Relevance-shifted tests for high dimensional data with small sample sizes

Von der Naturwissenschaftlichen Fakultät
der Universität Hannover
zur Erlangung des
akademischen Grades eines

Doktors der Gartenbauwissenschaften
- Dr. rer. hort. -

genehmigte

Dissertation

von

Dipl.-Ing. agr. **Cornelia Frömke**
geboren am 10.01.1977 in Hannover

2006

Referent:	Prof. Dr. Ludwig A. Hothorn
Korreferent:	PD Dr. Siegfried Kropf
2. Korreferent:	PD Dr. Frank Bretz
Tag der Promotion:	27.01.2006

Wenn ich weiß,
was Liebe ist,
ist es wegen Dir.
H. Hesse

Kurzfassung

In der Analyse von Microarray Daten ist es in der Forschung häufig von Interesse, Gene zu finden, welche sich zwischen zwei Behandlungsgruppen um ein bestimmtes k-faches in ihrer Aktivität unterscheiden. Oft wird ein Gen als signifikant erachtet, wenn die Punkt-Nullhypothese, dass kein Behandlungseffekt vorliegt, abgelehnt werden kann und gleichzeitig der Quotient der Behandlungsmittelwerte oder -Mediane das k-fache übersteigt. Dies ist statistisch nicht ausreichend. Stattdessen ist es angebracht, direkt die relevanzgeshiftete Hypothese, dass der Behandlungseffekt das k-fache nicht übersteigt, zu testen. Ziel dieser Arbeit ist die Entwicklung eines Testverfahrens, welches mittels eines Zweistichprobentests je Gen diese relevanzgeshiftete Hypothese überprüft.

Für die Entwicklung eines Testverfahrens muss die besondere Datenkondition von Microarrays beachtet werden. Aus ökonomischen Gründen ist die Anzahl an Beobachtungen in Microarray Datensätzen gering; eine Fallzahl von 8 oder weniger ist nicht ungewöhnlich. Neben der eingeschränkten Anzahl an Wiederholungen werden Tausende von Endpunkten simultan getestet. Diese hohe Dimensionalität führt zu einem extremen Multiplizitätsproblem. Auch kann nicht immer von gauss- oder lognormal-verteilten Daten ausgegangen werden. In der vorliegenden Arbeit wird eine Testprozedur vorgestellt, welche robust gegen Abweichungen der Gauß- oder Lognormalverteilung ist. Für spezielle Datenkonditionen weist das Verfahren eine hohe Güte auf. Dies ist essentiell für Microarrays, da durch eine fälschliche Nichtsignifikanz ein mögliches Kandidatengen von einer weiteren genetischen Analyse ausgeschlossen sein kann. Das Erzielen einer hohen Güte ist besonders dann eine Herausforderung, wenn die Fallzahlen gering sind, die Datenverteilung schief ist und ein multiples Testverfahren angewandt wird. Um diese Probleme zu überwinden, nutzt das Verfahren relevanzgeshiftete Tests, eingebunden in eine Prozedur mit datengeordneten Hypothesen. Diese Klasse von Prozeduren erwies sich als mächtig in der Analyse von Microarrays.

Neben dem nichtparametrischen relevanzgeshifteten Quotiententest mit datengeordneten Hypothesen werden auch zwei parametrische Versionen vorgestellt. Zusätzlich wird ein

nichtparametrischer Permutationsalgorithmus mit Quotiententests vorgeschlagen. Weiterhin werden zwei parametrische und ein nichtparametrisches Analogon für relevanzgeshiftete Tests auf Differenz beschrieben.
Für alle Prozeduren werden detaillierte Gütestudien im Vergleich mit Standardverfahren für die Analyse von Microarrays gezeigt. Alle neuen Methoden, welche in der Arbeit vorgestellt werden, sind Approximationen. Daher werden detaillierte Simulationsergebnisse der globalen Fehlerrate präsentiert.
Die in dieser Arbeit verwendeten Funktionen und Beispieldatensätze sind in einem R Paket verfügbar.

Schlagworte: k-fach, nichtparametrisch, Quotiententest, relevanzgeshiftete Hypothese

Abstract

In the analysis of microarray data frequently the research interest is to find genes which differ in their expression activity by a specific k-fold among two treatment groups. Although it is conventional practice, it is not sufficient to test the point-zero null hypothesis of no treatment effect and the ratio of treatment means or medians has to exceed the k-fold. Instead a relevance-shifted hypothesis has to be tested, that the treatment effect does not exceed the k-fold of interest. The aim of this work is the development of a testing procedure, which includes two-sample tests analyzing this relevance-shifted hypothesis for each gene.

To construct a testing procedure, the special data condition of microarrays has to be taken into account. For economical reasons microarray data tends to have a small number of observations; a sample size of 8 or less is not unusual. Besides the lack of repetitions, thousands of endpoints are tested simultaneously. This high-dimensionality leads to a massive multiplicity problem. Finally, a Gaussian or lognormal population distribution cannot always be assumed. In this thesis a testing procedure is proposed, which is robust against deviations from the Gaussian or lognormal distribution. Given certain data conditions, it provides a high power. This is essential for microarray data, as a miss of a possible candidate gene is severe as non-significant genes may be lost in further genetical analysis. The achievement of a high power is a challenge, if the sample size is small, the data distribution is skewed and a multiple testing method is applied. To overcome these problems, the new method uses a relevance-shifted test embedded in a procedure with a data-driven order of hypotheses. This class of procedures proved to be powerful for microarray data.

Besides the nonparametric relevance-shifted test on ratio with a data-driven order of hypotheses two parametric versions are given as well. Further a nonparametric permutation algorithm to test for a relevant ratio is presented. Finally two parametric and one nonparametric analog to test for a relevant difference are proposed.

For all procedures detailed power simulations in comparison with standard methods for the analysis of microarray data are shown. And as all new methods presented in this work

are approximations, detailed simulation results of the familywise error rate are given. An R-package with the functions and example data sets used in this work is available.

Keywords: k-fold, nonparametric, test on ratio, relevance-shifted hypothesis

Contents

List of Figures

Notations

α	Type I error rate
β	Type II error rate
FWER	familywise error rate
H_0	null hypothesis
H_1	alternative hypothesis
$i = 1, 2$	group index
$j = 1, \ldots, m$	endpoint index
$k = 1, \ldots, n_i$	repetition index
μ, σ, Σ	expected value, true standard deviation and covariance matrix
$\rho_{ijj'}$	true correlation among the endpoints of group i
$\bar{x}, s, S$	estimate of expected value, standard deviation and covariance matrix
$\tilde{x}$	empirical median
IQR	interquartile range
w_j	selector statistic
$\delta_{lower}, \delta_{upper}$	lower and upper relevance threshold in terms of difference
$\theta_{lower}, \theta_{upper}$	lower and upper relevance threshold in terms of ratio

Procedures with a data-driven order of point-zero hypotheses:

p-selector	parametric procedure
np-selector	nonparametric procedure

Procedures with a data-driven order of relevance-shifted hypotheses to test on difference:

shift selector	parametric procedure without a data transformation
δ-shift	parametric procedure with δ-shifted endpoints
random$^{\delta}$	parametric procedure with randomly generated data
np-δ-shift	nonparametric procedure with δ-shifted endpoints

Procedures with a data-driven order of relevance-shifted hypotheses to test on ratio:

Sasabuchi selector	parametric procedure without a data transformation
θ-shift	parametric procedure with θ-shifted endpoints
random$^{\theta}$	parametric procedure with randomly generated data
np-θ-shift	nonparametric procedure with θ-shifted endpoints

Other procedures:

minP	nonparametric relevance-shifted permutation algorithm for step-down minP adjusted p-values (for tests on relevant ratios)

Chapter 1

Introduction

In recent years biotechnological research has made it possible to obtain information about the activity of genes in a cell. This activity can be measured quantitatively in terms of expression levels by use of a technique called microarrays. Such a method to monitor the expression levels of thousands of genes simultaneously is applied for various experimental questions. Agricultural experimental questions include for example the local and systemic response of native tobacco *Nicotiana attenuata* to various herbivores (Heidel and Baldwin (2004)) or the response of tomato tissue to derivatives of the phytotoxin coronatine (Uppalapati *et al.* (2005)). In human medicine microarrays are used to find differentially expressed genes among for example two types of a specific cancer, such as differences between acute myeloid leukemia and acute lymphoblastic leukemia (Golub *et al.* (1999)). Other experimental questions of interest can be found in dermatology. For example studies of inflamatory skin diseases such as psoriasis are published (Kunz *et al.* (2004)).
Experiments are mainly done by one of two types of platforms: the oligonucleotide array GeneChip (Affymetrix, Santa Clara, CA) and the cDNA microarray proposed by Stanford University. The following brief introduction to the functionality of a microarray is given by focussing on oligonucleotide arrays. To measure the gene expression of a test subject, a messenger ribonucleic acid (mRNA) sample is taken. Passing through several processes the mRNA is transformed to complementary RNA (cRNA), labelled with a fluorescent dye

and applied on the array. On the surface of the array sequences of single-stranded deoxyribonucleic acid (DNA) called oligonucleotides corresponding to several thousands of genes are spotted via photolithography. The single-stranded cRNA from the test sample binds to the complementary oligonucleotides and the quantity of the gene expressions is then obtained by measuring the intensities of the fluorescent dye (Lockhart *et al.* (1996)). To achieve microarrays which are comparable to one another, these intensities have to be normalized in terms of for example background correction and variance stabilization (see Irizarry *et al.* (2003), Bolstad *et al.* (2003) or Geller *et al.* (2003) for details on normalization of oligonucleotide arrays). Finally a statistical analysis has to be applied to the normalized data.

The aim of this work is to find differentially expressed genes among two treatment groups of normalized data in a parallel two-sample group design. In the early beginnings of the application of microarrays the k-fold rule has been used to decide whether the activity of genes is dependent between the two groups. That is, if the ratio of treatment effects exceeds a certain threshold, then the difference in activity is significant. For example DeRisi *et al.* (1997) searched for a 2-fold change of gene expression compared to a control and Iyer *et al.* (1999) sought genes whose expression changed by a factor of 2.2 or more in at least two of the experiments. If the k-fold rule is the only criteria to declare genes a differentially expressed, then '*conclusions made on the basis of such fragile foundations are likely to prove misleading and premature*' as Miller *et al.* (2001) write in their article.
In current literature statistical tests are used to find differentially expressed genes. Usually the common t-test is used. This test evaluates for each gene the null hypothesis of no difference between the activity among the two treatments against the alternative hypothesis, that the activity is different from 0. Nevertheless authors may not only be interested in a general difference. Rather than testing this point-zero hypothesis, sometimes the aim is to find genes with a ratio larger than a specific k-fold. Examples are Halitschke *et al.* (2003), who declared genes as differentially expressed, if the p-value of the t-test is less than 0.05 and the ratio of treatment means exceeds a relevance threshold of 0.75 or 1.25.

A further approach to include a k-fold on interest is an analysis software for Microsoft Excel developed by the Stanford University and described in the article of Tusher *et al.* (2001) [1]. This program tests the individual genes based on the point-zero hypothesis. However an additional relevance threshold in terms of the fold change can be specified. The procedure declares genes as differentially expressed, if both the activities among the two groups are significantly different from 0 and the mean expression ratio exceeds the relevance threshold. While this algorithm correctly declares genes as significantly expressed from 0, it is not a valid proof for a k-fold change in expression. A claim of a significant fold change can only be justified by use of a statistical procedure, which directly tests the relevance-shifted hypothesis. Such a procedure for microarray data is proposed by Li and Wong (2001). They use a parametric confidence interval of the fold change based on the originally scaled data. However the application of parametric procedures on the non-transformed expression data is not common (Abruzzo *et al.* (2005)). The reason for a transformation of microarray data will be discussed in the further introduction.
In this work procedures are discussed, which directly test the relevance-shifted hypothesis and concern the special data conditions of microarray data. In particular methods are given, in which for each gene in the multivariate data set a relevance-shifted test in terms of either a relevant difference or a relevant ratio is embedded.

The challenge of the development of such a procedure is the data condition of a microarray experiment. First of all for each individual gene a relevance-shifted test among the two treatment groups has to be applied. With the use of a statistical test different types of errors can occur. According to Neyman-Pearson test theory two errors are of interest: the Type I error (α or false positive) occurs, if the null hypothesis is falsely rejected. The second error (β or false negative) happens, if a null hypothesis is falsely accepted. Commonly the Type I error is set to 0.05 when testing a single hypothesis i.e. for each gene independently. However this is not appropriate for the high-dimensional microarray data,

[1] The software is available at www-stat.stanford.edu/tibs/SAM.

as this results in testing several thousands of hypotheses. If one were to use for each of the thousands of hypotheses the same error rate of 0.05, the overall error of an experiment would increase dramatically. Several solutions for this problem exist.

The classical control for this multiple testing (or multiplicity) problem is the use of a procedure which controls the familywise error rate (FWER). This error rate is defined as the probability, that in the entire set of concerned hypotheses at least one falsely rejected hypothesis occurs. The FWER is the conventional error definition to protect against an inflating number of falsely rejected null hypotheses. Procedures which control the FWER are proposed by Hochberg and Tamhane (1987). The control of the FWER is of interest in this thesis. However it has to be noted that other error definitions exist, and two of them shall be briefly introduced here.

