#1

Felix Abramovich, Yoav Benjamini (Tel Aviv University, Israel),
David Donoho, Iain Johnston (Stanford University, USA)

Adapting to unknown sparsity by control of the False Discovery Rate

We consider the problem of recovering a high-dimensional vector observed in white noise, where the vector is known to be sparse, but the degree of sparsity is unknown. We consider three different ways of defining sparsity of a vector: using the fraction of nonzero terms; imposing power-law decay bounds on the ordered entries; and controlling the ℓ^p norm for p small. We obtain a procedure which is asymptotically minimax for ℓ^r loss, simultaneously throughout a range of such sparsity classes. The simultaneous asymptotic minimaxity is achieved by a data-adaptive thresholding scheme, based on controlling the *False Discovery Rate* (FDR). FDR control is a recent innovation in simultaneous testing, in which one seeks to ensure that at most a certain fraction of the rejected null hypotheses will correspond to false rejections. In our treatment, the FDR control parameter q plays an informative role in understanding how to achieve asymptotic minimaxity. Our results say that letting $q \to 0$ with problem size n is sufficient for asymptotic minimaxity, while keeping $q > 1/2$ prevents asymptotic minimaxity. To our knowledge, this relation between ideas in simultaneous inference and asymptotic decision theory is new. Our work suggest insights about a class of model selection procedures which has been introduced recently by several authors. These new procedures are based on complexity penalization of the form $2 \cdot \log\left(\frac{\text{potential model size}}{\text{model size}}\right)$. We exhibit a close connection to FDR-controlling procedures with q tending to 0, which strongly supports a conjecture of simultaneous asymptotic minimaxity of such procedures.

#2

Yekutieli D., Benjamini Y. (Tel Aviv University, Israel)

Genetic dissection of quantitative traits using the False Discovery Rate criterion

Genetic dissection of quantitative traits is achieved through a series of individual statistical tests, each testing the effect of the genetic structure in a given locus on one of many quantitative traits. The problem of multiple comparisons is the main statistical problem in quantitative trait mapping, some researchers even stressed that the resolution of this problem has important consequences on the future of the field. Correcting for multiplicity in a QTL study is very difficult due to the large number of hypotheses tested, often exceeding 100,000 tests and complex dependency structure. Dependency between trait measurements. For each trait, test statistics corresponding to closely located genetic loci are highly correlated. The presentation will focus on addressing the problem of multiple comparisons in quantitative trait mapping using the False Discovery Rate. If there are many QTLs the FDR thresholding is lower than the conventional Family Wise Error thresholding. The increase in power is particularly evident in large multiple comparison problems, where the conventional approach lacks power. I will show the validity of the existing FDR controlling procedures in QTL mapping. I will present the results of simulation studies and apply the FDR controlling procedures to real data. Finally a two stage FWE controlling modeling scheme will be presented. In the first stage the FDR criterion is used for screening promising QTLs. The second stage is a confirmatory study on the screened QTLs.

References:
1. Benjamini Y., Yekutieli D. "The control of the False Discovery Rate under dependence", Research paper of the Department of Statistics and OR Tel Aviv Uni. RP-SOR-97-04.
2. Churchill G. A., Doerge R. W., (1994) "Empirical Threshold Values for Quantitative Trait Mapping", Genetics 138: 963-971.
3. Lander E., Kruglyak L., (1995) "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results", Nature genetics 11 November 1995.
4. Yekutieli D., Benjamini Y., Resampling based FDR controlling multiple hypotheses testing. JSPI 82 (1999) 171-196.

#3

Yosef Hochberg (Tel Aviv University, Israel)

On Posterior P-values

Some typology of the the vast variety of inferential problems which cannot be specified apriory is discussed. General approaches for defining and evaluating posterior p-values are introduced. These are demonstrated with a particular problem involving testing a single null hypothesis with a series of same experiments due to repeats of "almost" significant p-values at earlier stages.

#4

Yoav Benjamini, Vered Madar (Tel Aviv University, Israel)

Non-Equivariant Simultaneous Confidence Intervals Less Likely to Contain Zero

The non-equivariant procedure presented by Benjamini and Stark [1996] yields simultaneous confidence intervals less likely to contain zero than the standard simultaneous procedures for many nonzero expectations of a set of independent random variables identically distributed up to their location parameters. We shall recall the procedure discussed in Benjamini and Stark [1996], and introduce a (slightly) more powerful non-equivariant procedure, as a modification to the procedure presented by Benjamini and Stark [1996].

References:
1. Benjamini, Y., and Stark, P.B., 1996. Non-Equivariant Simultaneous Confidence Intervals Less Likely to Contain Zero, J.Amer.Stat. Assoc., 91,329-337

#5

John D. Spurrier (University of South Carolina, USA),
Eleanne Solorzano (University of New Hampshire, USA)

Comparing More Than One Treatment to More Than One Control in Incomplete Blocks

The class of balanced treatment incomplete block designs is generalized to allow for comparison of k1 test treatments and k2 control treatments. The generalized class is equivalent to the class of balanced bipartite block designs considered by Jaggi, Gupta, and Parsad. Some results on design construction and A-optimality are given for small values of k1 and k2. Algorithms are developed for computing simultaneous confidence bounds for all test treatment versus control contrasts.

#6

Jason C. Hsu, James Rogers (Ohio-State University, USA)

Multiple comparisons of biodiversity

Ecological studies have often been incorrectly formulated so that the statistical error rate controlled is not the rate at which an error in decision is made. For example, water pollution monitoring under the U.S. Clean Water Act is currently formulated as a test-of-equalities problem, with proposals to reformulate it as a bioequivalence problem. Neither is correct; effluent toxicity trials are in fact non-inferiority trials. As another example, comparisons of biodiversity at the mesocosm level, based on indices such as Shannon's or Simpson's, currently treated most frequently as ANOVA problems, again often should be formulated as non-inferiority studies.
Using such non-exploratory studies with well defined errors in decision-making as examples, an outline of how a confidence set approach leads to statistical methods which control the error rate of decision-making will be indicated. These methods include average bioequivalence tests, intersection-union tests, stepdown methods with confidence sets, and multiple comparison with the best as special cases. Also indicated will be, in joint research with James Rogers, how second-order accurate, deterministic inference on Simpson's index can be achieved.

#7

Shanti S. Gupta (Purdue University, USA)

Empirical BAYES selection procedures for positive exponential family

In this paper, we are interested in the problem of simultaneous inference and selection from among $k \geq 2$ populations in comparison with a standard or control. The populations are denoted by $\pi_1, ..., \pi_k$. The random variable X_i associated with π_i is assumed to have the density $f(x_i|\theta_i) = c(\theta_i)e^{-x_i/\theta_i}\, h(x_i)$.
A nonparametric empirical Bayes approach is used to construct the selection procedure based on data from past *n* stages and the present stage. It is shown that this empirical Bayes

procedure is asymptotically optimal with a rate of order $O(n^{-1})$. The results are applicable to data arising from life-test experiments.

#8

C. Hirotsu (University of Tokyo, Japan)

A relationship between the isotonic inference and the changepoint analysis

The isotonic inference has many applications in industrial problems where there is a natural ordering in the levels of a treatment such as dose, temperature, time and so on. A changepoint model is also essential in the industrial process control. In the present paper we first demonstrate a relationship between the monotone hypothesis and the step type changepoint model in the normal means. It is simply that each of the corner vectors of the convex cone defined by the monotone hypothesis corresponds to the component hypothesis of the changepoint model. On the other hand a complete class of tests for the monotone hypothesis is shown to be all the tests that are increasing in every element of the projections of the observation vector onto those corner vectors. Then it happens that a statistic called max t and defined by the standardized maximum of those projections has been developed independently in two different streams of the isotonic inference and the changepoint analysis. It is actually the likelihood ratio test (lrt) statistic for the changepoint hypothesis. Those considerations are extended to various isotonic hypotheses including convexity, sigmoidicity and two-way ordered alternatives which induce slope change, inflection and two-way changepoint models as their corner vectors, respectively. The lrt for those changepoint hypotheses are easily derived and they become appropriate tests also for the original isotonic hypotheses by virtue of the complete class lemma. An exact and very efficient algorithm is introduced for calculating the distribution function of the max t type statistics. This will thus give a systematic way of approach to the isotonic inference other than the isotonic regression which is often too complicated excepting for the monotone hypothesis. Some power comparisons will also be given.

#9

Tetsuhisa Miwa (National Institute of Agro-Environmental Sciences, Japan)
A. J. Hayter (Georgia Institute of Technology, USA),
Wei Liu (University of Southampton, UK)

Exact calculations of the level probabilities in the unbalancedone-way models with applications to Bartholomew's test

An easy and quick procedure is presented to calculate the level probabilities under simple order of *k* independent normal random variables $Y_1, ..., Y_k$ with unequal variances. A crucial step in calculating the level probabilities is the calculation of orthant probabilities of the form $\Pr\{Y_1 < ... < Y_k\}$, where a recursive method (Hayter and Liu, 1996) and a cubic polynomial approximation method are employed.

These level probabilities have an application in the unbalanced one-way models for comparing *k* treatment effects, where Bartholomew (1959, 1961) proposed the likelihood ratio test for testing the homogeneity of the treatment effects against the simply ordered alternative hypothesis. Although there is some literature showing that Bartholomew's test has good

power properties, its null distribution for the unbalanced models has been difficult to calculate except for small *k*. The problem in the evaluation of this null distribution has been the difficulty in calculating these level probabilities. Our procedure to calculate the level probabilities allows the computation of the *p*-values and the critical points of Bartholomew's test for the unbalanced models.

References:
1. Bartholomew, D. J. (1959).A test of homogeneity for ordered alternatives, *Biometrika*, **41**, 36-48.
2. Bartholomew, D. J. (1961).Ordered tests in the analysis of variance, *Biometrika*, **48**, 325-332.
3. Hayter, A. J. and Liu, W. (1996).A note on the calculation of $\Pr\{X_1 < X_2 < ... < Y_k\}$, *Amer. Statist.*, **50**, 365

#10

Tony Hayter (Georgia Institute of Technology, USA),
Tetsuhisa Miwa (National Institute of Agro-Environmental Institute, Japan),
Wei Liu (University of Southampton, UK)

Combining the advantages of one-sided and two-sided multiple comparison procedures.

