolio of informative priors. We then show the connections between the false discovery – false non discovery rate approaches and, using posterior summaries, we explore different decision rules. The cost of erroneous decisions is dependent of the trade-off between False Discoveries and False Non discoveries (in terms of absolute number or rates). Different loss functions are discussed and the usual assumption that all false discoveries and false non discoveries are equally undesirable, is relaxed. We consider settings in which additional information on the relative importance / cost of false discoveries vs false non discoveries is modeled.

Practical implications are discussed with example in spatial statistics and Epidemiologic Surveillance. In particular, data from a cross-sectional study carried out in the Campania Region (IT) to study the spatial distribution of 16 parasites in 121 ovine farms were used. A multivariate tri-level hierarchical spatial Bayesian model was fitted to obtain estimate of posterior classification probabilities, and four decision rules were used to parasitological risk profiling. The second example came from a study on prevalence of lymphohematopoietic malignancies on the years 2001-2010, males and females, Sardinia Region (Italy), 377 municipalities. A tri-level hierarchical spatial Bayesian model was fitted to estimate for each municipality the posterior probability to belong to the set of non-divergent areas. We show how to incorporate prior information to explore the presence of a cluster of cases within a subset of municipalities.

Catelan D, Rinaldi L, Musella V, Cringoli G, Biggeri A. Statistical approaches for farm and parasitic risk profiling in geographical veterinary epidemiology. Stat Methods Med Res, 2012, 21(5): 531-43.
Catelan, D; Lagazio, C; Biggeri, A. A hierarchical Bayesian approach to Multiple Testing in Disease Mapping. Biometrical Journal, 2010, 52(6): 784–797.
Parmigiani G, Inoue L. Decision Theory: Principles and Approaches. Wiley, 2009: 136-139.

Presenters Index

M

N

P

R

S

T

W

Y

NOTES