According to Benjamini and Hochberg (1995) it is not necessary to control the FWER particularly for high dimensional testing problems. They give the example of a treatment and a control group, which are compared to each other in terms of several endpoints. The overall result, that the treatment is superior to the control may not be erroneous, even if some false positives occur. Here the FWER can be omitted in favor of the so-called false discovery rate (FDR), which is defined as the expected proportion of errors among the rejected null hypotheses. The FDR is less stringent than the FWER, that is, by use of the FDR the number of - truly and falsely - rejected hypotheses increases. For microarray data the FDR is used, because it is more acceptable to falsely declare some genes as significant than to miss possible discriminatory endpoints. Procedures which control the FDR and are proposed for the analysis of microarrays are given by for example Reiner *et al.* (2003). Well-known is the significance analysis of microarrays (SAM), which is based on the FDR and introduced by Tusher *et al.* (2001).

Another less stringent error rate is the k-FWER proposed by Victor (1982), Hommel and Hoffman (1988) and recently discussed by Lehmann and Romano (2005). As well as for the FDR the motivation for this error rate is the higher power (the probability that a test rejects a false null hypotheses and declares a gene as differentially expressed) due to the acceptance of more falsely rejected hypotheses. The k-FWER is defined as

the probability of rejecting at least k true null hypotheses. In the special case of $k = 1$ this error rate reduces to the FWER. A procedure controlling the k-FWER is used for comparison of the new methods proposed in this thesis.

However as the FWER is the classical error rate for multiplicity, the procedures discussed here are based on this rate. Hence the condition of the new procedures is the control of the FWER and in addition they have to show a superior power behavior compared to other FWE-controlling methods.

For the analysis of microarray data several procedures controlling the FWER are discussed by for example Dudoit *et al.* (2002), Dudoit *et al.* (2003) and Ge *et al.* (2003). Examples, which are used in this work for comparison, are the easily implemented α-adjustment according to Bonferroni (for application on microarray data see Shaffer (1995)) and the powerful (double-)permutation algorithm for step-down minP adjusted p-values proposed by Westfall and Young (1993). Especially the permutation algorithm proved to be useful for the analysis of microarrays, as it takes the correlation structure among the test statistics into account and it is hence more powerful. On the other hand this approach lacks power if the sample sizes are small, which is common for microarray data (see below). As will be shown in this work, in this case the permutation algorithm tends to become discrete. That is, the p-values cannot achieve a value from a continuous null distribution; but rather the possible outcomes are of a limited number. This can lead to an extreme loss of power.

In experiments with a small sample size the procedures with a data-driven order of hypotheses proved to be superior even to the permutation algorithm. This class of tests is introduced by Kropf (2000) and in more detail by Kropf *et al.* (2002). It is the basis of the procedures proposed in this work and hence the motivation and basic algorithm shall be briefly introduced here. The classical procedures which control the FWER and are used for microarray data correct for multiplicity by reducing the local error rate for the individual hypotheses, such as the α-adjustment of Bonferroni. Or, like the permutation algorithm, they adjust the result of the test, the p-value, for multiplicity. As will be

described in the further thesis, both types of correction can be disadvantageous for high dimensional data in combination with small sample sizes. The procedures with a data-driven order of hypotheses are in so far a new class of tests, as they compute the non-multiplicity corrected p-values and in addition for each hypothesis a so-called selector statistic directly from the data. Afterwards the null hypotheses are sorted in a decreasing order of these selector statistics. Starting with the null hypothesis corresponding to the largest selector the non-multiplicity corrected p-values are compared with the unadjusted α. As long as the p-values are less than α the corresponding null hypotheses are rejected. With the first exceeding of the α by a p-value, the corresponding null hypothesis is accepted and the procedure ends. All further null hypotheses in the order are automatically accepted. This approach is a multiple testing method. However with the data-driven order of the hypotheses by use of the selector statistic this is a multivariate test as well. In fact the procedure is derived from the so-called class of stabilized tests; these are multivariate tests and are introduced in the following chapter. Because it is a multivariate approach it has the advantage, that it concerns the correlation structure among the test statistics. Hence if certain conditions hold, it is more powerful compared to standard multiple testing methods.

To understand the challenges of the statistical analysis of microarrays especially concerning multiple testing, the typical data condition of these experiments has to be taken into account. With the simultaneous analysis of thousands of genes this kind of data is extremely high dimensional. However due to the comparably high cost of these experiments, the sample sizes per group and gene are small. Examples for the sample sizes and number of genes in a microarray experiment are given in the following table:

research area	sample size per group and gene	number of genes	source
human medicine	11, 27	7,129	Golub *et al.* (1999)
	8	5,548	Callow *et al.* (2001)
	24, 20	12,582	Armstrong *et al.* (2002)
	4, 5	13,824	Polacek *et al.* (2003)
agriculture	3	11,243	Schmidt *et al.* (2005)
	3	119	Giegé *et al.* (2005)

Various problems occur with this special type of data. Firstly classical multivariate tests, such as Hotelling's T^2 cannot be applied, because these tests require a sample size larger than the number of endpoints, that is genes. Furthermore the power of a statistical test can be extremely small, as with the small number of repetitions around 3 to 5 the number of degrees of freedom is small and hence p-values increase intensely (Kooperberg *et al.* (2005)).

In combination with the small number of replications, another property of microarray data sets is challenging. The data only seldom follows a normal distribution as described by Giles and Kipling (2003). Commonly it is assumed, that the data is approximately Gaussian distributed after a logarithmic transformation (Speed (2001)). That is, a log-normal distribution of the data is supposed. However articles can be found where this assumption is questioned. For example Hunter *et al.* (2001) describe in their article, that microarray data is often noisy and not normally distributed. Ma (2004) analyzes in his work the distribution of a microarray data set and describes the distribution of the individual gene expressions as non-normal with partly extreme values of skewness and kurtosis compared to the normal distribution due to outliers. Although the t-test is robust against a certain degree of non-normality, it requires among others Gaussian distributed data and lacks power if this assumption is violated.

The following two figures depict the power of the t-test and the distribution-free rank sum test, when the normal assumption is fulfilled and also if the samples follow an exponential distribution instead. To simulate the power two samples with 10 observations each are

generated. The true means are set to 100 and 100 plus the treatment effect and a standard deviation of 10 is selected. The nominal false positive rate is set to $\alpha = 0.05$. Furthermore each result is computed with 10,000 simulation runs. In the left graphic the random numbers are generated from the standard normal distribution and in the right one the samples follow the exponential distribution. As denoted above, the common assumption on microarray data is that after a logarithmic transformation the data is normal. The additional curve depicts a case where this assumption does not hold; it is the power of the t-test on the originally exponential distributed data with a logarithmic transformation. For the rank sum test the log-transformation is not necessary, as it gives exactly the same results.

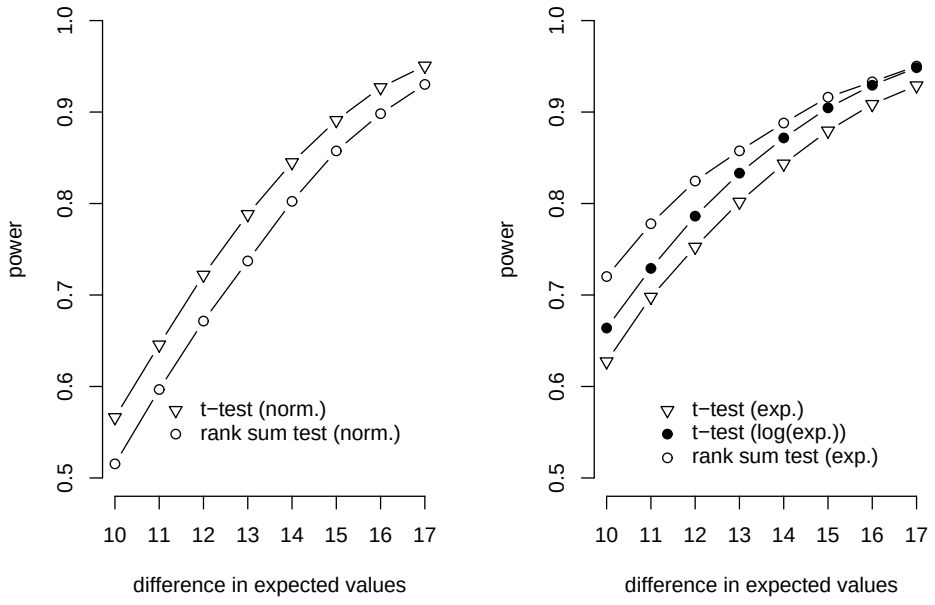

Figure 1.1: Power of t- and rank sum test for normal and exp. distributed samples

From the graphics it can be seen, that if the assumption of Gaussian distributed data is fulfilled, the t-test achieves a higher power compared to the distribution-free rank sum test. However if this requirement is not given, the rank sum test is superior; even if the

t-test is used on the log-transformed data.

For the the analysis of microarray experiments, several nonparametric procedures are proposed in the literature. For example Troyanskaya *et al.* (2002) compares the power of the permuted t-test, the rank sum test according to Wilcoxon and an ideal discrimination method based on Pearson's coefficient of correlation. However this article does neither concern a relevance shift nor the problem of small sample sizes in combination with the multiplicity correction. For example if the two-sided Wilcoxon's rank sum test is used in an experiment with a sample size of 5 for each group, then the smallest achievable two-sided p-value is $\binom{10}{5} = 0.0079$, for a sample size of 4 the smallest probability is 0.0286 and finally for a sample size of 3 it is 0.1. This is a consequence of the small sample sizes, because with a limited number of observations the null distribution of the rank sum test becomes discrete. This leads to a limited number of possible p-values. Assuming the experimental data set contains 3000 genes, the α-adjustment of Bonferroni is used and the false positive rate is set to 0.05, then the individual p-values have to be less than the comparisonwise error rate of 0.000017 to be significant, because with the Bonferroni correction the overall FWER is divided by the number of hypotheses to be tested. In this case no significant results can be found. The following graphic shall illustrate this problem. It is initially the same graphic as the right one above, but with an $\alpha = 0.001$.
With both a small error rate and sample size the rank sum test achieves a smaller power compared to the t-test, although the samples are taken from the exponential distribution. Beasley *et al.* (2004) take this problem into account. They suggest to use the maximum p-value from the t-test and a p-value computed by Chebyshev's inequality. This approach has the advantage, that even in experiments with small samples sizes significant results can be found. For example they achieved significant p-values in a simulated two-sample testing scenario with a sample size of 3 per group and $\alpha = 0.0005\%$. Another nonparametric approach concerning the small sample sizes is proposed by Neuhäuser and Senske (2004). They propose to use the Baumgartner-Weiß-Schindler test, which is less conservative than the rank sum test, because its permutation distribution is less discrete.

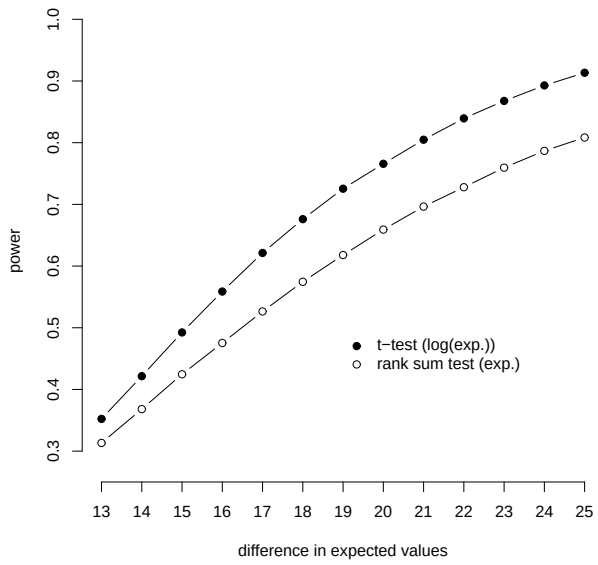

Figure 1.2: Power of t- and rank sum test for $\alpha = 0.001$

Summarizing, a procedure may be of interest, which uses a relevance-shifted test on ratio. Furthermore it has to control the FWER by use of a data-driven order of hypotheses and it shall be appropriate for skewed data. Such a procedure is the goal of this work. For its development the concept of stable tests is required, which is the basis for the procedures with a data-driven order of hypotheses. Stable tests are introduced in the second chapter. Afterwards the procedures with a data-driven order of point-zero hypotheses are discussed. First the parametric method proposed by Kropf *et al.* (2002) and afterwards the nonparametric analog given by Kropf *et al.* (2004) are introduced. Both the parametric and the nonparametric multiple testing procedures to test for a relevant difference are presented in chapter 4. Then two parametric procedures to test for a relevant ratio follow in chapter 5. The nonparametric procedure with a data-driven order of hypotheses to test for a relevant ratio is introduced in chapter 6. In the chapters 3 to 6 detailed explanations to the computation and examples are given. After the presentation of the algorithms in chapter

7 the power of the procedures compared with alternative FWER-controlling methods and some results of a k-FWER approach are shown graphically. Chapter 8 gives a summary of the results and conclusions of this thesis. Finally in appendix A detailed simulation results of the FWER are shown.