We consider the multiple comparison problems of making all pairwise comparisons among a set of treatment effects, and comparing a set of treatment effects with a control treatment, through the construction of simultaneous confidence intervals. One-sided procedures have the advantage of indicating the greatest number of significant differences in the direction of interest to the experimenter, whereas two-sided procedures provide both lower and upper bounds on the treatment differences. We present some new procedures which at the same specified confidence level combine the advantages of both the one-sided and the two-sided procedures. The new procedures are illustrated with some examples.

#11

J. Röhmel (Bundesinstitut für Arzneimittel , Germany)

Multiplicity in Clinical Trials - a Regulatory View

#12

George Y.H. Chi (FDA, USA)

Clinical Decision Rules and Multiple Endpoints

In clinical drug trials, one often observes a lack of a clear decision rule that is used to assess the effect of the drug based on the final outcome of the trial. The actual decision rule used may be either ad-hoc or post-hoc. Even in cases where a decision rule was defined, there may be a lack of optimality from the clinical perspective, or a lack of proper statistical support

structure from the statistical perspective in the decision rule. These may lead to serious inflation of type I error, either assessible or unassessible. This lack of a clear decision rule, or lack of optimality in the decision rule is essentially related to the problem of multiple endpoints. The purpose of this presentation is to illustrate some of these problems with practical examples, and to propose that a rational way of dealing with the multiple endpoints problem is to define clinical decision rules with proper statistical support structures that will provide the necessary basis for making valid statistical inference.

#13

Fortunato Pesarin (University of Padova, Italy)

Nonparametric combination of dependent partial tests

We deal with permutation approach of a variety of multidimensional problems of testing of hypotheses in a nonparametric framework. There are many multidimensional complex problems, frequently encountered in most applicational fields (agriculture, biology, clinical trials, engineering, the environment, experimental design, genetics, pharmacology, psychology, quality control, zoology, etc.), which are rather difficult to solve outside the permutation context, and in particular outside the method of nonparametric combination of dependent partial tests (Pesarin, 1992, 1999). Moreover, within parametric solutions based on normality of errors, it is sometimes impossible to obtain proper solutions. We mention, for instance, three such testing problems. One is related to the paired observations problem when scale coefficients are dependent on units, another is related to the two-way ANOVA, and the third to some multidimensional tests when the number of observed variables is higher than the sample size. In the first, within a parametric framework it is impossible to obtain estimates of standard deviations for each unit, whereas an exact effective permutation solution does exist. In the second it is impossible to obtain independent or even uncorrelated separate inferences for main factors and interactions, because all related statistics are compared with the same estimate of the variance of error components. Within the permutation approach, it is possible to obtain uncorrelated exact inferences in the general case and independent inferences under normality of errors. In the third, it is impossible to find estimates of the covariance matrix with more than zero degrees of freedom, whereas the nonparametric combination method allows for a proper solution, which is often asymptotically efficient. In a great variety of statistical analyses of complex hypotheses testing, when many response variables are involved or many different aspects are of interest, to some extent it is natural, and often convenient, first to process data by a finite set of $k > 1$ different *partial tests* (note that k is not necessarily equal to dimensionality q of responses). Therefore, they may be useful in a marginal or disjoint sense. But, when they are jointly considered, they provide information on a general overall (or global) hypothesis, which typically represents the true objective of the majority of multidimensional testing problems. combination in one (unidimensional) *combined* or *second-order* test, naturally arises. Multiple comparisons extensions of the above methodology are also discussed (Westfall and Young, 1993).

References:

1. Edgington E.S. (1995) *Randomization tests*, 3rd ed., Marcel Dekker, New York.
2. Good P. (1994) *Permutation Tests*. Springer-Verlag, New York.
3. Manly B.F.J. (1997) *Randomization, bootstrap and Monte Carlo methods in biology*. 2nd edition, Chapman and Hall, London.

4. Pesarin, F. (1992) A resampling procedure for nonparametric combination of several dependent tests. *Journal of the Italian Statistical Society*, 1, 87-101.
5. Pesarin, F. (1994) Goodness of fit testing for ordered discrete distributions by resampling techniques. *Metron*, LII, 57-71.
6. Pesarin, F. (1997a) An almost exact solution for the multivariate Behrens-Fisher problem, *Metron*, LV, 85-100.
7. Pesarin, F. (1997a) A nonparametric combination method for dependent permutation tests with application to some problems with repeated measures. *Proceedings of the ISI Satellite Meeting on Industrial Statistics*, C.P. Kitsos and L. Edler Eds., Pysica-Verlag, Heidelberg, 259-268.
8. Pesarin F. (1999). *Permutation testing of multidimensional hypotheses by nonparametric combination of dependent tests*. CLEUP, Padua.
9. Sprent P. (1998) *Data driven statistical methods*. Chapman and Hall, London.
10. Westfall P.H., Young S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.

 # 14

Dario Mazzaro, Fortunato Pesarin, Luigi Salmaso (University of Padova, Italy)

Repeated measures designs: a permutation approach and closed testing

We deal with permutation testing for multiresponse repeated measures designs and we consider a replicated unbalanced homoscedastic factorial design with fixed effects (Milliken, 1984) as the basic experimental plan. The design responses are measured in L time occasions. The usual linear model for single responses is: $\mathbf{Y} = \{y_{ji(l)r} = \mu_{(l)} + \alpha_{j(l)} + \beta_{i(l)} + \gamma_{ji(l)} + Z_{ji(l)r}\}$; $j = 1, 2$; $i = 1, 2$; $l = 1, ..., L$; $r = 1, ..., n_{ji}$; $\sum_{ji} n_{ji} = N$, where $y_{ji(l)r}$ are the experimental responses; $\mu_{(l)}$ is the population mean for the l-th measure; $\alpha_{j(l)}$ is the effect of the j-th level of factor A in the l-th measure; $\beta_{i(l)}$ is the effect of the i-th level of factor B in the l-th measure; $\gamma_{ji(l)}$ is the interaction effect between levels j and i of factors A and B in the l-th measure; $Z_{ji(l)r}$ are exchangeable experimental errors in the l measure from an unknown distribution P with zero mean and variance σ_l^2; finally, n_{ji} is the number of observations for each factor's levels combination. Thus, the total sample size is $L \cdot \sum_{ji} n_{ji} = L \cdot N$. The overall system of hypotheses is $H_0 : \cap_{l=1, ..., L} \{[H_{0A(l)}] \cap [H_{0B(l)}] \cap [H_{0AB(l)}]\}$, against the alternative $H_1 : \{H_0 \text{ is false}\}$, where the three partial hypotheses for each measure are $H_{0A(l)} : \{\alpha_{1(l)} = \alpha_{2(l)} = 0\}$ *vs* $H_{1A(l)} : \{\alpha_{1(l)} \neq \alpha_{2(l)}\}$, $H_{0B(l)} : \{\beta_{1(l)} = \beta_{2(l)} = 0\}$ *vs* $H_{1B(l)} : \{\beta_{1(l)} \neq \beta_{2(l)}\}$, $H_{0AB(l)} : \{\gamma_{11(l)} = \gamma_{12(l)} = \gamma_{21(l)} = \gamma_{22(l)} = 0\}$ *vs* $H_{1AB(l)} : \{H_{0AB(l)} \text{ is false}\}$, so that, the null hypothesis H_0 is true if all three partial sub hypotheses are true. Let us consider the three partial tests for effects in every measure. For example, the l-th permutation test for the effect of factor A is constructed by a linear combination of the two following statistics: $T^*_{A|1,(l)} = w_{11} \sum_{r=1}^{n_{11}} y^*_{11(l)r} - w_{21} \sum_{r=1}^{n_{21}} y^*_{21(l)r}$, $T^*_{A|2,(l)} = w_{12} \sum_{r=1}^{n_{12}} y^*_{12(l)r} - w_{22} \sum_{r=1}^{n_{22}} y^*_{22(l)r}$, where the weights w_{ji} are defined as: $w_{11} = (n_{22} - 2\nu^*)/(n_{11} - 2\nu^*)$, $w_{12} = (n_{22} - 2\nu^*)/(n_{12} - 2\nu^*)$, $w_{21} = (n_{22} - 2\nu^*)/(n_{21} - 2\nu^*)$, $w_{22} = (n_{22} - 2\nu^*)/(n_{22} - 2\nu^*)$, we jointly consider the L measures, then the permutation solution is based on the nonparametric combination methodology (Pesarin, 1999). It is worth noting that the new permutation approach, presented here, is highly robust, with respect to departures from normality of error terms in the linear

model for responses, since it is conditioned to the sufficient statistic represented by the data matrix. A comparative simulation study has been performed in order to evaluate the power of such exact tests. Multiple comparisons for the above tests are also discussed (Westfall and Young, 1993).

References:
1. Milliken G. A., Johnson D. E. (1984). *Analysis of messy data, designed experiments (vol. 1)*. Van Nostrand Reinhold Company, New York.
2. Pesarin F. (1999). *Permutation testing of multidimensional hypotheses by nonparametric combination of dependent tests*. CLEUP, Padua.
3. Pesarin F., Salmaso L. (1999). *Exact permutation testing on effects in replicated 2^k factorial designs*. Working Paper series, Department of Statistics, University of Padua. Submitted for publication.
4. Westfall P.H., Young S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.

#15

Fong Wang-Clow (Genentech, Inc., USA)

Alternative Approaches of Multiple Comparisons in Clinical Trials

There are a lot of types of multiple comparison in clinical trial. This presentation will provide four types of multiple comparisons and their applications in clinical trials. This presentation also shows how the needs of medical and clinical drive the use of the alternative approaches as opposing to the classical approaches.

References:
1. Bauer, P. (1991) "Multiple Testing in clinical Trials" Statistics in Medicine Vol. 10, 871-890
2. Dunnett, W.C. and Gent,M. (1996) "An alternative to the use of two-sided tests in clincial trials" Statistics in Medicine Vol.15, 1729-1738
3. Fleming, T.R. (1990) "Evaluation of active control trials in AIDS" Journal of Acquired Immune Deficiency Syndromes 3(Suppl.2):S82-87
4. Koch,G.G. Sansky,S.A. (1994) "Statistical considerations for multiple confirmatory protocols" DIA Workshop

#16

Juliet Popper Shaffer (University of Berkeley, USA)

Directional vs. Nondirectional Inference: Exploration of Their Relations and Suggested Compromises

In many multiple testing problems, point hypotheses are tested.Yet many researchers feel that point hypotheses (especially hypotheses of null effects) are unrealistic in most if not all situations, because a null hypothesis, for example that the mean of a control group is equal to the mean of a treatment group, is never (or almost never) exactly true. There are two ways of interpreting the test of a single hypothesis concerning the value of a parameter: as a test of a point null hypothesis at level α, or as a corresponding test of a pair of hypotheses concerning the sign of the parameter, each such directional hypothesis at level $\alpha/2$ (in a symmetric case) and at some level $\leq \alpha$ in a nonsymmetric case. In multiple testing, the relationship between

the two approaches (point hypothesis vs. directional pair) is more complex, and depends on the characteristics of the multiple procedure. In stepwise tests, it is not even clear that the familywise error rate remains $\leq \alpha$. (See Finner, 1999.) Even if it does, the relations between the directional and nondirectional levels is more complex than in testing single hypotheses. Some examples of relationships will be discussed, and possible compromise procedures, taking both the direction and nondirectional points of view into consideration, will be explored.

References:
1. Finner, H. (1999). Stepwise multiple test procedures and controlof directional errors. *Annals of Statistics* 27, 274-289.

#17

Qian Li, Laura Lu, Mo Huque (FDA, USA)

A Flexible Multiplicity Adjustment Approach and Its Application

We propose a general multiplicity adjustment approach that is flexible enough to allow one to specify decision criteria more freely compared to some of the conventional approaches, such as Bonferroni and Simes approaches. In fact, Bonferroni and Simes approaches are special cases of this flexible approach. Due to the flexibility of this approach, decision criteria can be chosen based on the direction of alternatives for the purpose of enhancing power. An example is illustrated for the application of this approach. In this example, several decision criteria are discussed to control the overall type I error when more than one studies were conducted for the same efficacy claim in clinical trial setting.

#18

William C. Horrace (University of Arizona,USA)

On the Ranking Uncertainty of Labor Market Wage Gaps

This paper uses multiple comparison methods to perform inference on labor market wage gap estimates from a regression model of wage determination. The regression decomposes a sample of workers' wages into a human capital component and a gender specific component; the gender component is called the gender differential or wage gap and is sometimes interpreted as a measure of sexual discrimination. Using data on fourteen industry classifications (e.g. retail sales, agriculture), a new relative estimator of the wage gap is calculated for each industry. The industries are then ranked based on the magnitude of these estimators, and inference experiments are performed using "multiple comparisons with the best" and "multiple comparisons with a control". The inference indicates that differences in gender discrimination across industry classifications is statistically insignificant at the 95% confidence level and that previous studies which have failed to perform inference on gender wage gap order statistics may be misleading.

#19

Ullrich Munzel (University of Goettingen),
Ludwig Hothorn (University of Hannover, Germany)

Nonparametric Multiple Comparisons in the Presence of Ties

Multiple comparisons are considered in a general nonparametric one-way layout, which includes continuous distributions as well as discontinuous distributions. The so called normalized version of the distribution function is used to define generalizations of the well known Mann-Whitney effect for pairwise comparisons. The corresponding effect estimator is shown to be asymptotically equivalent to a sum of independent, uniformly bounded random variables. This asymptotic argument is used to show the asymptotic normality and to estimate the correlation matrix of the estimators under alternative. Thus, it is possible to derive simultaneous confidence intervals of the effects as well as multiple test procedures for a nonparametric generalization of the Behrens-Fisher-Problem. The application to the many-to-one problem and to the all-pairs problem are discussed. Moreover, the correlation structure of the effect estimators is examined under the hypothesis of homogenitiy, i.e. the pairwise equality of the underlying distributions. The resulting test procedures for the many-to-one problem and the all-pairs problem have a product correlation structure and are generalizations of Steel's asymptotic test for the many-to-one problem (Steel,1959) and of Steel's and Dwass' (Steel, 1960; Dwass, 1960) asymptotic test procedure for the all-pairs problem, respectively.

References:

1. Brunner and Munzel (1999). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. Biometrical Journal, to appear.
2. Steel, R.D.G.(1959). A multiple comparison rank sum test: Treatments versus control. Biometrics 15, 560-572.
3. Steel, R.D.G.(1960). A rank sum test for comparing all pairs of treatments. Technometrics 2, 197-207.
4. Dwass, M (1960). Some k-sample rank-order tests. Contributions to Probability an Statistics (Eds. I. Olkin et al.) Stanford UniversityPress, 198-202.

#20

Hans-Helge Müller and Helmut Schäfer (University of Marburg, Germany)

Monitoring clinical trials:a general statistical principle for design changes

A general method is presented that allows to change statistical design elements such as the residual sample size during the course of an experiment or to include an interim analysis for early stopping when no formal rule for early stopping was foreseen, to increase or reduce the number of planned interim analyses, and to make other changes, without affecting the type I error risk. The method may be applied at the time of a pre-planned interim analysis or for administrative interim looks. The method is described in the general context of statistical decision functions and is based on the conditional rejection probability of a decision variable. The method is illustrated in an non-inferiority trial comparing extra-corporal shock wave therapy to standard surgical procedure in patients with tendinosis calcarea.

References:

1. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. Biometrics 1994;50:1029-1041. Correction in Biometrics 1996;52:380.
2. Proschan MA, Hunsberger SA.Designed extension of studies based on conditional power. Biometrics 1995;51:1315-1324.
3. Lehmacher W, Wassmer G. Adaptive sample size calculation in group sequential trials. Biometrics 1999;55: issue December, in press, temporary 131-135.
4. Fisher LD. Self-designing clinical trials. Stat Med 1998;17:1551-1562.
5. Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. Biometrics 1999;55:190-197.
6. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. Submitted to Biometrics.
7. Brannath W, Posch M, Bauer P. Recursive combination tests. Submitted to Biometrics.

#21

Peter H. Westfall (Texas Tech University),
Russell D. Wolfinger, Randall D. Tobias (SAS Inc., USA)

New Bayesian and Frequentist Software Solutions for Multiple Inferences

Our recent book "Multiple Comparisons and Multiple Tests using the SAS(R) System" contains a number of software solutions that are relatively new, and easy to implement using available SAS(R) procedures and macros. These include various Bayesian inferences for functions of mean and variance parameters in random effects models (possibly heteroscedastic), including (i) calculation of posterior probabilities of rankings, (ii) simultaneous Bayesian credible intervals, and (iii) Bayesian decisions using the loss function approach. Bayesian software for calculating posterior probabilities of point null hypotheses in free combination applications also is made available. On the frequentist side, we present software based on Shaffer's "method 2"that utilizes correlations from general models and sets of contrasts, as well as software for determining required sample sizes and estimating directional error rates.

#22

Daniel Q. Naiman (Johns Hopkins University, USA)

Tubes and Inclusion-Exclusion Probability Inequalities

In multiple comparisons, a key problem is to estimate or bound the probability of a union of events, and inclusion-exclusion plays an important role in attacking such problems. Naiman and Wynn (1997) introduced the notion of an *abstract tube* and described why it is relevant and useful this context. This notion will be reviewed and key properties of abstract tubes will be described. In particular, associated with any abstract tube is an inclusion-exclusion identity and corresponding truncation inequalities. Classical inclusion-exclusion arises as a special case, but there are theorems to the effect that these inequalities are typically weaker than can be obtained when a *smaller* tube is used instead. Recent new abstract tubes, some due to Dohmen (1999A, 1999B) and others building on the work of Naiman and Wynn (1992), all with applications to multiple comparisons and reliability will be presented.

References:

1. Dohmen, K. (1999A).‘‘An improvement of the inclusion-exclusion principle.” Arch. Math. (Basel) **72** no. 4, 298--303.
2. Dohmen, K. (1999B).‘‘Improved inclusion-exclusion identities and inequalities based on a particular class of abstract tubes.” Electron. J. Probab. **4** no. 5, 12 pp. (electronic)
3. Naiman, D. and Wynn, H.P. (1992).‘‘Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics.” *Annals of Statistics* **20** 43-76.
4. Naiman, D. and Wynn, H.P. (1997). ‘‘Abstract Tubes, Improved Inclusion-Exclusion Identities and Inequalities, and Importance Sampling.” *Annals of Statistics* **25** 1954-1983.

#23

G. C. Bhimani , Prashant R. Makwana (Saurashtra University, India)

Estimation and Comparison of Survival Rate of Patients with Thalassaemia (Major) under Different Treatment

The Life Table estimates of survival rate is obtained for the patients of T halassaemia - Major, residents of Saurashtra and Kutch region. The Standard Error of the estimates is obtained with Greenwood, Peto and direct method. Further the Confidence Bands are obtained with the Asymptotic Kolmogrov and Hall-Wellner method. Also the comparison of two drugs for Iron chillation is done using the Product Limit Estimation procedure of Kaplan and Meier. On the Basis of these estimates comparison of survival curves is done with several Non Parametric Tests.

#24

Ruediger Vollandt, Manfred Horn (University of Jena, Germany)

Sample size determination for multiple many-one and pairwise comparisons of proportions

We address the problem of sample size determination in many-one and pairwise multiple comparisons of proportions which are arcsin-root transformed. For one- and two-sided many-one comparisons, we provide the least favorable configuration which minimizes the all-pairs power. Corresponding explicit sample size formulas are given, also for the case of prior knowledge on the underlying success probabilities. The solutions for pairwise comparisons are restricted to the case $k = 3$.

References:

1. Hayter, A.J. and Liu, W. (1990)."Power asessment for tests of equalityof several proportions," Commun. Statist. - Theory Meth., 19, 19-30.
2. Hayter, A.J. and Lui, W. (1992). "A method of power assessment for tests comparing several treatments with a control," Commun. Statist.-Theory Meth.., 21, 1871-1889.
3. Horn, M. and Vollandt, R. (1998)."Sample sizes for comparisons of k treatments with a control based on different definitions of the power," Biometrical Journal,40, 589-612.
4. Liu, W. (1996)."On some single-stage, step-down and step-up procedures for comparing three normal means," Computational Statistics & Data Analysis, 21, 215-227.