It should be noted, that although the aim of this work is the construction of procedures for the analysis of microarrays, they are general multiple testing methods for multivariate designs and can be applied for other experimental questions as well.

The entire software used in this work is implemented in R, version 2.0.1; as this statistical analysis software is widely used for the analysis of microarray data. A printout of the software is omitted in this work, but an R package with a selection of algorithms is available on request. For the software the R packages 'exactRankTests' (Hothorn and Hornik (2004)), 'multtest' (Pollard *et al.* (year not specified by authors)) and 'DAAG' (Maindonald and Braun (2004)) are required.

Chapter 2

Introduction to stabilized tests

Many experimental questions require different observations from experimental units. If more than one kind of observation is taken from a subject, the statistical term for such a sample is multivariate data and the individual observations are denoted as endpoints. A multivariate data set is characterized by a set of m endpoints taken from n experimental units. In the one-sample case all endpoints of an object k ($k = 1, \ldots, n$) can be summarized in the column vector $\mathbf{x}_k$. The entire data set can be written as all n column vectors combined in a matrix $\mathbf{X}$ with the dimensions $m \times n$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}. \tag{2.1}$$

Column k denotes the m endpoints taken of the experimental unit k and row j ($j = 1, \ldots, m$) represents the n independent measurements. Hence each row denotes an endpoint and for each endpoint the columns represent the repetitions.

In the two-sample randomized parallel design the data set contains the samples 1 and 2 with the index $i = 1, 2$. And the data matrix $\mathbf{X}$ subsumes all m-dimensional vectors $\mathbf{x}_{ik}$

to a $m \times N(N = n_1 + n_2)$ matrix. It is denoted by

$$\mathbf{X} = \begin{pmatrix} x_{111} & \cdots & x_{11n_1} & x_{211} & \cdots & x_{21n_2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{1m1} & \cdots & x_{1mn_1} & x_{2m1} & \cdots & x_{2mn_2} \end{pmatrix}. \tag{2.2}$$

In this work statistical tests for the difference or ratio are used on the multivariate data to find differences in the locations between the groups for each endpoint. Except as otherwise stated, differences or ratios in arithmetic means are of interest. For all further parametric tests it is assumed, that the m-dimensional observation vectors are multivariate normal distributed with expected values dependent on the treatment group and unknown, but equal covariance matrices:

$$\mathbf{x}_{ik} \sim N_m(\mu_i, \Sigma = (\sigma_{jj'}) = (\rho_{jj'}\sigma_j\sigma_{j'})),\ k = 1, \ldots, n_i, \tag{2.3}$$

where $\rho_{jj'}$ denotes the correlation among the endpoints for group i. In the univariate case the estimator of the expected value μ_i for the treatment group i is computed by $\bar{x}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{ik}$. And the estimator of the variance σ_i^2 is calculated by $s_i^2 = \frac{1}{n_i-1}\sum_{k=1}^{n_i}(x_{ik}-\bar{x}_i)^2$. For the multivariate data the mean vector $\mu_\mathbf{i}$ including all m means for group i is estimated with

$$\bar{\mathbf{x}}_i = \begin{pmatrix} \bar{x}_{i1} \\ \bar{x}_{i2} \\ \vdots \\ \bar{x}_{im} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_i}\sum_{k=1}^{n_i} x_{i1k} \\ \frac{1}{n_i}\sum_{k=1}^{n_i} x_{i2k} \\ \vdots \\ \frac{1}{n_i}\sum_{k=1}^{n_i} x_{imk} \end{pmatrix} \tag{2.4}$$

and the covariance matrix $\mathbf{\Sigma}_i$ is estimated with $\mathbf{S}_i$ by

$$\mathbf{S}_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i} \begin{pmatrix} (x_{i1k} - \bar{x}_{i1})^2 & (x_{i1k} - \bar{x}_{i1})(x_{i2k} - \bar{x}_{i2}) & \ldots & (x_{i1k} - \bar{x}_{i1})(x_{imk} - \bar{x}_{im}) \\ (x_{i2k} - \bar{x}_{i2})(x_{i1k} - \bar{x}_{i1}) & (x_{i2k} - \bar{x}_{i2})^2 & \ldots & (x_{i2k} - \bar{x}_{i2})(x_{imk} - \bar{x}_{im}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{imk} - \bar{x}_{im})(x_{i1k} - \bar{x}_{i1}) & (x_{imk} - \bar{x}_{im})(x_{i2k} - \bar{x}_{i2}) & \ldots & (x_{imk} - \bar{x}_{im})^2 \end{pmatrix}. \tag{2.5}$$

Here it is assumed, that $\Sigma_i = \Sigma_{i'} = \Sigma$ for $i \neq i'$.

In the statistical analysis of multivariate data at least two problems occur. First, a restriction to the classical multivariate tests is that the number of repetitions has to be larger

than the number of endpoints ($N > m$). The second problem is that with an increasing correlation among the endpoints the power can decrease. Both problems will be discussed by the use of Hotelling's T^2 test.

2.1 Simulated power of Hotelling's T^2 test

The classical test to analyze data sets consisting of two samples and multiple endpoints is Hotelling's T^2 test. In the two sample case this method tests whether there is at least one endpoint different between two independent groups. More formally, the test compares the m endpoints within the two treatment groups concerning the mean vectors $\mu_\mathbf{1}$ and $\mu_\mathbf{2}$. The ith mean vector $\mu_\mathbf{i} = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im})'$ includes the individual expected values of the different endpoints. The null hypothesis (H_0) states the equality of the two mean vectors: $H_0 : \mu_\mathbf{1} = \mu_\mathbf{2}$, and the alternative hypothesis (H_1) represents the experimental interest, which is to show at least one difference in the location vectors of the two populations : $H_1 : \mu_\mathbf{1} \neq \mu_\mathbf{2}$. For equal covariance matrices Hotelling's T^2 test statistic is given by:

$$T^2 = \frac{n_1 \cdot n_2}{N}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)' \cdot \mathbf{S}^{-1} \cdot (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1), \tag{2.6}$$

where the covariance matrix $\mathbf{S}$ is computed with $\mathbf{S} = \frac{1}{N-2}\left((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2\right)$ (Hartung & Elpelt) (1999). The test rejects the null hypothesis at significance level α if $T^2 > \frac{m(N-2)}{N-m-1} \cdot F_{df_1=m,df_2=N-m-1,1-\alpha}$, where $F_{df_1=m,df_2=N-m-1,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the F-distribution with m and $N - m - 1$ degrees of freedom.

As described above, a number of endpoints less than the sample size and redundancies in the endpoints are problematic in the analysis of multivariate data. The following simulated example shows this problem. It is motivated by Kropf (2000), where the author uses instead of a two-sample Hotelling's T^2 the analog for the one-sample case. For many tests closed expressions exist to compute the power. Alternatively or if no equations exist, these probabilities can be estimated by computer simulations. To clarify the problems occurring with multivariate testing graphically, two samples with five repetitions each ($n_1 = n_2 = 5$) and a varying number of endpoints shall be tested with Hotelling's T^2. For the first

treatment group all expected values are set to $\mu_{\mathbf{1}} = (0, \ldots, 0)'$ and the mean vector for group two is $\mu_{\mathbf{2}} = (1.5, \ldots, 1.5)'$. The theoretical standard deviation for each endpoint of the ith treatment group is $\sigma_{ij} = 1$ and the correlation $\rho_{ijj'}$ between the endpoints j and j' $(j \neq j')$ of group i is variable. The following graphic presents the simulated power of Hotelling's T^2 dependent on the number of endpoints. For four different correlations the power curves are shown. Each power result is obtained with 10,000 simulation runs and a fixed seed is used.

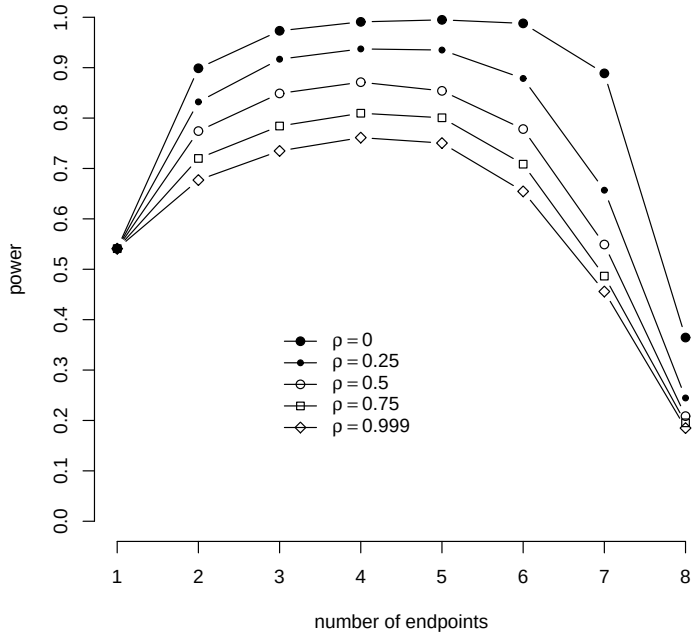

Figure 2.1: Hotellings T^2 test: comparison of power for varying number of endpoints and correlations

For only one endpoint - and therefore no differences between the curves for different $\rho_{ijj'}$ - all power lines are the same. In this case Hotelling's T^2- test reduces to the two sided, two-sample t-test, which is given by

$$t = \frac{\mid \bar{x}_2 - \bar{x}_1 \mid}{s_{pool} \cdot \sqrt{\frac{1}{n_2} + \frac{1}{n_1}}} \sim t_{df,1-\alpha/2} \tag{2.7}$$

with $s_{pool} = \sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(N - 2)}$ denoting the pooled standard deviation ("STUDENT" (1908)). The t-test is t-distributed with $df = N - 2$ degrees of freedom and $\alpha -$ quantiles are taken from the upper tail.
For all curves, and therefore independent of the correlation, the power increases with the inclusion of up to four endpoints. If more endpoints are included in the analysis, then the power decreases. With eight endpoints the power lines in the graphic end, because a number of endpoints equal to or larger than the total number of experimental units results in a singular and thus noninvertible covariance matrix. If the endpoints are correlated to a small degree then with each included endpoint valuable information is gained. Hence the power increases until the procedure lacks power due to the increased number of endpoints. However with an increasing correlation, the gain in additional information decreases. Although the power increases, it does not reach the same level as the power of the less correlated endpoints.

To overcome the classical problems of multivariate procedures LÄUTER *et. al.* (1996) introduced a class of so-called stabilized tests for the analysis of dimensional data. These tests exactly control the α-level, have a higher power compared to other procedures under special restrictions and can be used for high-dimensional data sets with small sample sizes. First the general idea of the theory is shortly presented. For this purpose the so-called standardized sum test is introduced. Detailed explanations, other special test versions, more general theorems and their proofs can be found in LÄUTER *et. al.* (1996) and LÄUTER (1996).

2.2 Stabilized parametric two-sample tests

Basically a stabilized test consists of two steps. In the first step the high-dimensional data vectors are transformed into univariate or low-dimensional scores by use of a linear data transformation with a data-dependent weight vector. Afterwards these scores are tested with a classical univariate or multivariate test. The following example is restricted to

univariate scores, which are compared with the common two-sample t-test.

In contrast to the original data vectors the scores are neither Gaussian distributed nor independent. However if these scores are derived by following special rules, then they can be compared with the t-test, which controls exactly the Type I error rate. The score z_{ik} consists of the data vector $\mathbf{x}_{ik}$ and a m-dimensional weight vector $\mathbf{d}$ and it is computed as

$$z_{ik} = \mathbf{d}'\mathbf{x}_{ik}. \tag{2.8}$$

The weight vector $\mathbf{d}$ may be a function of the sums of squares and cross products matrix

$$\mathbf{W} = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})' \tag{2.9}$$

where $\bar{\mathbf{X}}$ is the $2 \times N$ matrix of total means. For it's computation an N-dimensional vector of ones $\mathbf{1}_N$ is required:

$$\bar{\mathbf{X}} = \frac{1}{N}\mathbf{X}\mathbf{1}_N\mathbf{1}'_N. \tag{2.10}$$

Furthermore it has to be ensured, that $\mathbf{d} \neq \mathbf{0}$ with probability 1.

With the scores z_{ik} the t-test can be computed in the usual way:

$$t = \frac{\bar{z}_2 - \bar{z}_1}{s_z\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.11}$$

where $\bar{z}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} z_{ik}$ and $s_z^2 = \frac{1}{N-2}\sum_{i=1}^{2}\sum_{k=1}^{n_i}(z_{ik} - \bar{z}_i)$ and the test statistic follows a t-distribution with $N - 2$ degrees of freedom.

The matrix $\mathbf{W}$ itself is the sum of the matrices $\mathbf{H}$ and $\mathbf{G}$. The first one represents the deviations from the null hypothesis:

$$\mathbf{H} = \frac{n_1 n_2}{N}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)'. \tag{2.12}$$

And matrix $\mathbf{G}$ includes the residual errors:

$$\mathbf{G} = \sum_{i=1}^{2}\sum_{k=1}^{n_i}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)'. \tag{2.13}$$

Hence it is called the total sums of squares and cross products matrix. Both $\mathbf{H}$ and $\mathbf{G}$ follow a central Wishart distribution with the same matrix parameter $\mathbf{\Sigma}$ and degrees of

freedom 1 and $N-2$ under H_0.