#25

Shun-Yi Chen Tamkang University, Taiwan),
Hubert J. Chen (University of Georgia, USA)

A Single-Stage Procedure for Testing Homogeneity of Means Against Ordered Alternatives Under Unequal Variances

In this paper we apply a single-stage procedure described by Chen and Chen (1998) to test the equality of normal means against ordered alternatives, $H_a : \mu_1 \geq \mu_2 \geq ... \geq \mu_I$, in one-way layout when variances are unknown and unequal. Tables of percentage points and the power under a specific alternative needed for implementation are given. Relation between the single-stage and the two-stage test procedures is discussed.

References

1. Bishop, T.A., and Dudewicz, E.J. (1978), ‘‘Exact Analysis of Variance With Unequal Variances: Test Procedures and Tables,” *Technometrics*, 20, 419-430.
2. Chen, S.Y., and Chen, H.J. (1998), ‘‘Single-Stage Analysis of Variance UnderHeteroscedasticity,” *Communications in Statistics: Simulation and Computation*, 27(3), 641-666.
3. Ghosh, B. K. (1975), ‘‘On the Distribution of the Difference of Two t-Variables,” *Journal of the American Statistical Association*, 70, 463-467.
4. Marcus, R (1980),‘‘A Two-Stage Procedure for Testing Homogeneity of Means Against Ordered Alternatives in Analysis of Variance With Unequal Variances,” *Communications in Statistics - Theory and Methods*, A9(9), 949-963.
5. Williams, D.A. (1977), ‘‘Inference Procedures for Monotonically Ordered Normal Means,” *Biometrika*, 64, 9-14.

#26

J. Läuter (University of Magdeburg, Germany)

Analysis of multiple endpoints - confidence regions and model selection

The development of the spherical multivariate tests [Läuter 1996, Läuter, Glimm and Kropf 1998] has provided many possibilities of exact inference for clinical studies with multiple endpoints. The tests are applicable for the mean-value comparison of several populations with an unknown covariance matrix. They are especially suitable for cases with a high number of variables p and a small sample size n. Thus, the large dimension p can compensate for a too small sample size n to a certain extent to avoid numerical and statistical instability. In the talk, a new method of the calculation of linear principal- component scores is presented which is based only on the within-sample covariances and yields nevertheless an level-alpha test in every case. This method can be applied for the determination of exact confidence intervals of linear scores.

References:

1. Läuter, J. (1996): Exact t and F tests for analyzing studies with multiple endpoints. Biometrics 52, 964-970.
2. Läuter, J., Glimm, E., and Kropf, S. (1998): Multivariate tests based on left-spherically distributed linear scores. The Annals of Statistics 26, 1972-1988.

#27

T. Lang, A. Auterith , P. Bauer (University of Vienna, Austria)

Trendtests with Adaptive Scoring

The concept of adaptive two-stage designs is applied to the problem of testing the equality of several normal means against an ordered (monotone) alternative. The likelihood-ratio-test proposed by Bartholomew is known to have favourable power properties when testing against a monotonic trend. Tests based on contrasts provide a flexible way to incorporate available information regarding the pattern of the unknown true means through appropriate specification of the scores. The basic idea of the presented concept is the combination of Bartholomews test (first stage) with an "adaptive score test" (second stage) which utilizes the information resulting from isotonic regression estimation at the first stage. Several results of an extensive Monte Carlo simulation study will be reported concerning the power behaviour of these combination tests. This approach will be of special interest if, e.g., sample size reassessment is incorporated.

#28

Markus Neuhäuser (Byk Gulden Pharmaceuticals, Germany),
Frank Bretz (University of Hannover, Germany)

Nonparametric all-pairs multiple comparisons

Nonparametric all-pairs multiple comparisons based on pairwise rankings can be per-formed in the one-way design with the Steel-Dwass procedure. To apply this test, Wilcoxon's rank sum statistic is calculated for all pairs of groups; the maximum of the rank sums is the test statistic. For large sample sizes we introduce a generalization of the Steel-Dwass procedure for unbalanced designs and provide exact calculations of the asymptotic critical values. It should be noted that the method proposed by Critchlow and Fligner (1991, Commun. Statist. - Theory Meth. 20, 127-139) gives approximate critical values only in case of unbalanced sample sizes. For small sample sizes we recommend to use the new statistic according to Baum-gartner, Weiß, and Schindler (1998, Biometrics, 54, 1129-1135) instead of Wilcoxon's rank sum for the multiple comparisons. We show that the resultant proce-dure can be less conservative and, according to simulation results, more powerful than the original Steel-Dwass procedure. We also investigate the behaviour of the procedure in case of heteroscedasticity. We illustrate the methods with example data.

#29

Yosef Hochberg and Michael C. Mosier (Tel Aviv University , Israel)

Intersection-Union Procedures For Some Restricted Models

Intersection tests for Union hypotheses (IU tests) have been proposed for use in some multiple comparisons problems, see for instance Berger (1982), Laska and Meisner (1988), and Berger and Hsu (1996). In this paper, sharper IU tests are given for some restrictive multiple comparison problems. The restriction assumed in all these problems is that the multiple effects under consideration are all of the same sign. One and two sided examples where such assumptions are meaningful are discussed. Critical values for implementing the

sharper procedure are given for selected values of the underlying study parameters, in addition to a method of using SAS to give a very good approximation to the critical value for any desired case.

References:
1. Berger, R. L. (1982). Multiparameter Hypothesis Testing and Acceptance Sampling. Technometrics 24, 295-300.
2. Berger, R. L. and Hsu, J. C. (1996). Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets. Statistical Science, 11, 283-319.
3. Laska, E.M. and Meisner, M.J. (1989). Testing whether an identified treatment is best. Biometrics, 45, 1139-1151.

#30

Siegfried Kropf, Uwe Schmidt, Marilene S. Jepsen (University of Leipzig, Germany)

Two-Stage Adaptive Design in a Clinical Trial with Three Study Arms and Multiple Endpoints, Including a Test of Non-Inferiority

Three different types of heart-lung machine systems should be compared in a clinical trial with regard to their impact on blood coagulation and immune system parameters. These were a standard version (A) and two modifications (B) and (C). It should be shown that B and C are 'superior' to A and that the more economical version C is 'not inferior' to the expensive version B. The practical circumstances allowed for a randomized three-armed double blind trial. However, the prior knowledge was not sufficient to plan a study with fixed sample size or an usual sequential design, such that we startet a two-phase adaptive trial. There are four different sources of statistical multiplicity in this trial.

- The first one is the simultaneous consideration of blood coagulation and immune system. Here, we decided to treat both questions separately without special adjustment.
- Each of these two physiological categories is described by several variables (multiple endpoints). This problem is taken up by the application of so-called stable multivariate tests (Läuter, Glimm and Kropf, 1996).
- We have multiple comparisons between the three groups. More precisely, there are two tests of superiority of one treatment with respect to another one and one test of non-inferiority of a treatment. The three comparisons are carried out as tests with a priori ordering of hypotheses. The test of non-inferiority is transformed into a test of a contrast of all three treatments. This latter problem is extended to a series of modified hypotheses similar as in Bauer, Röhmel, Maurer, and Hothorn (1998).
- For the two-phase adaptive design, the methodology of Bauer and Köhne (1994) is used. The paper describes, how these basic techniques are combined. We utilize proposals by Kropf, Hothorn and Läuter (1997) to carry out multiple comparisons with multiple endpoints and modify an approach of Bauer and Kieser (1999) to treat multiple hypotheses in a two-phase design. The adaptation of the multivariate tests after phase I is used to reduce the laboratory costs in phase II (if necessary). Critical assumptions of the statistical methods and the complications resulting from the combination of techniques are discussed. The results of the trial are considered as given after phase I together with the conclusions for phase II. This allows for a trade-off between the expected gain and the related costs for the investigation of the remaining unanswered partial questions.

References:

1. Bauer, P.; Kieser, M. (1999). Combining Different Phases in the Development of Medical Treatments within a Single Trial. Statistics in Medicine 18, 1833-1848.
2. Bauer, P.; Köhne, K. (1994). Evaluation of Experiments with Adaptive Interim Analysis. Biometrics 50, 1029-1041.
3. Bauer, P.; Röhmel, J.; Maurer, W.; Hothorn, L. A. (1998). Testing Strategies in Multi-Dose Experiments Including Active Control. Statistics in Medicine 17, 2133-2146.
4. Kropf, S.; Hothorn, L. A.; Läuter, J. (1997). Multivariate Many-to-One Procedures with Application in Pre-Clinical Trials. Drug Information Journal 31, 433-447.
5. Läuter, J.; Glimm, E.; Kropf, S. (1996). New Multivariate Tests for Data with an Inherent Structure. Biometrical Journal 38, 5-22. Erratum: Biometrical Journal 40, 1015.

#31

Nancy L. Geller, (National Heart, Lung, Blood Institute, USA)

Multiple endpoints in clinical trials

In many clinical trials, there are several endpoints of comparable importance, rather than one primary endpoint. In stroke treatment, a number of scales have been used to measure improvement in outcome and no one scale is believed to assess all dimensions of recovery. The restriction to one primary endpoint when designing or analyzing such a clinical trial may be inappropriate. We discuss several hypothesis tests for multiple endpoints in two-armed clinical trials as well as their implementation, including group sequential monitoring (1). Procedures to follow the primary hypothesis test in order to determine which individual endpoints differ will also be described (2). We apply the results to the NIH t-PA Stroke clinical trial, originally published in 1995 in the New England Journal of Medicine (3).

References:

1. Tang, D.-I., Geller, N.L. and Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. Biometrics 49:23-30.
2. Tang, D.-I. and Geller, N.L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. Biometrics 55:1188-1192.
3. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995). Tissue plasminogen activator for acute ischemic stroke. New England Journal of Medicine 333:1581-87.