For the standardized sum test (SS-test) the weight vector is computed by

$$\mathbf{d} = (\text{Diag}(\mathbf{W}))^{-1/2}\mathbf{1}_N. \tag{2.14}$$

Alternatively the scores are calculated with

$$z_{ik} = \sum_{j=1}^{m} \frac{x_{ijk}}{\sqrt{\sum_{i=1}^{2}\sum_{g=1}^{n_i}(x_{ijg} - \bar{x}_j)^2}}. \tag{2.15}$$

The following graphic shows the power of the stabilized SS-test in comparison to the classical Hotelling's T^2. It is generated under the same conditions as figure 2.1.

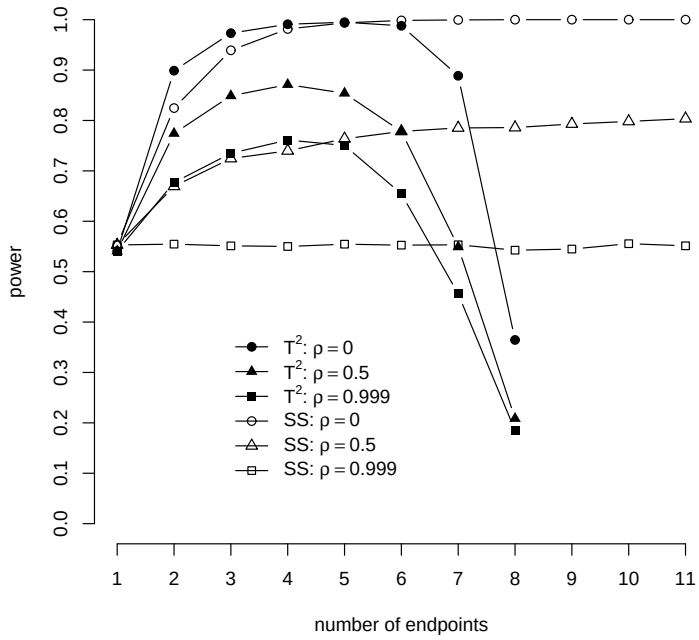

Figure 2.2: Hotelling's T^2 and standardized sum test: comparison of power

It can be seen, that the SS-test is not as powerful as Hotelling's T^2 if the number of endpoints is less than 7. However, if more endpoints are observed, then Hotelling's T^2 looses power and is finally not computable anymore. But even with 11 endpoints, the

power of the SS-test stays constant on its highest level. It has to be noted that the application of the SS-test is even possible with a further inclusion of endpoints.
An additional advantage of the SS-test not shown here is the option for one-sided testing. This is not available for Hotelling's T^2.
The further procedures discussed in this work are used for the analysis of multivariate data. However instead of a global statement as possible with Hotelling's T^2 or the SS-test these procedures give local decisions. Hence differences among the groups for each individual endpoint are of interest. As Kropf (2000) showed in his work the theory of stabilized tests can be used for the construction of a multiple testing procedure as well. In this thesis procedures are presented, which are based on the idea of Kropf (2000).

Chapter 3

Testing procedures for point zero-hypotheses

In the former chapter the idea of stabilized tests was introduced. The stabilized tests can be used for the construction of multiple testing procedures as well. For the first time Kropf (2000) presents such tests and among others further works extends this method to parametric (Kropf and Läuter (2002)) and nonparametric procedures (Kropf *et al.* (2004)). The new procedures discussed here are derived from such multiple testing procedures. Hence before the new methods are shown, a parametric and a nonparametric multiple testing method based on the theory of stabilized tests are presented.
However before the algorithms and their explanation are given, the concept of the familywise error rate and a special type of power are required. After the algorithms a small example data set and two microarray experiments are analyzed with these methods and finally graphical simulation results of the power compared to standard multiple testing methods are given.

The familywise error rate and the power: In the introduction the Type I error rate and the power were briefly introduced. These definitions are valid for testing a single null hypothesis. For multiple testing the definitions have to be extended. If more than one null hypothesis is tested, the familywise error rate (FWER) is used. It is the probability of

rejecting at least one null hypothesis when all null hypotheses are considered. Out of m null hypotheses let m' hypotheses be true and the other $m - m'$ false. Then the FWER is defined as

$$\text{FWER} = P(\text{reject at least one of } H_{0,1}, H_{0,2}, \ldots, H_{0,m'} \mid H_{0,1}, H_{0,2}, \ldots, H_{0,m'} \text{ are true}). \tag{3.1}$$

The FWER can be controlled in the weak and in the strong sense. A multiple testing procedure controls the FWER in the weak sense, if the FWER is less than or equal to the selected α in the special case that all null hypotheses are true. Further, the FWER is protected in the strong sense, if the probability to reject any true null hypotheses is less than or equal to α even in the case where some null hypotheses are false and some are true.

As denoted in the introduction the definition for the power is the probability of rejecting a false null hypothesis. This holds for the case of testing a single hypothesis. Concerning multiple hypotheses the definition is more complex. Westfall *et al.* (1999) describes four types:

- complete power = P(reject all null hypotheses that are false)
- minimal power = P(reject at least one null hypothesis that is false)
- individual power = P(reject a particular null hypothesis that is false)
- proportional power = average proportion of false null hypotheses that are rejected.

The proportional power represents the goal of the experiments discussed here, as reflects the goal of such studies: to find as many false null hypotheses as possible (Westfall & Krishen (2001)). For microarray data this definition of power is widely accepted, see for example Kropf & Läuter (2002) and Dudoit *et al.* (2003). As it is the aim to control the FWER here, the minimal power could have been used as well. If all null hypotheses observed in an experiment are true, then the minimal power converges to the FWER. However, as it does not distinguish between rejecting one null hypothesis or more

than one, it cannot reflect the goal of the experiment. Further the complete power is not appropriate, as it demands that the multiple testing procedure is able to reject all false null hypotheses. The individual power considers one specific null hypothesis. Hence unless the analyst has for example a priori knowledge of one specific endpoint and the interest lies in this hypothesis only, the individual power is not useful for multivariate data.

3.1 Procedures

All multiple testing procedures based on the stabilized tests which are discussed in this work are split in two parts. In the first part for each endpoint the p-value of a two-sample test and a data-dependent selector statistic are calculated. In the second step the p-values are sorted in descending order of their corresponding selector statistics and sequentially the p-values are compared with the unadjusted α. The null hypothesis of an endpoint is rejected, if the corresponding p-value is equal or less than α and all former null hypotheses in the ordering have been rejected as well. Thus the procedure stops at the first non-significance and all further null hypotheses are accepted automatically.
Therefore both the parametric and the nonparametric procedure shown here are similar except that one uses a parametric test and a selector based on the deviations of the individual observations to the arithmetic mean and the other one consists of a nonparametric test and the selector is a robust estimator of dispersion.

3.1.1 Parametric procedure

With the following parametric procedure proposed by Kropf and Läuter (2002) it will be tested for each endpoint whether the two samples have a significant difference. In terms of hypotheses it will be tested, whether the null hypothesis $H_{0,j} : \mu_{2j} - \mu_{1j} = 0$ can be rejected in favor of $H_{1,j} : \mu_{2j} - \mu_{1j} \neq 0$. The algorithm of the two-sided two-sample parametric multiple testing procedure with a data-driven order of hypotheses for independent samples is:

1. Compute independently for each endpoint the two-sided p-value p_j by use of the two-sided two-sample t-test for independent samples:

$$t_j = \frac{\mid \bar{x}_{2j} - \bar{x}_{1j} \mid}{s_{pool,j} \cdot \sqrt{\frac{1}{n_2} + \frac{1}{n_1}}} \tag{3.2}$$

with the two sample means of the jth endpoint $\bar{x}_{ij} = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ijk}$ and the standard deviation of the pooled samples $s_{pool,j} = \sqrt{\frac{\sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_{ij})^2}{N-2}}$.
Furthermore calculate the selector statistic

$$w_j = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_j)^2, \tag{3.3}$$

with the total mean of both samples per endpoint $\bar{x}_j = \frac{1}{N} \sum_{i=1}^{2} \sum_{k=1}^{n_i} x_{ijk}$.

2. Sort the variables according to their selector statistic in decreasing order. To reject $H_{0,j} : \mu_{2j} - \mu_{1j} = 0$ p_j has to be less than the unadjusted α. At the first non-significance the procedure stops and all further null hypotheses are accepted automatically.

This procedure can be applied with one-sided tests as well. In this case the one-sided p-values are used, but the selector statistic remains the same. However it has to be noted, that by the application of one-sided tests the procedure can lack power. With the same selector statistics the order of hypotheses is exactly the same as for two-sided testing. If an endpoint has a significant result in the opposite direction, then although it's unadjusted p-value is large, the selector is large as well. Hence the procedure can abort prematurely.

The connection to the stabilized tests is the derivation of the selector statistic. The jth selector statistic $w_j = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_j)^2$ corresponds to the jth diagonal element of the matrix of the total sums of squares and cross products $\mathbf{W}$:

$$\mathbf{W} = \mathbf{H} + \mathbf{G} = \frac{n_1 n_2}{N} (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)' + \sum_{i=1}^{2} \sum_{k=1}^{n_i} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)'. \tag{3.4}$$

The weight vector $\mathbf{d}$ is a vector of zero elements and a one element at the position of the maximum diagonal elements of $\mathbf{W}$. Hence the first endpoint corresponding to the maximum

selector statistic is tested with a two-sample test. If the null hypothesis of this endpoint is correctly rejected, then the procedure continues with the second largest selector. But if the null hypothesis is incorrectly rejected, then it has to be ensured that this occurs with a probability α. Otherwise if the null hypothesis is accepted, the procedure stops and no Type I error can occur.

In this procedure the selector statistic reflects the variance among both samples per endpoint and it is therefore an estimate of the scale. If the difference in means is close to 0 the selector achieves a comparatively small value, because the squared deviations to the total mean of both samples is small. And with an increasing difference between the two samples the selector increases. Two small examples shall illustrate the effect. Note that the index j is omitted. The two samples $x_{1k} = 3, 3.5, 4$ and $x_{2k} = 3.1, 3.6, 3.9$ are given. Clearly both have similar observations. The difference in means is $\bar{x}_2 - \bar{x}_1 = 0.033$ and the selector statistic $w_j = 0.828$ indicates the small dispersion. In the second example the former samples are used as well, but with $x'_{2k} = x_{2k} + 1 = 4.1, 4.6, 4.9$. In this case the second sample has clearly larger values compared to the first one. The difference in means is 1.033 and the selector statistic increases to $w_j = 2.428$. Hence if a difference in location among the groups exists, the endpoint achieves a large selector. Otherwise the selector does reflect the variances of the individual groups only.

This procedure is powerful, if the variances among the endpoints are homogeneous. If this is not given, the FWER is still controlled, but the procedure loses power. For example if the single values of the two samples x_{1k} and x_{2k} are multiplied by 2, then the difference in means increases to 0.067, the selector however achieves a value of 3.313. Although the difference in means is less than in the second example, the selector is larger. Hence with heterogeneous variances among the endpoints a lack of power may arise, because endpoints under H_0 with a p-value greater than α and a high variance leading to a large selector can stop the procedure prematurely, which may result in false acceptances of other null hypotheses. Summarizing the procedure is particularly appropriate for endpoints with equally scaled observations.

In all further procedures with a data-driven order of hypotheses the selector is a function

of both the treatment effect and the variance of the the samples.

3.1.2 Nonparametric procedure

In data sets with several endpoints it is likely that the assumption of Gaussian distributed data can not be fulfilled. As discussed in the introduction this is especially true for high-dimensional data such as microarrays, where usually thousands of endpoints are tested. In this case the use of a nonparametric procedure is reasonable. This section presents the nonparametric multiple testing procedure with a data-driven order of hypotheses proposed by Kropf *et al.* (2004).

For the application of the procedure, it is assumed that the independent m-dimensional sample vectors $\mathbf{x}_{1k} > 0$ and $\mathbf{x}_{2k} > 0$ follow the continuous distribution functions $F_m(\mathbf{x})$ and $G_m(\mathbf{x})$. These functions are considered to be equal except for a shift in location: $G_m(\mathbf{x}) = F_m(\mathbf{x} - \boldsymbol{\Delta})$ with the vector of the treatment effects $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_m)'$.

The procedure tests whether the treatment effect of endpoint j is unequal to 0. In terms of hypotheses, it is tested whether the null hypothesis $H_{0,j} : \Delta_j = 0$ can be rejected in favor of $H_{1,j} : \Delta_j \neq 0$.

The procedure uses the rank sum test according to Wilcoxon (1945) instead of the t-test as the two-sample test. As this test is more complex than the t-test, it will be started with the algorithm of this test. For sake of simplicity it is presented by use of two samples without the index for the jth endpoint.