#32

Rand R. Wilcox (University of South California, USA)

Pairwise comparisons of trimmed means for two or more groups

The paper takes up the problem of performing all pairwise comparisons among *J* independent groups based on 20% trimmed means. Currently, a method that stands out is the percentile-t bootstrap method where the bootstrap is used to estimate the quantiles of a Studentized maximum modulus distribution when all pairs of population trimmed means are equal.However, a concern is that in simulations, the actual probability of one or more type I errors can drop well below the nominal level when sample sizes are small. A practical issue is whether a method can be found that corrects this problem while maintaining the positive features of the percentile-t bootstrap. Three new methods are considered, one of which achieves the desired goal. Another method, which takes advantage of theoretical results by

Singh (1998), performs almost as well but is not recommended when the smallest sample size drops below 15. In some situations, however, it gives substantially shorter confidence intervals. Some results on comparing dependent groups are reported as well.

#33

Peter Reitmeir (National Research Center for Environment and Health, Germany)

On the use of Bootstrap Cut-Off Tests in a closed testing procedure

Tests for a global hypothesis can be derived by simultaneous consideration of p values based on tests for the corresponding single hypotheses. These so called cut-off tests, however, are highly conservative procedures, because the dependencies among the single tests are not incorporated in the test decision. In applying resampling methods Bootstrap cut-off tests lead to remarkable improvements. The construction of the tests are mainly based on a prespecified weight vector. This enables the selection of powerful tests for alternatives with any false single hypothesis or for alternatives, where at least a given number of false hypotheses contributes to the rejection of the global hypothesis. For the application of these tests in a closed testing procedure some modifications are necessary. Several strategies for the selection of Bootstrap cut-off tests are investigated. Especially for the common problem of all pairwise comparisons or for the case of ordered alternatives the proposed tests regard logical dependencies among false single hypotheses. By Monte Carlo simulations comparisons with Shaffer's step down procedure or with the free step down resampling procedure (Westfall, Young) are given and recommendations for practical use are discussed.

References:

1. Reitmeir P., Wassmer G.: Resampling based methods for the analysis of multiple endpoints in clini-cal trials. Statistics in Medicine 18: 3453-3462 (1999).
2. Shaffer J.P.: Modified sequentially rejective multiple test procedures. JASA 81: 826-831 (1986).
3. Wassmer G., Reitmeir P.,.Kieser M., Lehmacher W.: Procedures for testing multiple endpoints in clinical trials: an overview. Journal of statistical planning and inference 82: 69-81 (1999).
4. Westfall P.H., Young S.S.: Resampling-based multiple testing. Wiley, New York (1993).

#34

Jiayang Sun (Case Western Reserve University, USA)

Multiple Comparisons for Infinite Number of Parameters

When the number of parameters of interest is infinite or extremely large, such as, in the case of quantifying the uncertainty in the estimate of a regression curve or a response surface or a map or an image, itself, the standard multiple comparison methods for a finite number of parameters often lead to an infinite confidence bound or a test too conservative to be useful. In these cases, the methods designed for a continuous domain must be used. The Scheffe's method is a classical approach for such a purpose. It provides a simultaneous confidence bound for a regression function when errors are Gaussian, independent and homoscedastic, and the predictor space is unconstrained, i.e. the domain of interest is the whole q dimensional Euclidean space. In practice, we are often interested in functions defined on an interval or other restricted domains and i in other more general cases than the

Gaussian. Thus the Scheffe's bound is also too conservative or inadequate in these cases and there have been attempts to provide good informative bounds in many important applications. In this talk, I'll introduce some modern techniques for simultaneous inferences and compare them with classical ones and others. Applications include simultaneous confidence bands for linear regression and nonparametric regression with homoscedastic, and heteroscedastic errors, growth and response curves with structured covariance matrices, and generalized linear models. Some "tricks" for these various models will be shown, real data examples and new (free) softwares will be provided.

 # 35

A. J. Sankoh (Wyeth-Ayerst Research, USA)

Non-superiority Clinical Trials and Multiplicity: An Interpretation Issue

Interpreting the efficacy results from active control clinical trials is not easy. This difficulty is compounded when drug efficacy is demonstrated on the basis of clinical evidence from none traditional superiority clinical trials. This is because such efficacy interpretation depends on the quantification of a clinically and statistically acceptable minimal margin of inferiority d by which the effectiveness of the new drug can be reduced and still be viewed clinically relevant and statistically significant compared to no treatment. The quantification of d requires a clear understanding of the data upon which the approval of the active comparator R was based. The general premise for such quantification is that the clinical trials that formed the basis for the approval of the reference active comparator were placebo controlled randomized clinical trials. In other words, the quantification depends on the validity of the assumption that there was a clinically meaningful or sizable and statistically significant treatment difference DR-P between the reference (R) and placebo (P) treatments in the clinical trials on which the approval of R was based. Efficacy interpretation becomes hopelessly even more complicated when such clinical trials designed to demonstrate non-superiority drug effect have multiplicity components (due to multiple endpoints and/or multiple comparisons) in them. Multiplicity due to multiple comparisons could arise when more than one active experimental dose is included in the design and/or more than one clinical efficacy objective is being investigated. We discuss in this presentation the difficulty in interpreting the efficacy results of non-superiority clinical trials in the presence of multiplicity with a special focus on type I error rate and power of the tests.

#36

Nairanjana Dasgupta, Francis G. Pascual (Washington State University, USA)

Exact unconditional tests for comaparing several logistic regression slopes to a standard

In experiments of life sciences application of logistic regression techniques are fairly common. Often it is imperitive to compare the slopes of a series of logistic regressions(arising from applications of different treatments) to that of a standard. We present an exact method for performing this many-to-one comparison based on simple functions of the sufficient statistics without conditioning on the nuisance parameters. We compare our proposed method with asymptotic methods like Reiersol~(1961)based on Minimum Logit Chi-squres, sequentially rejective Bonferroni (Holm,1971) based on Wald statistics and a step-down method based on likelihood ratio tests and show that our method

outperforms its competitors in terms of both Type I errors and marginal power (Spurrier, 1992). The research was motivated by problems in plant pathology and Environmental sciences and we include these as our data examples.

References:
1. Holm, S., (1979). A simple sequentially rejective test procedure. *Scandanavian Journal Statistics* 9, 65-70.
2. Reiersl, O. (1961). Linear and nonlinear multiple comparisons in logit analysis. *Biometrika* 48, 359-365.
3. Spurrier, J. D. (1992). Optimal designs for comparing the variances of several treatments with that of a standard treatment. *Technometrics* 34, 332-339.

 # 37

Gerhard Hommel (University of Mainz, Germany)

Adaptive modifications of hypotheses after an interim analysis

It is investigated how one can modify the hypothesis/es in a study after an interim analysis such that the type I error rate is controlled. If only a global statement is desired, a solution was given by Bauer (1989). If individual statements should be made, the formal application of the closure test may lead to an excessive type I error rate; two proposals for a correction are given. For a general multiple testing problem, by Kieser, Bauer and Lehmacher (1999) and Bauer and Kieser (1999) solutions are given, by means of which the set of hypotheses can be reduced after the interim analysis. If weights for the tests within each of two stages are chosen, the same idea can be applied. Since it is allowed that a hypothesis has weight 0 in the first stage, but a weight > 0 in the second stage, a formal way has been found to include additional hypotheses in the second stage. Nevertheless, the scientific reason of such an inclusion has to be discussed very critically.

References:
1. Bauer,P.(1989). Multistage testing with adaptive designs. Biom. und Inf. in Med. und Biol. 20, 130-136.
2. Bauer,P. and Kieser,M.(1999). Combining different phases in the development of medical treatments within a single trial. Stat. in Med. 18, 1833-1848.
3. Kieser,M., Bauer,P. and Lehmacher,W.(1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. Biom.J. 41, 261-277.
4. Westfall,P.H., Krishen,A. and Young,S.S.(1998). Using prior information to allocate significance levels for multiple endpoints. Stat. in Med. 17, 2107-2119.

 # 38

Brent R. Logan, Ajit C. Tamhane (Northwestern University, USA)

Comparison of Two Treatments Based on Multiple Endpoints

Clinical trials often compare two treatment groups on the basis of multiple endpoints. Frequently, the treatment is assumed to have a one-directional effect on each of the endpoints. In such trials the researcher is interested in establishing not only an overall treatment difference, but also on which endpoints there is a significant treatment effect. In the case where the treatment groups are assumed to have equal covariance matrices, two

methods stand out in the literature: OBriens global OLS test statistic, applied in a closed testing procedure, and Westfall and Youngs (WFY) bootstrap procedure to adjust the single endpoint p-values. In this talk, we investigate further through simulation the properties of these methods. In addition, we propose and compare a hybrid of these two based on the T_max principle of Hothorn. It is concluded that individual p-value adjustments, either through the WFY bootstrap or the hybrid approach, are generally more effective in identifying treatment differences on individual endpoints. Finally, extensions to the unequal covariance matrices case are proposed and compared in a simulation study.

#39

Ajit C. Tamhane, Brent R. Logan (Northwestern University, USA)

Multiple test procedures for identifying the minimum effective and maximum safe doses simultaneously

The therapeutic window is a range of doses of a drug that are both effective and safe. Since generally the efficacy increases with the dose level while the safety decreases, the determination of the therapeutic window reduces to finding the minimum effective and maximum safe doses (MINED and MAXSD). This problem is addressed in the present paper. A bivariate normal model is assumed for the efficacy and safety endpoints. The MINED is defined as the lowest dose that exceeds the mean efficacy of the zero dose by a specified threshold. Similarly the MAXSD is defined as the highest dose that does not exceed the mean toxicity of the zero dose by a specified threshold. Single-step and step-down multiple test procedures are proposed to identify the MINED and MAXSD. These procedures control the type I familywise error probability of declaring any ineffective dose as effective or any unsafe dose as safe at a preassigned level α. The critical points of the exact normal theory procedures depend on the correlation coefficient between the efficacy and safety variables. This difficulty can be side-stepped by using the Bonferroni approximation to the exact critical values which amounts to treating the efficacy and safety testing as two separate families, each with type I familywise error probability controlled at level $\alpha/2$. This approximation is shown to be not very conservative. Another way to avoid this difficulty as well as to relax the assumption of bivariate normality is to use the bootstrap versions of the exact normal theory procedures. The different Bonferroni normal theory and the bootstrap procedures are compared in a simulation study. A real data example is provided to illustrate the procedures.

#40

V. Guiard (FBN Dummerstorf, Germany)

Multiple Test Problems in Detecting of Genes - a small overview

In order to detect quantitative trait loci (QTL), influencing a special trait, for every position of the chromosomes it will be tested whether there exists a QTL or not. For a simple situation, if the test statistic is considered as a function on the position on a chromosome, it varies according to the square of an Orenstein-Uhlenbeck diffusion process. For more complicated situations the overall error will be controlled by means of a permutation test. Both apperoaches assume the overall null hypothesis that there is no Gene influencing the trait of interest. For every inheritable trait there will by be at least one Gene on the genom. Therefore Weller etal applied the concept of the false discovery rate for detecting QTL.

References:
1. Doerge, R.W.;Churchill, G.A. (1996): Permutation tests for multiple loci affecting a quantitative character. Genetics 412: 285-294
2. Lander, E.S.; Botstein, D. (1989): Mapping mendelian factors underlying quantitativetraits using RFLP linkage maps. Genetics 121: 185-199
3. Lander, E.;Kruglyak, L. (1995): Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics 11: 241-247
4. Weller, J. I.; Song, J. Z.; Heyen, D. W.; Lewin, H. A.Ron, M. (1998): A new approachto the problem of multiple comparisons in the genetic dissection of complex traits.Genetics. 150:1699-706

#41

H.P. Piepho (University Kassel, Germany)

Multiple treatment comparisons in linear models when the standard error of a difference is not constant

Users of analysis of variance (ANOVA) procedures are accustomed to an ANOVA table, followed by a table of means. When the underlying fixed effects linear model is variance-balanced, i.e. the standard error of a difference is constant for all pairwise comparisons, non-significant differences can be indicated by underlining. Unfortunately, when the design is unbalanced, i.e. the standard error of a difference is not constant over pairs of treatments, it may turn out to be impossible to consistently represent significant differences by underlining. The same problem occurs, e.g., in linear mixed models and in generalized linear models. This paper proposes a simple, conservative approach, which allows a lines-representation of treatment comparisons. The price for the improved display of results is a potential loss of significances, though a loss of more than one significance is rarely observed in practice. Very frequently, there is no such loss at all. Lost significances may be reported separately.

References:
1. Piepho HP 1999 Multiple treatment comparisons in linear models when the standard error of a difference is not constant. submitted

 # 42

James J. Chen (FDA, USA)

Weighted P-Value Adjustments for Animal Carcinogenicity Trend Test

A typical animal carcinogenicity experiment routinely analyzes approximately 10-30 tumor sites. This paper proposes using weighted adjustments by assuming that each tumor can be classified as either Class A or Class B based on prior considerations. The tumors in Class A, which are considered as more critical endpoints, are given less adjustment. Two weighted methods of adjustments are presented: the weighted p adjustment and weighted α adjustment. The power to detect a dose effect increases if a treatment-dependent tumor is analyzed as in Class A tumors, and the power decreases if it is analyzed as in Class B tumors. A data set from an National Toxicology Program (NTP) two-year animal carcinogenicity experiment with thirteen tumor types/sites observed in male mice was analyzed using the un-weighted and weighted methods. The un-weighted adjustment concluded that there was no statistically significant dose-related trend. Using the FDA classification scheme for the

weighted adjustment analyses, two rare tumors (with background rates of 1% or less) were analyzed as Class A tumors and eleven common tumors (with background rates higher than 1%) as Class B. Both weighted analyses showed a significant dose-related trend for one rare tumor.

 # 43

Daniel T. Voss (Wright State University, USA)

Analysis of unreplicated factorial experiments

We are interested in methods of analysis of unreplicated factorial experiments which provide strong control of error rates. Small fractions of 2^k factorial experiments are useful for screening many factors when few non-negligible effects are anticipated. Such screening experiments often utilize designs which are nearly saturated, saturated, or super-saturated, providing few or no error degrees of freedom. Lacking an independent variance estimator, a saturated design can be analyzed by comparing the relative magnitudes of either the normalized parameter estimates or the corresponding sums of squares. Many methods have been proposed for the analysis of saturated designs (see Hamada and Balakrishnan, 1998). Proposed methods are increasingly reflecting methods and ideas from multiple comparisons, but few of the methods are known to control error rates over all parameter configurations. Kinateder, Voss and Wang (2000) reviewed methods known to control error rates and discussed related open problems. In this talk, we will provide a current review of open problems and known results concerning the control of error rates in the analysis of saturated, nearly saturated, and super-saturated designs.

References:

1. Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: A review with some new proposals. Statistica Sinica, 8, 1--41.
2. Kinateder, K. K. J., Voss, D. T., and Wang, W. (2000). Analysis of saturated and super-saturated factorial designs: A review. In N. Balakrishnan (Ed.), Proceedings of the Indian International Statistical Association 1998 International Conference. Newark, New Jersey: Gordon and Breach. In press.

 # 44

Gudrun Bernhard (Novartis Pharma AG, Switzerland),
Markus Klein, Gerhard Hommel (University of Mainz, Germany)

Global and multiple test procedures using ordered p-values

We looked at global and multiple tests for the combination of n hypotheses that are based only on the ordered p-values of the individual tests for each of the n hypotheses.Two different situations were considered:

- Arbitrary dependencies among the test statistics; for this situation, Röhmel and Streitberg (1987) provided a general class of global tests.
- Independent test statistics; for this situation, a general class of global tests was described by Kornatz (1994) using recursive formulas.

Multiple test procedures that are based on global tests can be developed using the concept of critical matrices for closed tests (Wei Liu, 1996).

References:
1. Klein,M. (1998). Multiple Testprozeduren: Step-down und Step-up. Diploma thesis, Mainz.
2. Kornatz,C. (1994). Allgemeine Schrankentests und ihre Anwendung bei aufeinanderfolgenden Studien. Diploma thesis, Mainz.
3. Liu, Wei (1996). Multiple tests of a non-hierarchical finite family of hypotheses. JRSS B 58, 455-461.
4. Röhmel,J., Streitberg,B. (1987). Zur Konstruktion globaler Tests. EDV in Med. und Biol. 18, 7-11.

 # 45

F. Bretz (University of Hannover, Germany),
A. Genz (Washington State University, USA),
L.A. Hothorn (University of Hannover, Germany)

Studentized multiple contrast tests: distribution under the null and the alternative

We consider multiple comparison procedures with and without order restrictions. Such procedures include the tests of Williams, Hayter, Hirotsu, Rom et al., and many others. One main problem of all of these approaches is their limited numerical availability under the null and the alternative hypotheses. We show that all of the above approaches can be represented as studentized multiple contrast tests. We introduce new methods of computing the arising central and non-central multivariate t-distributions. Numerical issues are discussed and the new approaches are compared to existing ones. The results indicate that we are able to compute p-values (and hence quantiles) and power values (and hence sample sizes) robustly and reliably at moderate accuracy levels (about 4 significant digits) within a few seconds of workstation time for multiple comparison problems with up to 20 groups. A data example illustrates the application and the availability of the presented techniques.

References:
1. Bretz, F. (1999) Powerful Modifications of Williams' Test on Trend. Ph.D. thesis, University of Hannover.
2. Genz, A. and Bretz, F. (1999) Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation, 63, 361-378
3. Genz, A. and Bretz, F. (2000) Methods for the Computation of Multivariate t-Probabilities. (submitted)
4. Hothorn, L.A., Neuhäuser, M. and Koch, H.F. (1997) Analysis of randomised dose-finding studies: closure test modifications based on multiple contrast tests. Biometrical Journal, 39, 467-479.

#46

Klaus Straßburger, Guido Giani, Helmut Finner (Deutsches Diabetes-Forschungsinstitut an derHeinrich-Heine-Universität, Germany)

Stepwise Partitioning Procedures

Within a one-way layout with normally distributed observations Tong (1969) proposed a multiple decision procedure for partitioning a given set of treatments into two subsets with the purpose of separating good and bad treatments. The qualities good and bad are defined in terms of the mean difference from a control. Tong's procedure is optimal in the sense that it maximizes the minimum probability of a correct partition (MPCP) within the class of so-called natural procedures (Giani and Straßburger, 1997). In this contribution we show that the optimality of Tong's procedure cannot be extended to a larger class of multiple decision rules. In fact, a non-natural partitioning procedure will be presented, which leads to a greater MPCP than Tong's procedure. Although the new procedure has a stepwise structure, it substantially differs from the well-known step-up and step-down selection procedures used for comparisons with a control. After a discussion of the theoretical results, the minimum total sample sizes necessary for Tong's and the new stepwise procedure to control the probability of correct partition at a preassigned confidence level P are compared with each other. It turns out that in practically relevant situations ($P \geq 95\%$) Tong's procedure has nearly the same efficiency as the new decision rule.

References:
1. Tong, Y.L. (1969): On Partitioning a Set of Normal Populations by Their Locations with Respect to a Control. Annals of Mathematical Statistics 40, 1300-1324.
2. Giani, G., Straßburger, K. (1997): Optimum partition procedures for separating good and bad treatments. Journal of the American Statistical Association 92, 291-298.

#47

Egbert Biesheuvel (Solvay Pharmaceuticals, The Netherlands), Ludwig Hothorn (University of Hannover, Germany)

Many-to-one comparisons in a stratified design maintaining the overall alpha level

Dunnett (1955) described a multiple comparison procedure for many-to-one comparisons in the one-way layout assuming normal distributed data. Cheung and Holland (1991) extented the Dunnett procedure to the situation of a stratified design. In this latter situation, the correlation matrix has a block product-moment structure. Nowadays different algorithms exist to handle multivariate t-distributions. Percentage points, confidence intervals and power can be computed/calculated within SAS, even in case of unbalanced data. The performance of this method in comparison to resampling techniques (bootstrap, permutation and parametric simulation) is investigated for different kind of data. The bootstrap and permutation techniques are calculated within PROC MULTTEST and the Edwards and Berry %-rejection method within PROC MIXED. In addition, a proposal of a non-parametical technique for many-to-one comparisons in a stratified design will be discussed. Extentions to a step-down procedure and how to proceed in situations with a non product-moment correlation structure will be briefly mentioned.