Exact rank sum test: To compute the exact rank sum test sort the combined samples x_{1k} and x_{2k} in an increasing order and rank them. Denote the ranks of the second treatment group by r_{2k}. Calculate the sum of the r_{2k}:

$$W = \sum_{k=1}^{n_2} r_{2k}. \tag{3.5}$$

The two-sided null hypothesis $H_0 : \Delta = 0$ is rejected, if either $W \geq w_{\alpha/2}$ or $W \leq n_2(n_1 + n_2 + 1) - w_{\alpha/2}$, where $w_{\alpha/2}$ denotes the upper tail probabilities from the null distribution of the Wilcoxon rank sum test and values for the quantile are given in for

example HOLLANDER AND WOLFE (1999) with $n_2 \leq n_1$. If $n_1 < n_2$ the rank sum is taken from the ranks of the first sample and the critical value of $n_2(n_1+n_2+1)-w_{\alpha/2}$ changes to $n_1(n_1+n_2+1)-w_{\alpha/2}$. The assumption of $n_2 \leq n_1$ is done throughout this work. In the one-sided case to test on increase this test reduces to $H_0 : \Delta \leq 0$ and it is rejected with $W \geq w_\alpha$. For testing against a decrease $H_0 : \Delta \geq 0$ is rejected if $W \leq n_2(n_1+n_2+1)-w_\alpha$. If samples are tested which include equal observations (ties), then the conditional null-distribution can be computed as proposed by HOLLANDER AND WOLFE (1999).

Asymptotic rank sum test: In addition to the exact version the asymptotic one is presented as well for reason of completeness. However as the normal approximation requires a sample size larger than 25 in one group (BÜNING AND TRENKLER (1994)), it may not be appropriate for the analysis of microarray data using small sample sizes.
To compute the asymptotic rank sum test the expectation and the variance of the statistic W are needed. The expectation of W is computed with

$$E(W) = \frac{n_2}{N} \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik} \tag{3.6}$$

and the variance of W is given by

$$Var(W) = \frac{n_1 n_2}{N^2(N-1)} \left\{ N \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik}^2 - \left(\sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik} \right)^2 \right\}. \tag{3.7}$$

Then the large-sample approximation of the Wilcoxon test is

$$W^{approx} = \frac{W - E(W)}{\{Var(W)\}^{1/2}}. \tag{3.8}$$

The expected value and the variance are taken from HOTHORN AND HORNIK (2002). As the authors stated, the equation is valid for tied samples as well. To reject the two-sided null hypothesis $|W^{approx}|$ has to be greater or equal to $z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ denotes the $\alpha/2$-quantile from the right tail of the standard normal distribution. In the one-sided case H_0 is rejected if $W^{approx} \geq z_{1-\alpha}$ (increase) or $W^{approx} \leq z_\alpha$ (decrease), where z_α denotes the α-quantile from the left tail of the standard normal distribution.

Nonparametric procedure with data-driven order of hypotheses: As the two-sample test the rank sum test is used. However, a selector statistic which is robust against deviations from the normal distribution is required as well. A nonparametric statistic to measure the 'total variance' is the interquartile range (IQR), which is here the difference of the 75% to the 25% quantile among the two samples per endpoint. To compute the quantiles, the pooled samples are ordered, such that $x_{ijk}^{(1)} \leq \ldots \leq x_{ijk}^{(N)}$. Following the default definition of R (see help file for 'quantile'), the quantile is given by $q_{(h-1)/(N-1)} = x_{ijk}^{(h)}$, with $h = 1, \ldots, N$ (Parzen (2004)). By use of the rank sum test and the IQR the two-sided testing procedure is computed as:

1. Compute independently for each endpoint the p-value of the two-sided rank sum test according to Wilcoxon (1945) either exact or asymptotic.
   Furthermore calculate the interquartile range as selector statistic
   $$IQR = q_{75} - q_{25}. \tag{3.9}$$
2. Sort the variables according to their IQR in decreasing order.
3. To reject $H_{0,j} : \Delta_j = 0$ the jth p-value has to be less than the unadjusted α. At the first non-significance the procedure stops and all further null hypotheses are accepted automatically.

Both the parametric and the nonparametric procedures control exactly the FWER. In the following chapter similar procedures including relevance-shifted tests are presented. As no proof of them is provided, the empirical control of the FWER is given only. However the procedures reduce to the approaches introduced in this chapter, when the relevance thresholds are set to 0 (test on relevant differences) or 1 (test on relevant ratios) and if in one case the data is logarithmized (see chapter 6). In the appendix simulation results of the FWER are shown. As the reduction to the above discussed procedures is tested as well, the two procedures above can be used as a benchmark for the simulated FWER of the new procedures. The simulations of the FWER are listed in the appendix starting on page 155.

Adjusted *p*-values: To present results of the procedures with a data-driven order of hypotheses, a table listing the p-values and the selector statistics is necessary. However, by reporting a p-value, which is adjusted for multiplicity, the selector statistics can be omitted. Here the adjusted p-value can be directly compared with the chosen FWER without regarding the ordering of the hypotheses.

Irrespective of the multiplicity correction, an adjusted p-value is defined as the smallest overall significance level at which the corresponding hypothesis can be rejected by use of a multiple testing method (Wright (1992)). For the procedures with a data-driven order of hypotheses the (unadjusted) p-values are sorted according to their selector statistic in decreasing order and denoted as $p_{(1)}, \ldots, p_{(m)}$. Then the adjustment is as follows: $p^{adj}_{(1)} \leftarrow p_{(1)}$ and $p^{adj}_{(j)} \leftarrow \max(p^{adj}_{(j-1)}, p_{(j)})$.

3.2 Examples

In this section three example data sets are analyzed by use of the parametric and the nonparametric procedures with a data-driven order of hypotheses. The first data set is a small one including nine endpoints. This small data set is chosen as an example to show all results of the procedures with a data-driven order of hypotheses. The other two experiments are microarray data sets. Because of their size, only a part of the results can be shown.

For comparison the results of the corresponding two-sample tests with another multiplicity correction is given. For this purpose the α-adjustment according to Bonferroni is used. This method is probably the best known and easiest correction. Following Sachs (1997) each hypothesis is tested against $\alpha/\#$ hypotheses, which is α/m for the two-sample multivariate data discussed here. Alternatively the Bonferroni adjusted p-values for the m hypotheses can be computed by $p^{bon}_j = \min(p_j \cdot \ \#\ \text{hypotheses}, 1)$. An advantage of this method is the control of the FWER in the weak and the strong sense for many settings and its easy application. This method can however be extremely conservative because, among other reasons, it ignores the stochastic dependencies among the endpoints. This is a considerable

disadvantage for the analysis of the multivariate data discussed here, because as it will be seen in chapter 7 the genes in a microarray data set are correlated and therefore their test statistics as well.
In all analysis in this and the following chapters it is tested two-sided by use of an error rate of $\alpha = 5\%$.

3.2.1 Possum data set

This data set is published by Lindenmayer *et al.* (1995) and it is available in the R package 'DAAG' from Maindonald and Braun (2004). In this experiment morphological measurements are taken from mountain brushtail possums (*Tichosurus caninus*) on seven locations in Australia. The individual morphological endpoints are the length of the head (hdlngth), the distance across the widest part of the head (skullw), the total body length (totlngth), the distance from the base to the tip of the tail (taill), the distance from the heel to the tip of the largest toe (footlgth), the length of the ear conch (earconch), the distance from medial canthus to lateral canthus of right eye (eye), the body girth (chest) and the belly girth (belly).
Only a part of this data set is used here. In particular only male animals are observed and the treatment groups correspond to the second and the third site. The following table lists the here used subset of the original data set:

endpoint	site 2	site 2	site 2	site 2	site 2	site 2	site 2	site 2	site 3	site 3	site 3	site 3
hdlngth	90.6	94.4	93.3	92.4	85.3	85.1	90.7	91.4	90.1	98.6	95.4	97.6
skullw	55.7	57.9	59.3	56.0	54.1	51.5	55.9	54.4	54.8	63.2	59.2	61.0
totlngth	85.5	85.0	88.0	80.5	77.0	76.0	81.0	84.0	89.0	85.0	85.0	93.5
taill	36.5	35.5	35.0	35.5	32.0	35.5	34.0	35.0	37.5	34.0	37.0	40.0
footlgth	73.1	71.2	74.3	68.4	62.7	70.3	71.5	72.8	66.0	66.9	69.0	67.9
earconch	53.1	55.5	52.0	49.5	51.2	52.6	54.0	51.2	45.5	44.9	45.0	44.3
eye	14.4	16.4	14.9	15.9	13.8	14.4	14.6	14.4	15.0	17.0	15.9	15.8
chest	26.0	28.0	25.5	27.0	25.5	23.0	27.0	24.5	25.0	28.0	29.5	28.5
belly	28.5	35.5	36.0	30.0	33.0	27.0	31.5	35.0	33.0	35.0	35.5	32.5

The first table shows the analysis of the data set by use of the parametric procedure. In the first column the name of the endpoint is listed. Afterwards the difference in means, the selector statistic and the test statistic follow. Then the unadjusted and the adjusted p-values are given and finally the Bonferroni corrected p-value are shown.

Results of the parametric procedures:

endpoint	difference in means	selector statistic	test statistic	unadjusted p-value	adjusted p-value	Bonferroni adjusted p-value
totlngth	6.000	270.563	2.345	0.041	0.041	0.369
hdlngth	5.025	194.143	2.304	0.044	0.044	0.395
earconch	-7.463	173.380	-7.726	$1.595 \cdot 10^{-5}$	0.044	$1.44 \cdot 10^{-4}$
footlgth	-3.087	123.690	-1.608	0.139	0.139	1.000
skullw	3.950	119.457	2.312	0.043	0.139	0.390
belly	1.938	96.229	1.078	0.307	0.307	1.000
taill	2.250	44.563	2.085	0.064	0.307	0.573
chest	1.938	38.729	1.867	0.091	0.307	0.823
eye	1.075	10.389	2.054	0.067	0.307	0.604

While with the α-adjustment only one significant p-value can be observed, three significant endpoints can be found with the parametric procedure with a data-driven order of hypotheses. Theoretically four significant endpoints could have been observed, however the procedure is aborted prematurely because the fourth endpoint ('footlgth') has a p-value larger than α and a comparatively high selector.

Results of the nonparametric procedures:

endpoint	difference in medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value	Bonferroni adjusted p-value
earconch	-7.35	7.350	10.0	0.004	0.004	0.036
totlngth	4.50	5.250	37.0	0.077	0.077	0.691
skullw	4.30	4.525	36.0	0.109	0.109	0.982
hdlngth	5.45	4.175	36.0	0.109	0.109	0.982
footlgth	-3.95	4.175	15.0	0.073	0.109	0.655
belly	1.75	4.000	30.5	0.489	0.489	1.000
chest	2.50	2.625	35.5	0.113	0.489	1.000
taill	2.00	1.875	35.5	0.109	0.489	0.982
eye	1.35	1.500	36.5	0.073	0.489	0.655

Only one significant endpoint can be found by both the Bonferroni adjustment and the procedure using a selector statistic, because already all other unadjusted p-values are larger than 5%. Nevertheless the procedure with a data-driven order of hypotheses gives smaller adjusted p-values compared to the conservative α adjustment. In addition, the ordering of the hypotheses is similar in comparison with the parametric analysis: three of the first four endpoints in the ordering occur in both cases. While the parametric procedure finds more significant results, the advantage of the nonparametric approach is, that it lists the endpoint 'earconch' at the top of the order. Considering the unadjusted p-values it is the most discriminating endpoint of the data set.

3.2.2 TSHR mutation data set

The second example is a microarray data set. It consists of patients with autonomously functioning thyroid nodules (AFTNs), which are benign tumors producing more thyroid hormones compared to the healthy tissue. Approximately 60% of the AFTNs are caused by a mutation of the thyrotropin receptor (TSHR). The goal of interest in this experiment is the identification of other causes leading to the pathogenesis of AFTNs. For this purpose 15 patients are observed, where 10 have a mutation of the TSHR and the remaining 5 have not. In total 12,625 genes are observed. The experiment was performed by use of Affymetrix

GeneChips and the data is normalized. Further information about the experiment and results are published by Eszlinger *et al.* (2004).

Results of the parametric procedures: The first analysis of the microarray data set is done by application of the testing procedures on the originally scaled data. As discussed in the introduction commonly a microarray data set is logarithmized prior the analysis. However for sake of comparison the non-logarithmized data shall be analyzed as well.
By use of the two-sample t-test without any multiplicity correction 1,266 significant genes can be found. With the α-adjustment of Bonferroni only endpoint # 10018 with an adjusted p-value 0.034 is significant. By use of the parametric procedure with a data-driven order of hypotheses no significantly discriminatory genes can be found. The following table presents the results of the first ten endpoints of the procedure with the data-driven order of hypotheses:

endpoint	difference in means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
1702	18999.786	6185936136	1.772	0.0998	0.0998
2071	16438.726	5739625540	1.556	0.144	0.144
12575	-8276.965	4947876639	-0.793	0.442	0.442
1539	16176.468	4751679940	1.710	0.111	0.442
12569	-5416.750	4696048346	-0.526	0.608	0.608
7940	-23469.326	4211279601	-3.170	0.007	0.608
5132	18945.422	3990932797	2.359	0.035	0.608
3304	-18418.397	3838254166	-2.330	0.037	0.608
3305	-17431.033	3784863556	-2.179	0.048	0.608
12573	-10074.488	3646527578	-1.153	0.270	0.608

Out of the first ten endpoints with the highest selector statistics four have an unadjusted p-value less than 0.05. However the procedure stops prematurely because the gene with the largest selector is not significant. It has to be noted, that the endpoint # 10018, which is significant with the Bonferroni correction, does not appear in this list.
However this result is expected, because the intensities of gene expression cover the range

of values close to 0 and up to values of around 50,000. Depending on the intensities the variance varies as well. Hence the selector statistics are biased by the variance heterogeneity among the endpoints. The following graphic depicts the dependency of the variance on mean for the patients with a mutation of the TSH receptor.