References:
1. Cheung, S.H. and Holland, B. Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics, 47, 21-32 (1991)
2. Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1096-1121 (1955)
3. Edwards, D. and Berry, J.J. The efficiency of simulation-based multiple comparisons. Biometrics, 43, 913-928 (1987)
4. Genz, A. and Bretz, F. Numerical computation of multivariate t-probabilities with application to power calculations of multiple contrasts. Working paper (1999)

#48

I.Novikov (The Chaim Sheba Medical Center, Israel)

Multiple comparisons in one applied ranking problem.

In the talk we consider the use of multiple comparisons procedures in the following applied statistical problem. Given the results of customers survey for a large company with a network of branches, we want to identify a) the three "best" branches, which will get bonuses; b) all the branches, where the results of the survey were substantially below the median level, which have to be additionally investigated and possibly re-organized. To distribute bonuses among the three "best" branches in a fair way, we propose the procedure for testing differences among them, which takes into account that they are the three best points among a given number of observations.The number of the 'worst' branches is not fixed from the beginning and must be determined via some decision making procedure, which also uses the distribution of extreme observations.

#49

Michael Weichert (Cap Gemeni, Germany)

Determining all logical dependencies among pairwise hypotheses using graphs

In a one-way layout one often is interested in the pairwise copmarison of some or all the treatments. When conducting more than one test on the data one is confronted with the problem of substaining the multiple level of the experiment. One way of controlling the multiple level is to use the closure testing principle. When constructing the closure testing system one has to determine the logical dependecies among the hypotheses. The logical dependecies are of two kinds. First the intersection of some hypotheses may include some other ones. (e.g. $\mu_1 = \mu_2 \cap \mu_2 = \mu_3 \rightarrow \mu_1 = \mu_3$) Second an intersection hypothesis could be empty, so it is not part of the closure test system. (e.g. $\mu_1 \geq \mu_2 \cap \mu_2 \geq \mu_3 \cap \mu_3 > \mu_1$) The problem of dertermining these dependencies is solved by mapping the relevant intersection hypotheses onto a graph. When only twosided hypotheses are of interest, an undirected graph is used, otherwise a directed graph is used. The results can be extended to handle systems of hypotheses including shifted hypotheses and equivalence hypotheses.

References:
1. Berhard, G. (1991), Computergestützte Durchführung von multiplen Testprozeduren - Verbesserte Algorithmen und Powervergleich-, Dissertation, Universität Mainz

2. Hommel, G. (1999), Concepts for the description of logical relationships among hyotheses, Vortrag auf der AG Sitzung "Multiple Verfahren", Mainz
3. Marcus, R., Perlitz, E., Gabriel, K.R. (1976), On closed testing procedures with special reference to ordered analysis of variance, Biometrica 63
4. Weichert, M. (2000), Robuste Mittelwertvergleiche mit gartenbaulichen Anwendungen, Dissertation, Universität Hannover

 # 50

Anat Reiner (Tel Aviv Univeristy, Israel)

Using the False Discovery Rate Criteria for Simultaneous Hypothesis Testing in Epidemiological Research

Some well established statistical methods have been broadly used for the purpose of analyzing epidemiololgical data. The methods selected depend mostly on the data structure and the data collecting method, ignoring possible insufficiency of the procedure under certain circumstances. An analysis containing multiple comparisons is an example to a situation in which cautious considerations need to be made before applying an analytical technique and interpreting its results. This is due to the increased type I error arisen by simultaneously performing multiple statistical tests, and the possible loss of power that might occur as a result of attempting to control the increase of the type I error. In search of relevant cases for implementation of the FDR criterion for multiple comparisons, we attempted to identify the typical statistical procedures applied for dealing with problems addressed by researchers of epidemiology, through a survey of randomly sampled articles out of the 1993 to 1995 volumes of the American Journal of Epidemiology and the American Journal of Public Health. It was recognized from the survey that one of the most widely used analytical tools is the multiple logistic regression model fitting procedure, that aims to predict the probability of attaining a cerain medical condition, and also produces estimates of the odds ratios for the subgroups of interest, therefore involving multiple hypotheses testing. It was therefore concluded that focusing the discussion and analysis of the multiple testing problem in the cases where a logistic regression procedure is applied will yield quite a good coverage of the problem as it faced by epidemiological reaearch activity. Ottenbacher(1998), who analyses the size of the type I error in a sample of published epidemiological articles, enhances the need to apply procedures that deal with multiple comparisons, and suggests reducing the significance level by using a more conservative criteria, that will take multiplicity into account. He mentions the Bonferroni method as an example, with the drawback of its resulting in a drastic loss of power. He mentions the Benjamini and Hochberg method (1995) with a similar drawback, and suggests the alternative of using a less conservative criteria than the FWER (Family-wise Error Rate). In fact Ottenbacher fails to recognize that the Benjamini and Hochberg method adopts exactly the same idea: using a less conservative criteria that still provides sufficient information concerning the type I error. Moretheless, the criteria it suggests to control, the FDR (False Discovery Rate), which is the expected rate of false rejections, is structurally defined and theoretically supported. On this ground it became worthwhile to study the performance of methods that control the FDR in different scenarios that represent the various data structures confronted in epidemiological research. For this purpose, simulative databases were created, containing multiple explanatory variables and one dichotomeous dependent variable. Each database was defined using a unique combination of characteristics, such as sample size, number of multiple hypotheses, proportion of false hypotheses, extent of significance and type of dependency between the test statistics. Overall, 48 different data configurations were created. the odds-ratios from 1 were calculated. The

hypotheses were tested using each of 10 different methods to set a corrected significance level, given a desired type I error. 5 of the methods controlled the FWER, and 5 of them controlled the FDR. Data was repeatedly simulated and modeled 2500 times, for each type of configurations. Averages and standard deviations were calculated for the FDR and the test power. These measurements were used to thoroughly investigate the performance of each method. The performances of the methods that control the FDR were compared to the methods that conrol the FWER, and also compared against each other. Results show a consistent advantage of the methods that control the FDR in terms of test power. The optimal data characteristics in terms of power gain that resulted from using the FDR criteria is a high proportion of false hypotheses, accompanied by a high total number of hypotheses and a low significance of them, in case of independence or positive dependence, and a high significance in situations of a relatively low power, as in the case of general dependency. In case of independence or positive dependence, the Benjamini and Hochberg original method, and their later developed adaptive method, always achieve the best results in terms of absolute power, and sesitivity of power to conditions that yield low power by definition. In case of general dependence, the Benjamini and Liu method always achieves the best results in terms of absolute power.

#51

Paul N. Somerville (University of Central Florida, USA),
Frank Bretz (Hannover University, Germany)

Obtaining Critical Values for Simultaneous Confidence Intervals andMultiple Testing

Fortran 90 and SAS-IML programs are presented which enable a user to obtain, on demand, critical values for 15 different multiple contrast procedures, some of which are as yet unpublished. In addition, critical values can be calculated for procedures corresponding to any specified set of contrasts. The estimates of population may be correlated, provided the estimated variance covariance of the means is included in the input. The programs are not limited to the randomized on-way layout but are applicable to procedures for which the estimates are obtained by multiple regression, and include incomplete block designs and missing value cases. Accuracies of approximately 3 decimal places may be obtained in 2 or 3 seconds using any Pentium Processor.

#52

Guido Giani, Klaus Straßburger, and Helmut Finner (Deutsches Diabetes-Forschungsinstitut an der Heinrich-Heine-Universität Düsseldorf, Germany)

Separate - A Program Package for Multiple Comparisons

Separate is a menu-driven software package devoted to designing and analysing experiments for multiple comparisons with the "best" and with a control. In addition, support is given in certain problems of testing for equivalence and for difference. For normally distributed data three designs, the one-way layout, the randomized bock design, and the crossover design without carry-over effects, are supported by the program. For each of the two options of determining sample sizes (option I) and calculating probability levels of correct decisions being achieved with given sample sizes (option II), the experimenter has to specify threshold values to characterize treatments as good, bad, or equivalent to the control. To cover the

unknown variance case, the treatment qualities good, bad, or equivalent are defined in terms of standardized mean differences from the best or the control. Besides the classical indifference zone approach of Bechhofer (1954), the subset selection formulation of Gupta (1956) supplemented by additional power requirements, and further related approaches, the problem of discriminating between good and bad treatments and, if intended, those being equivalent to the control is dealt with. Option I facilitates simultaneous as well as seperated control of all the kinds of multiple errors at designated levels, whereas under option II the respective minimum multiple error probabilities being achieved for given sample sizes are numerically calculated (Giani and Straßburger 1997, 2000). For most of the implemented decision rules the program also gives the least favorable parameter configuration at which the minimum probability of correct selection or correct discrimination is attained. For the discrimination problems this configuration is the solution of a complex optimization task and depends in general on the parameters of the underlying procedure and all the specifications made in advance. Finally, it should be mentioned that also single-step and step-down procedures are implemented for various subset selection objectives. Besides this, the software offers facilities to handle the described discrimination problems under distribution models with scale parameter. At the present, for the one-way layout procedures for discriminating with respect to variances under the normal model and with respect to incidences in exponentially distributed data are available.

1. References:
   Bechhofer, R.E. (1954): A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations With Known Variances. Annals of Mathematical Statistics 25, 16-39.
2. Gupta, S.S. (1965): On Some Multiple Decision (Selection and Ranking) Rules. Technometrics 7, 225-245.
3. Giani, G., Straßburger, K. (1997): Optimum Partition Procedures for Separating Good and Bad Treatments. Journal of the American Statistical Association 92, 291-298.
4. Giani, G., Straßburger, K. (2000): Multiple Comparison Procedures for Optimally Discriminating Between Good, Equivalent, and Bad Treatments With Respect to a Control. Journal of Statistical Planning and Inference 83, 413-440.