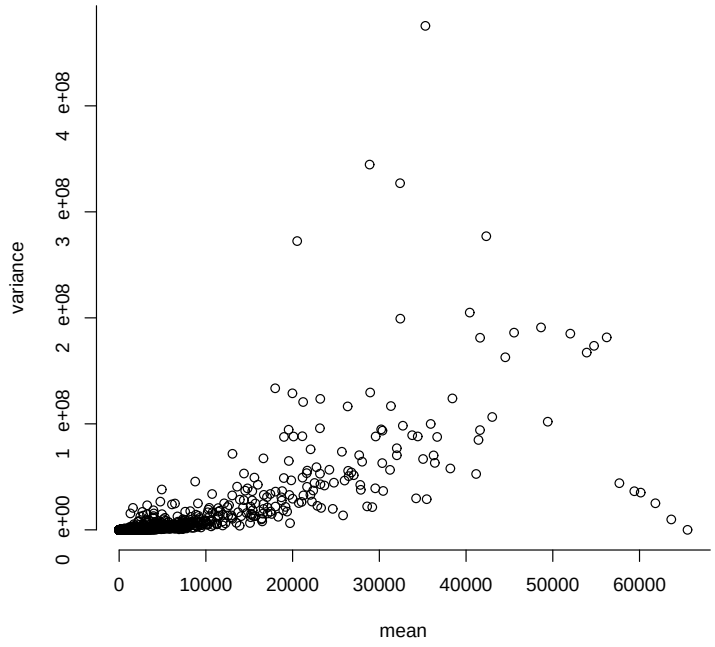

Figure 3.1: Variance vs. mean dependency for non-logarithmized data

In the literature it is commonly assumed, that by use of the logarithmic transformation the variance becomes independent from the mean intensities (Speed (2005), Geller *et al.* (2003) and Lu (2004)). As this is essential for the procedure with a data-driven order of hypotheses, a second analysis is done with the logarithmized data. Throughout this work the natural logarithm is used. It has to be noted however that the assumption of the data has changed. It is now assumed, that after the logarithmic transformation the data is Gaussian distributed with homogeneous variances. Hence prior the data transformation a log-normal distribution of the populations is presupposed.

With the local tests 1,241 significant genes can be found. That is 25 significant genes less than without the logarithmic transformation. By use of the Bonferroni correction no significant genes are achieved. And for the parametric procedure with a data-driven order of hypotheses the first ten endpoints with the highest selectors are:

endpoint	difference in means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
11321	-2.038	67.543	-1.830	0.090	0.090
8435	-2.317	45.023	-2.928	0.012	0.090
6022	0.178	42.465	0.180	0.860	0.860
4145	-2.808	41.828	-4.688	0.0004	0.860
12177	1.147	41.525	1.239	0.237	0.860
7940	-2.662	40.011	-4.327	0.0008	0.860
3839	2.502	39.971	3.768	0.002	0.860
11876	1.769	39.292	2.167	0.049	0.860
8273	-2.487	38.019	-3.925	0.002	0.860
2148	1.353	36.819	1.606	0.132	0.860

As well as for the non-logarithmized data no significant results can be found. The number of endpoints with an unadjusted p-value less than 0.05 has increased however, from four to six. In addition it has to be noted, that all ten endpoints in this list are other genes than the ones listed for the analysis with non-logarithmized data.

One reason for the lack of power of the procedure with a data-driven order of hypotheses can be an insufficient variance stabilization due to the logarithm. Although the logarithm is commonly used to stabilize the variance, there are those who find this approach questionable. Durbin *et al.* (2002) and Huber *et al.* (2002) for example discuss this problem. In both articles the authors describe the variance versus mean dependency as linear, if the mean intensities are moderately large. However this constant coefficient of variation is not given for genes with a small intensity. Here, it is assumed that the variance is constant. If the intensity of the gene is in between these extremes, the variance structure becomes more complex.

The following figure supports the assumption, that the logarithm is not a sufficient ap-

proach for the stabilization of the variance, for at least the procedures with a data-driven order of hypotheses. The figure is initially the same as the former one, but the data is logarithmized.

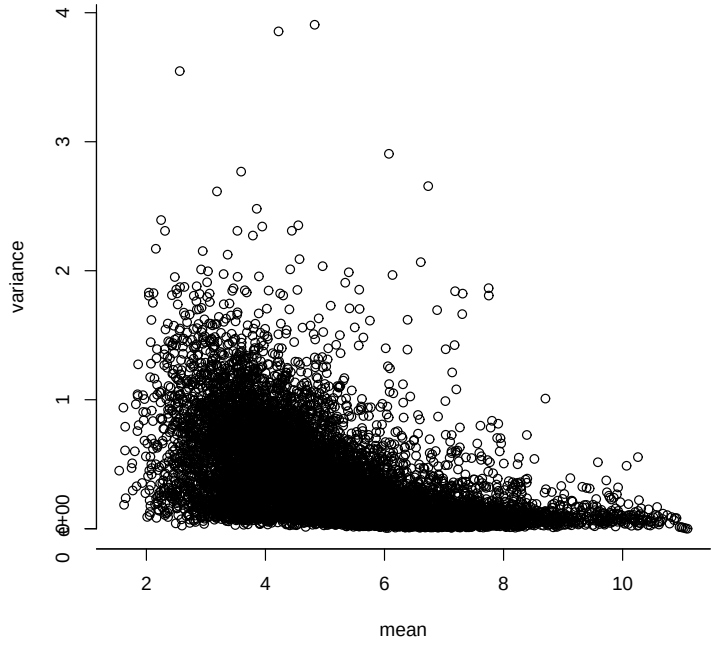

Figure 3.2: Variance vs. mean dependency for logarithmized data

Due to the logarithmic transformation the variance becomes moderately constant for large mean intensities. However for small gene expressions the variance inflates. Approaches to stabilize the variance more sufficiently are described by Durbin *et al.* (2002) and Huber *et al.* (2002). But as this is beyond the scope of this work a further investigation of such methods is omitted.

Results of the nonparametric procedures: It is conceivable, that the absence of significant findings is due to the application of parametric procedures. In this section the nonparametric procedures are applied to the data.

First the data is analyzed on the original scale. Without a multiplicity correction in total 1,023 significant genes can be found. However with both the Bonferroni correction or the procedure with a data-driven order of hypotheses no significant endpoints are achieved. With the logarithmized data the number of significant genes for the local tests and the Bonferroni correction stay the same because the result of the rank sum test is independent from the application of this transformation. The logarithm has however an impact on the computation of the selector; but even with this transformation the nonparametric procedure with a data-driven order of hypotheses does not find any differentially expressed genes. The following two tables list the ten endpoints with the highest selector statistics.

Analysis based on original data

endpoint	difference in medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value
12573	-14826.332	27625.60	71.0	0.296	0.296
1637	8103.799	25331.70	85.0	0.591	0.591
6190	16769.580	25200.54	90.5	0.212	0.591
12575	-12610.240	24908.27	73.0	0.437	0.591
1539	24165.513	23714.04	95.0	0.075	0.591
5132	15213.176	23049.45	97.0	0.040	0.591
11555	8364.084	22950.42	89.0	0.310	0.591
9686	14652.715	22878.27	89.5	0.260	0.591
2071	17429.607	22477.94	88.0	0.329	0.591
7107	3440.071	18818.10	86.0	0.513	0.591

Analysis based on log-transformed data

endpoint	difference in medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value
6022	3.671	3.001	81	0.953	0.953
7940	0.069	2.928	56	0.001	0.953
4145	0.057	2.818	58	0.005	0.953
3839	10.189	2.781	103	0.003	0.953
2808	2.568	2.744	82	0.859	0.953
12337	0.902	2.659	81	1.000	1.000
2148	3.404	2.635	93	0.129	1.000
9207	0.897	2.626	75	0.594	1.000
7892	0.201	2.575	71	0.310	1.000
5589	5.370	2.533	88	0.371	1.000

With the logarithmic transformation more unadjusted p-values can be found compared to the analysis on the original scaled data. In the analysis of the logarithmic data the endpoint with the largest selector has a large p-value. If this endpoint had not stopped the procedure, three significant genes would have been found.

3.2.3 TNFα data set

The second microarray data set is a subset of the published data from Polacek *et al.* (2003). In this subset, genes are sought which have a differential expression among tumor necrosis factor α (TNFα) stimulated human cells and non-stimulated cells. The entire data set consists of transcriptional profiles generated from amplified and unamplified mRNA. Here the amplified data is used only.

Each treatment group contains five replicates. These are repetitions pooled from several dishes. Hence this data set consists of technical and not of biological repetitions and it is here analyzed for illustration purpose only. Originally filter arrays of 13,824 genes are used. Here endpoints with a sample size less than 2 in at least one group are removed and 13,224 genes remained for the analysis. Furthermore the data set is normalized. Among

other techniques the logarithm to the base of 10 is used. Thus an analysis based on the originally scaled data is omitted.

Results of the parametric procedures: In this data set many significantly expressed genes can be found. By use of the t-test without any multiplicity correction 4,501 null hypotheses are rejected. As a high proportion of the endpoints are highly significant 135 differentially expressed genes are found in combination with the Bonferroni adjustment. Due to this high number of extremely small unadjusted p-values the conservative correction is superior to the procedure with a data-driven order of hypotheses, which is aborted after finding nine significant genes:

endpoint	difference in means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
5979	1.756	7.779	30.165	$1.583 \cdot 10^{-09}$	$1.583 \cdot 10^{-09}$
13618	1.660	7.059	17.968	$9.440 \cdot 10^{-08}$	$9.440 \cdot 10^{-08}$
11600	1.649	6.827	40.775	$1.441 \cdot 10^{-10}$	$9.440 \cdot 10^{-08}$
8563	1.612	6.610	21.691	$2.151 \cdot 10^{-08}$	$9.440 \cdot 10^{-08}$
13585	1.614	6.546	39.803	$1.746 \cdot 10^{-10}$	$9.440 \cdot 10^{-08}$
6629	1.604	6.499	28.522	$2.469 \cdot 10^{-09}$	$9.440 \cdot 10^{-08}$
13653	1.567	6.381	14.299	$5.582 \cdot 10^{-07}$	$5.582 \cdot 10^{-07}$
10284	1.437	6.197	2.582	0.049	0.049
7599	1.492	5.588	42.060	$1.125 \cdot 10^{-10}$	0.049
1689	1.082	5.169	2.230	0.067	0.067

Results of the nonparametric procedures: If it is assumed, that the underlaying populations of the samples are not Gaussian distributed and nonparametric procedures are preferred, then 3,288 significant results are achieved with unadjusted rank sum tests. That is 1,213 less differentially expressed genes compared to the unadjusted t-tests. The Bonferroni adjustment results in no discriminatory endpoint - all adjusted p-values are 1. Due to the small sample sizes this result is expected, as the smallest possible unadjusted p-value is 0.0079. Hence an adjusted p-value less than 1 can not be achieved with the Bonferroni correction. However the nonparametric procedure with a data-driven order of

hypotheses finds six differentially expressed genes:

endpoint	difference in medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value
5979	1.728	1.697	40	0.008	0.008
11600	1.665	1.653	40	0.008	0.008
13585	1.640	1.623	40	0.008	0.008
13618	1.660	1.593	40	0.008	0.008
6629	1.565	1.546	40	0.008	0.008
8563	1.551	1.520	40	0.008	0.008
7018	1.590	1.510	15	0.100	0.100
13653	1.569	1.491	40	0.008	0.100
7599	1.450	1.439	40	0.008	0.100
11277	1.381	1.343	40	0.008	0.100

Endpoint # 7018 stops the procedure prematurely. If $\alpha = 10\%$ would have been chosen, then 11 discriminatory genes could have been found. For comparison the parametric analog would have resulted in 12 rejections of null hypotheses. Summarizing both procedures with a data-driven order of hypotheses have similar results: irrespective of the selected α, out of the first ten endpoints with the highest selector found by the parametric procedure with a data-driven order of hypotheses, eight can be found here as well.

3.3 Simulation results of the proportional power

In this section results of the simulated proportional power of the two procedures are presented. Additionally to the two procedures with a data-driven order of hypotheses results of local tests are shown as well. These are the t-tests or rank sum tests without a correction for multiple testing. Furthermore results from the α-adjustment according to Bonferroni are given. In the graphics the α-adjustment with either the t-test or the rank sum test and the local tests are abbreviated as 'Bonferroni' and 'local' for the use of the parametric or the nonparametric two-sample tests. The procedures with a data-driven order of hypotheses are denoted as 'p-selector' and 'np-selector' for the parametric and nonparametric

version. If not stated otherwise 50 endpoints are analyzed in each simulated experiment, where five of them are differently expressed and 45 are under the null hypothesis. For endpoints under H_1 three endpoints have true means of $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \tau$ and the remaining two variables have $\mu_{1j} = 100 + \tau$ and $\mu_{2j} = 100$, where τ denotes the true treatment effect. All endpoints under H_0 have $\mu_{1j} = \mu_{2j} = 100$. In each scenario two-sided hypotheses are tested. The FWER is set to 5% and each simulation result is computed with 10,000 simulation runs. In all scenarios the random numbers are taken from the standard normal distribution.