#53

Helmut Finner (Deutsches Diabetes-Forschungsinstitut an der Heinrich-Heine-Universität Düsseldorf, Germany),
Markus Roters (Universität Potsdam, Germany)

Multiple hypotheses testing and expected type I errors

In this paper we investigate the behaviour of the expected number of type I errors EV_n (say) of multiple test procedures for n hypotheses $H_1, ..., H_n$, where the random variable V_n denotes the number of type I errors. Special attention will be focussed on procedures controlling a multiple level α and the case that all hypotheses are true. We consider (i) single-step, step-down and step-up procedures based on independent p-values, (ii) test procedures based on exchangeable test statistics and (iii) test procedures based on the range statistics. The behaviour of EV_n will be studied especially for the case that n tends to infinity.

References:
1. Finner, H. Roters, M.(1994).On the limit behaviour of the joint distribution function of order statistics. The Annals of the Institute of Statistical Mathematics 46, 343-349.

2. Finner, H.& Roters, M.(1998) Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. The Annals of Statistics 26, 505-524.
3. Finner, H.& Roters, M.(1999).Asymptotic comparison of the critical values of step-down and step-up multiple comparison procedures.Journal of Statistical Planning and Inference 79, 11-30.
4. Finner, H.& Roters, M (2000).On the critical value behaviour of multiple decision procedures. Scandinavian Journal of Statistics. 27, 1-11.
5. Finner, H.& Roters, M. (2000). Asymptotic sharpness of product-type inequalities for maxima of random variables with applications in multiple comparisons.Statistica Sinica, in revision.
6. Finner, H.& Roters, M.(2000).Multiple hypotheses testing and expected type I errors.t In preparation.
7. Spjotvoll, E. (1972). On the optimality of some multiple comparison procedures. The Annals of Mathematical Statistics 43, 398-411.

#54

C. Hirotsu, S. Aoki (University of Tokyo, Japan)

Test for the association between the disease and alleles

The association analysis between the disease and alleles is one of the simple methods for localizing the susceptibility genes. For revealing the association, several statistical tests have been proposed without discussing explicitly the alternative hypotheses. We therefore specify two types of alternative hypotheses: 1. there is only one susceptibility gene in the locus; 2. there is an extension or shortening of alleles associated with disease, and derive exact maximal chi-squared type tests for the respective hypotheses. We also propose to combine those two tests when the prior knowledge in not sufficient enough to specify one of those two hypotheses. Inparticular those ideas are extended to the three-way association between the disease and bivariate allele frequencies at two closely linked loci.

#55

Yoav Benjamini (Tel Aviv University, Israel)

The multiplicity problem in scientific research: what is being done about it - and what could be done instead.

The multiplicity problem will be reviewed in some areas of scientific research, where it raises special difficulties. I shall give examples of the way by which the problem is being handled currently - mostly by ignoring it. Dificulties with traditional FWE controlling procedures which might cause this response will be discussed. Recent developments within the FDR control approach will be presented, making this approach especially atractive for these scientific areas. In some cases FWE control is crucial - so I shall discuss how to combine FDR control and FWE control for that purpose.

 # 56

Sanat K. Sarkar (Temple University, USA)

False Discover Rate of a Generalized Step-Up-Down Multiple Testing Procedure

This paper provides a theoretical understanding of how and why the notion of False Discovery Rate (FDR) works in a general stepwise multiple testing procedure. Also, it broadens the scope of the FDR by covering not only procedures that are more general than a step-up or step-down procedure but also situations where the test statistics are not necessarily independent.

References:

1. Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society, Ser B*, 57, 289-300.
2. Sarkar, S. K. (1998), Some Probabilty Inequalities for Ordered MTP_2 Random Variables: A Proof of the Simes' Conjecture, *Annals of Statistics*, 26, 494-504.
3. Tamhane, A.C., Liu, W, and Dunnett, C. W. (1998). A Generalized Step-Up-Down Multiple Test Procedure, *Canadian Journal of Statistics*, 26, 353-363.

#57

Y. Benjamini, Y. E. Kling (Tel Aviv University, Israel)

Aspects of Multiplicity in Statistical Process Control (SPC)

The "Multiplicity Problem", is reviewed in the context of Statistical Process Control (SPC). Disregard of this issue, as is common practice, results in an inflated rate of false alarms. We examine the appropriateness of two overall error measures: the Familywise Error Rate (FWE) and the False Discovery Rate (FDR). We discuss a few selected SPC configurations that give rise to multiplicity: Multiple criteria on the same chart, controlling several aspects of a process, controlling multiple attributes of a product, and controlling the quality of the final product

#58

Martin Posch and Peter Bauer (University of Vienna, Austria)

Interim Analysis and Sample Size Reassessment

This paper deals with the reassessment of the sample size for adaptive two stage designs based on conditional power arguments utilizing the variability observed at the first stage. Fisher's product test for the p-values from the disjoint samples at the two stages is considered in detail for the comparison of the means of two normal populations. We show that stopping rules allowing for the early acceptance of the null hypothesis which are optimal with respect to the average sample size may lead to a severe decrease of the overall power if the sample size is a priori underestimated. This problem can be overcome by choosing designs with low probabilities of early acceptance or by mid-trial adaptations of the early acceptance boundary using the variability observed in the first stage. This modified procedure is negligibly anti-conservative and preserves the power.

References:
1. Bauer, P. and Köhne, K. (1994). Evaluations of experiments with adaptive interim analyses. Biometrics, 50, 1029--1041.
2. Birkett, M.~A. and Day, S.~J. (1994). Internal pilot studies forestimating sample size. Statistics in Medicine, 13, 2455--2463.
3. Browne, R.~H. (1995). On the use of a pilot sample for sample size determination. Statistics in Medicine,14, 1933--1940.
4. Cui, L., Hung, H. M.~J., and Wang, S. (1999). Modification of sample size in group sequential clinical trials. Biometrics, 55,321--324.
5. Kieser, M. and Friede, T. (1999). Re-calculating the sample size of clinical trials in internal pilot studies with control of the type {I} error rate. Statistics in Medicine, to appear
6. Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. Biometrics, 55, 1286-1290
7. Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. Biometrical Journal, 41, 689--696.
8. Proschan, M.~A. and Hunsberger, S.~A. (1995). Designed extension of studies based on conditional power. Biometrics, 51, 1315--1324.
9. Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. Statistics in Medicine, 9, 65--72.

#59

Jørgen Hilden, Michael Weis Bentzon (University of Copenhagen, Danmark)

Bonferroni with contextual P-value transforms: a means to gain power

Consider a k-faced null hypothesis, the jth subtest P-value being P_j. Bonferroni corrections (BC), when needed, are wasteful of power. Our present aim is to suggest a method of improving power by exploiting the subject-matter context: we deemphasize those subtests that provide little hope of revealing something interesting because the standard error (SE_j) is large compared with the largest effect that might realistically exist (K_j). The standard BC procedure gives an Overall P-value, or Overall Attained Significance Level, $OASL_{BC} = k*\min_j\{P_j\}$. As an extension, let $f_j(.)$ be an increasing, preferably continuous, function through the origin, and $g_j(.)$ its inverse. When applied to summary statistic $T = \min_j\{f_j(P_j)\}$, the Bonferroni inequality implies $P_0\{T =< t\} =< SIGMA_J\ g_J(t)$, so the associated OASL = $SIGMA_J\ g_J(\min_j\{f_j(P_j)\})$. This OASL reduces to $OASL_{BC}$ when the transform functions are all the same (or $k = 1$). Now, for some index j, the context may indicate that the realistic alternative to H_{0j}: $m_j = 0$ is $H_{(alt)j}$: $0 < m_j < K_j$, K_j being small relative to the associated SE_j. In most models the likelihood ratio, LR_j(data y) = $p\{y|H\_0j\}$ / $\sup_{(alt)j}\{p\{y|m_j\}\}$, is a natural starting point for combined inference and implicitly is a monotone function of (the conventional one-sided) Pj. This suggests defining $T = \min_j\{LR_j\}$ and using the extended Bonferroni rule. The (one-sided) BC procedure is reproduced as long as K_j's are effectively infinite. As K_j/SE_j approaches zero for a certain part of the j's, i.e. as the hope of useful information about these m_j's dwindles, the procedure focusses on the remaining, fewer, subtests, thereby recuperating power. The extended Bonferroni scheme can be built into sequential rejection schemes.

 # 60

IDENTIFYING EFECTIVE AND SAFE TREATMENTS

P. Bauer, W.Brannath and M.Posch
(Department of Medical Statistics, University of Vienna, Austria)

We consider the situation where k treatments and a (zero) control are compared with respect to efficacy and safety. For efficacy the null hypotheses for the many one comparisons in terms of the paramenter of interest are defined as H_{oi}^{E}: $\mu_i \leq \mu_0$, $i = 1, ..., k$. Here μ_0 and $\mu_1, ..., \mu_k$ denote the parameter under the control (μ_0) and the k treatments, respectively. For safety the shifted one sided null hypotheses H_{oi}^{S}: $\theta_i \leq \theta_0 + \delta$ are investigated, where δ is the prefixed safety margin for the corresponding parameter of interest.
A treatment i is considered to be effective if H_{oi}^{E} is rejected and is considered to be safe if H_{oi}^{S} is rejected. If both H_{oi}^{E} and H_{oi}^{S} are rejected it is considered to be effective and safe.
By considering only the k sub-families (H_{oi}^{E}, $H_{oi}^{E} \cup H_{oi}^{S}$), $i = 1, ..., k$, the multiple levels applied within the sub-families can be adjusted in a stepwise way. Within the sub-family a hierarchical procedure with a fixed sequence of testing is used. This multiple level-procedure can also be applied to the problem of simultaneously establishing superiority of a treatment to a (zero) control and Δ-equivalence to an active control and is more powerful than the procedure by Bauer et al. (1998). If order restrictions are assumed to hold among the parameters of interest a split strategy by applying adjusted multiple levels within the two sub-families (H_{oi}^{E}, $i = 1, ..., k$) and (H_{oi}^{S}, $i = 1, ..., k$) can be applied. If all treatments are found to be effective or all treatments are found to be safe this leads to an improvement of the Bonferroni-splitting.
A possible generalization to continuous families with the corresponding confidence intervals is given.

Bauer, P., Röhmel, J., Maurer, W. and Hothorn, L. (1998) Testing strategies in multi-dose experiments including active control. Statist. Med., 17, 2133-46.

(Notice: The abstract numbers are time-sequentially)