3.3.1 Power with increasing treatment effect

The first six graphics show the proportional power for increasing treatment effects and correlations among the endpoints. In the first three figures results of the parametric tests are presented; in the left graphic all endpoints per group are uncorrelated ($\rho_{ijj'} = 0.01$), in the middle one the correlation is set to $\rho_{ijj'} = 0.5$ and in the right graphic $\rho_{ijj'} = 0.999$. The lower three figures are the same as the upper one but for nonparametric tests. For all simulated experiments the remaining parameters are set as $n_i = 7$ and $\sigma_{ij} = 10$.

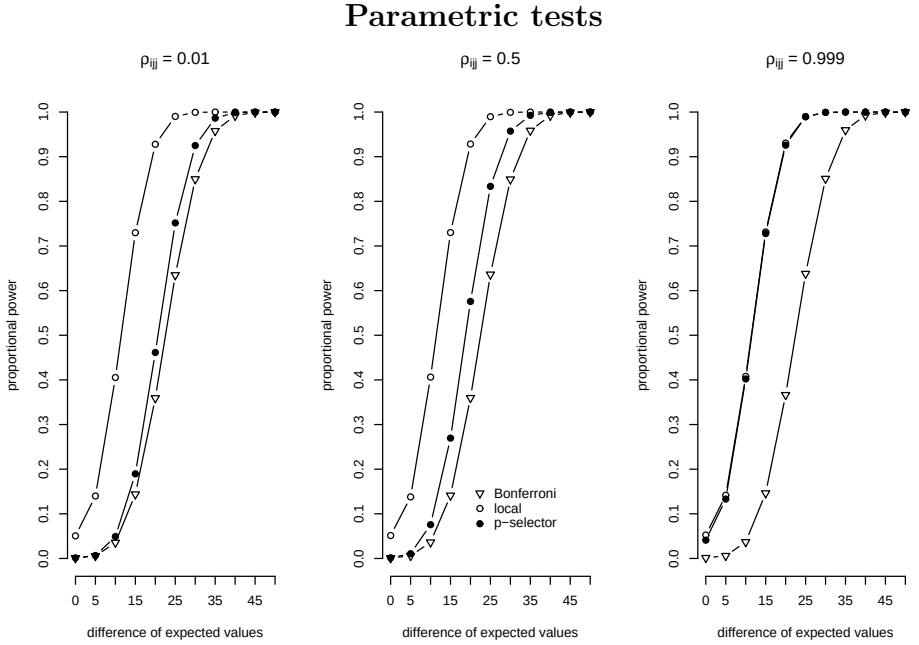

Figure 3.3: Parametric tests for point-zero hypotheses: Power for increasing treatment effect with different correlations

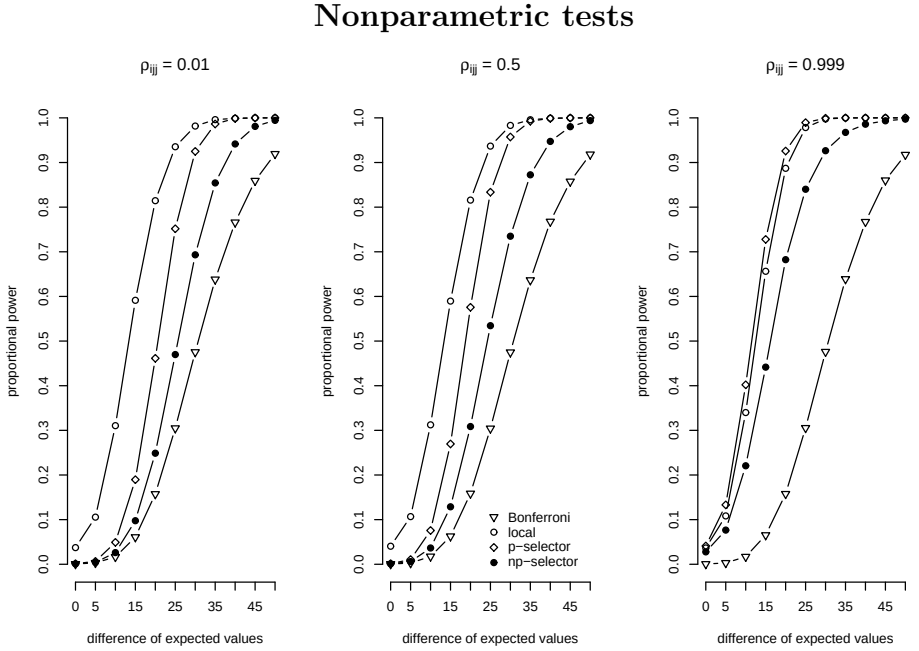

Figure 3.4: Nonparametric tests for point-zero hypotheses: Power for increasing treatment effect with different correlations

In all six scenarios the procedures with a data-driven order of hypotheses show a higher power compared to the Bonferroni adjustment and, as expected, a lower power than the local tests. While the power stays constant for the local tests and the α-adjustment, the procedures with a data-driven order of hypotheses gain power with an increasing correlation among the endpoints. In the extreme case of $\rho_{ijj'} = 0.999$ the difference in power between the parametric local tests and the parametric procedures with a data-driven order of hypotheses is marginal. As it can be expected for Gaussian distributed data, the parametric tests achieve generally a higher power compared to the nonparametric tests. This is true as well for all further simulations in this chapter.

3.3.2 Power for varying sample sizes and different levels of α

These graphics present the dependency of the power on the sample size per group and the selected α. The true difference in means for the endpoints under H_1 is set to $\tau = 15$, the standard deviation is $\sigma_{ij} = 10$ and the correlation is set to $\rho_{ijj'} = 0.3$.

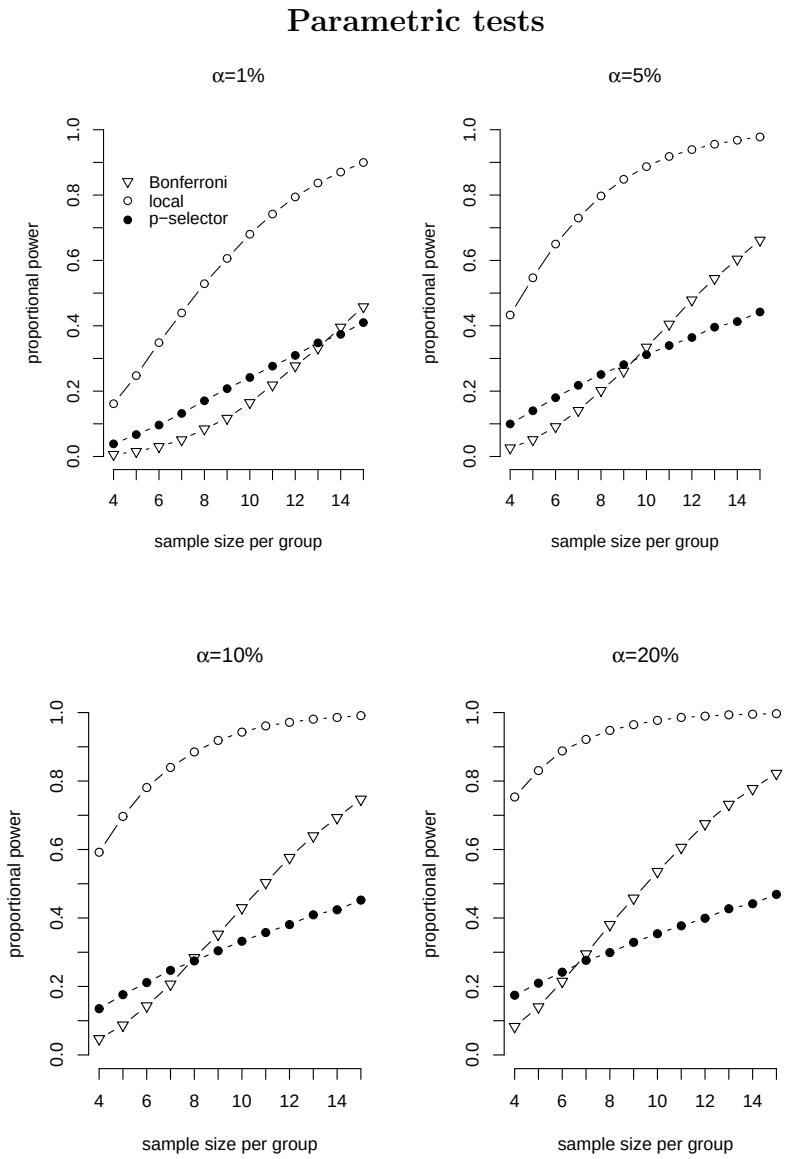

Figure 3.5: Parametric tests for point-zero hypotheses: Power for different sample sizes per group and α

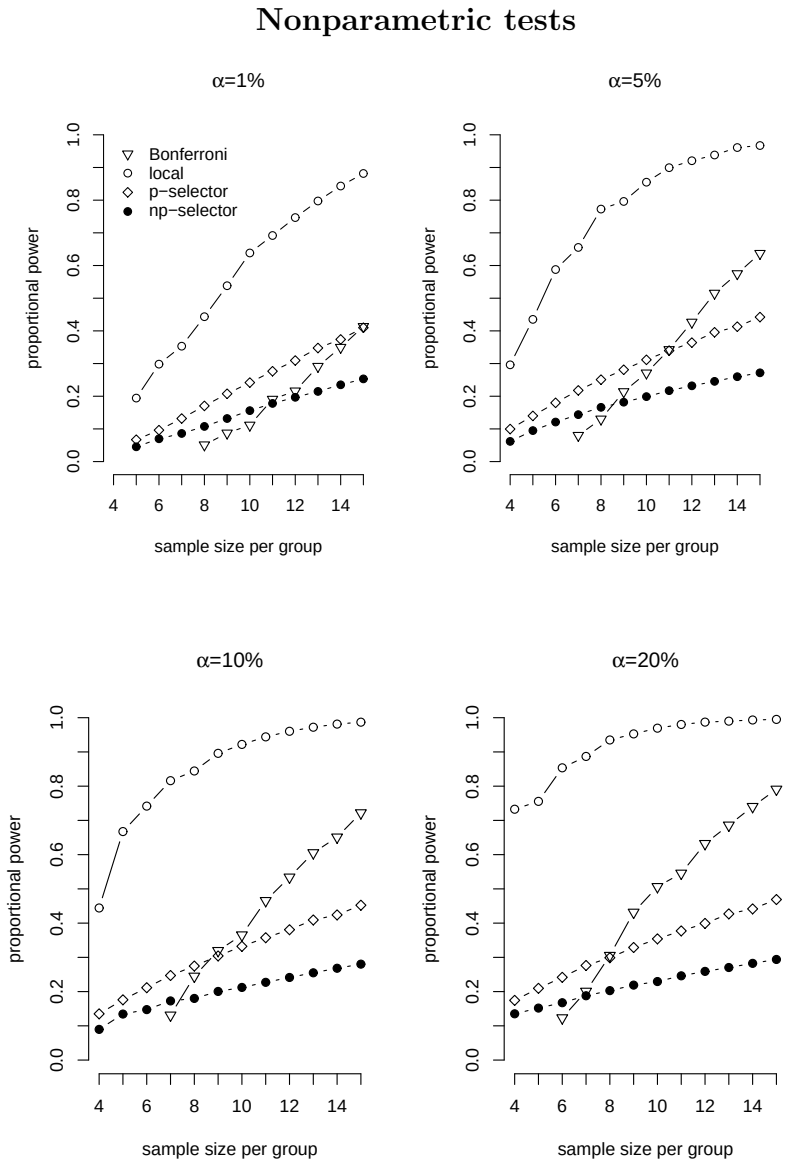

Figure 3.6: Nonparametric tests for point-zero hypotheses: Power for different sample sizes per group and different levels of α

The graphics show an expected behavior of the power curves. With an increasing sample size the power increases. However the Bonferroni adjustment benefits more from the increase in sample sizes than the procedures with a data-driven order of hypotheses, because with an increasing sample size the absolute test statistic becomes larger and the critical value becomes smaller - which is essential for the conservative α-adjustment. The selector procedures do not require extreme test statistics or critical values because the α is not reduced due to multiple testing.

With an increasing α the power increases for all four methods. The highest gain in power shows Bonferroni, because this procedure adjusts the α - an increase of the Type I error has a direct effect. For the procedures with a data-driven order of hypotheses the p-values are sorted in an independent order of the α. Thus the probability to find significant p-values increases with an increasing α, but the order of the p-values stay the same compared to a lower Type I error.

These graphics clearly show that the α-adjustment of Bonferroni is not appropriate for high-dimensional data with small sample sizes combined with the selection of a small α. For the parametric tests it lacks power and for the nonparametric tests it is even impossible to compute any significant result because of the discreteness of the rank sum test. However if α is set to 10% or even higher, the adjustment can achieve a comparably good power. This may be taken into account for experiments which are planned for screening.

3.3.3 Adapted expected difference in means and varying sample size

As it could be seen from the latter graphic for small sample sizes the procedures with a data-driven order of hypotheses are more powerful than the α-adjustment according to Bonferroni. The next graphic visualizes the influence of the samples size on the multiple testing methods while the test statistics stay constant. In particular an adapted expected true difference in means is computed, which decreases with increasing sample sizes.
As a reference curve local test results are plotted. The 'adapted difference' is selected, such that the proportional power of the local test is around 80% in a simulation setting with a true difference of means of 20, $n_1 = n_2 = n = 5$, $\sigma_{ij} = 10$ and $\rho_{ijj'} = 0.3$.
To compute the adapted expected true difference in means the non-centrality parameter of the t-test is used:

$$\nu = \frac{\mu_2 - \mu_1}{\sigma\sqrt{\frac{2}{n}}} \Leftrightarrow \mu_2 - \mu_1 = \nu \cdot \sigma \cdot \sqrt{\frac{2}{n}}. \tag{3.10}$$

With the input of the above proposed simulation parameters the adapted expected difference in means of 20 is:

$$120 - 100 = 3.162278 \cdot 10 \cdot \sqrt{0.4}. \tag{3.11}$$

For any sample size n the difference becomes:

$$\mu_2 - \mu_1 = \sqrt{3.162278^2} \cdot 10 \cdot \sqrt{\frac{2}{n}} = \sqrt{10} \cdot \sqrt{\frac{200}{n}} = \sqrt{\frac{2000}{n}}. \tag{3.12}$$

By the use of the latter equation the true means for endpoints under H_1 are set to $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \sqrt{2000/n}$.

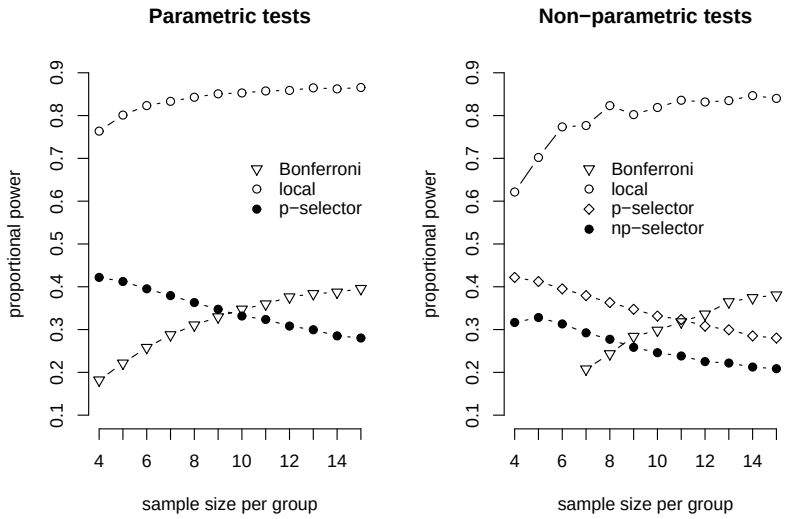

Figure 3.7: Tests for point-zero hypotheses: Power for different sample sizes with adapted true ratio

While the test statistics stay constant, the power of the Bonferroni adjustment and of the local tests increases with an increase of the sample sizes because the degrees of freedom increase. And with increasing degrees of freedom the critical values and the p-values become smaller, respectively. However the power lines of the procedures with a data-driven order of hypotheses show a monotone decrease (parametric) or a small increase and afterwards a monotone decrease (nonparametric). The reason for this decrease is that a high selector is received only with a large treatment effect. An increase in sample sizes has no impact. As the treatment effect decreases with the increasing sample size the procedures with a data-driven order of hypotheses lack power.

3.3.4 Simulations with increasing disturbance

The following graphic shows the power for varying variances among the endpoints. For each endpoint and group separately the true standard deviation is computed as $\sigma_{ij} = 10 + u \cdot d$ ($u \sim U(-5, 5)$), where u takes values from 0 to 2 in steps of 0.1 units and U(-5,5) is the

uniform distribution on the interval from -5 to 5. The other parameters are $\mu_{2j} - \mu_{1j} = 20$ and $\rho_{ijj'} = 0.3$, $n_1 = n_2 = 7$.

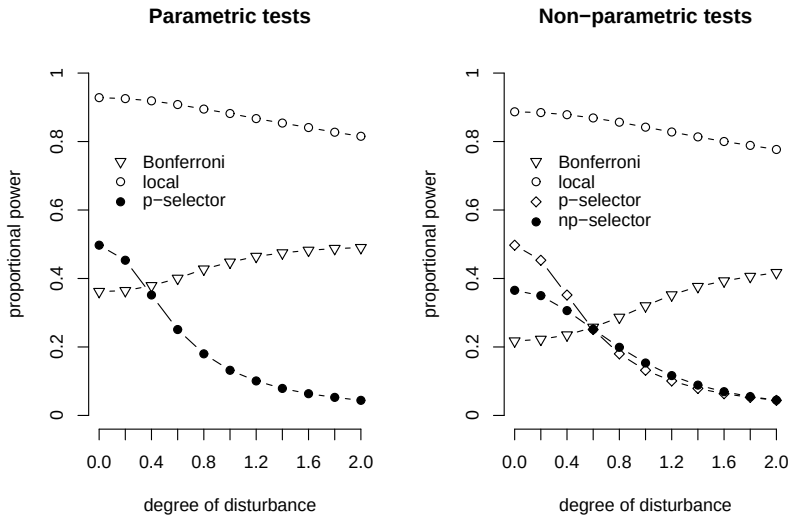

Figure 3.8: Tests for point-zero hypotheses: Power for increasing disturbance

On the left end of the curves the true standard deviation is 10. For this scenario the power for the local tests is between 90% and 100%. With the increasing disturbance factor the local tests loose power, because for some endpoints the variance increases. The variance decrease for the other endpoints has no impact, because these endpoints already show a significance (as the variance is lower than 10 - and with 10 the power is close to 100%). In contrast to the local tests the Bonferroni adjustment gains power, because with $\sigma_{ij} = 10$ less than 50% of the endpoints under H_1 show a non-significant result. Hence for more than 50% of the endpoints the power can increase with a reduction of the variance. At the right end of the graphics the power of the Bonferroni methods converges to 50% - with a disturbance of 2 50% of the endpoints profit from the variance reduction and the other 50% have non-significant results as the variance is too large.

The procedures with a data-driven order of hypotheses lose power constantly, because the selectors are sensitive to the variance. If the variance per endpoint increases, then the treatment effect, which is equal to the variance of the pooled groups per endpoint, is

masked by the individual variances of the two samples.

In the right graphic the proportion of the power of the parametric procedure with a data-driven order of hypotheses compared to the nonparametric one changes: the nonparametric procedure is more robust for an increasing variance heterogeneity among the endpoints.

Chapter 4

Relevance-shifted testing procedure on difference

This chapter introduces the procedures with a data-driven order of hypotheses for tests on relevant differences. Instead of testing the point-zero hypothesis as in the former chapter, it is now of interest whether the treatment effect is significantly smaller or larger than a-priori chosen relevance thresholds. Both algorithms for a parametric and a nonparametric relevance-shifted approach with a data-driven order of hypotheses are shown.

Before the approaches are introduced some remarks concerning all further procedures with a data-driven order of hypotheses irrespective of parametric or nonparametric versions and test on difference or ratio have to be given. As denoted above all further procedures include a relevance-shifted test. In combination with such a test the selector statistics from the former chapter can not be used, because those procedures exceed the FWER in a high degree. For example if the relevance-shifted t-test (introduced below) is used with the selector statistic $w_j = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_j)^2$, then in some tested scenarios empirical error rates around 50% and even higher can be observed. Hence selectors corresponding to the applied test statistic have to be developed. All motivations and problems of appropriate selector statistics discussed here are valid for the procedures in the further chapters.

In this work first testing procedures for relevant differences are presented. However the

entire work on relevance-shifted tests with a data-driven order of hypotheses has its origin in a selector for the parametric test on ratio. An equation for this selector can be derived from the theory of LÄUTER *et al.* (1996), as it will be presented in section 5.1.1 of chapter 5. This selector considers the relevance shift in the hypotheses system and in combination with an appropriate test on ratio it is an exact α-level test. The selector statistic however has two disadvantages. First if the aim is to test two-sided the analyst needs to select two relevance thresholds; one for a relevant under-expression and one for the relevant over-expression. But the selector allows only the use of one threshold. Furthermore the selector regards only one margin of the null hypothesis. In contrast to the point-zero hypothesis, where the null hypothesis states that the difference in treatment means is 0, the relevance-shifted hypothesis covers the ratio of the treatment effects between 1 and the relevance threshold. But with the derived selector the testing procedure focuses only on the margin of the hypothesis, that is the point where the ratio of the means is equal to the relevance threshold. This results in a lack of power, as it will be shown in chapter 5 for the parametric relevance-shifted test on ratio with a data-driven order of hypotheses. To overcome these two problems modifications had to be constructed which control the FWER and achieve a higher power than the standard multiple testing methods. These modifications are generated in an empirical way. Hence no proofs can be presented and the empirical control of the FWER is shown by simulations. Furthermore all procedures with a data-driven order of hypotheses for relevant differences or ratios are based on the ideas of the modified procedures with a data-driven order of hypotheses. Therefore all testing procedures are approximations.

4.1 Parametric procedures

The relevance-shifted t-test: As the aim of this work is testing against a relevance threshold, a relevance-shifted t-test is used, which includes the relevance threshold δ_{side} and takes the values $\delta_{lower} \leq 0$ and $\delta_{upper} \geq 0$. Throughout this work the thresholds are set to the special case of $-\delta_{lower} = \delta_{upper}$. In the univariate case the null hypotheses to

test for a relevant increase is $H_0 : \mu_2 - \mu_1 \leq \delta_{upper}$. One-sided against a relevant decrease is tested as $H_0 : \mu_2 - \mu_1 \geq \delta_{lower}$ and for two-sided testing the following null hypothesis $H_0 : \delta_{lower} \leq \mu_2 - \mu_1 \leq \delta_{upper}$ is rejected in favor of $H_1 : \mu_2 - \mu_1 < \delta_{lower}$ or $\mu_2 - \mu_1 > \delta_{upper}$. The critical values are the same as for the unshifted t-test. However compared to the t-test (equation (2.7) on page 16), the test statistic changes to:

$$t^{\delta}_{side} = \frac{\bar{x}_2 - \bar{x}_1 - \delta_{side}}{s_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{4.1}$$

While for one-sided testing equation (4.1) is used with δ_{lower} or δ_{upper} depending on the direction of interest, for the two-sided test both test statistics are computed. The null hypothesis $H_0 : \delta_{lower} \leq \mu_2 - \mu_1 \leq \delta_{upper}$ is rejected if either $t^{\delta}_{lower} \leq t_{df,\alpha/2}$ or $t^{\delta}_{upper} \geq t_{df,1-\alpha/2}$.

Here the two-sided p-values are denoted by p^{δ}.

For data sets with multiple endpoints the relevance-shifted t-test is computed for each endpoint separately. Thus beside the inclusion of index j for the jth endpoint, the test statistic stays the same.

4.1.1 The shift-selector procedure

The procedure with a data-driven order of point-zero hypotheses uses the two-sample t-test for unpaired data. In comparison the procedure with a data-driven order of hypotheses for relevance-shifted tests uses the t-test including a relevance shift. A straight-forward approach is the inclusion of the relevance threshold(s) in the computation of the selector statistics as well. In particular a data shift prior to the computation of the selector has to be done.

This procedure is abbreviated as 'shift-selector' in the further thesis.

1. Select relevance thresholds $\delta_{lower} \leq 0, \delta_{upper} \geq 0$.
2. For each endpoint j:
   (a) Compute the two-sided p-value p^{δ}_j of the relevance-shifted t-test by use of equation (4.1).

(b) Shift the data of the first treatment group to:

$$x^*_{1jk} = \begin{cases} x_{1jk} + \delta_{upper}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} \geq 0 \\ x_{1jk} + \delta_{lower}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} < 0. \end{cases} \tag{4.2}$$

(c) Calculate the selector statistic

$$w_j = \sum_{k=1}^{n_1} (x^*_{1jk} - \bar{x}^*_j)^2 + \sum_{k=1}^{n_2} (x_{2jk} - \bar{x}^*_j)^2. \tag{4.3}$$

with the mean of the combined samples per endpoint computed as $\bar{x}^*_j = (\sum_{i=1}^{n_1} x^*_{1jk} + \sum_{i=1}^{n_2} x_{2jk})/N$.

3. Sort the m p-values for decreasing selectors w_j.

4. Compare the jth ordered p-value with the unadjusted α. It is significant, if $p^\delta_j <$ unadjusted α.

5. Stop at the first non-significance and accept for all further endpoints the null hypothesis.

For one-sided testing the data transformation changes to $x^*_{1jk} = x_{1jk} + \delta_{upper}$ (increase) or $x^*_{1jk} = x_{1jk} + \delta_{lower}$ (decrease), respectively. The remaining procedure stays the same except that one-sided p-values are used.

As it will be shown in section 4.3 this procedure controls empirically the FWER in the weak and in the strong sense. It can not however be recommended in practice because of its low proportional power. The following graphics show the proportional power of this procedure in a two-sided testing scenario where five out of 50 endpoints are under the alternative hypothesis. Three of the endpoints have a difference in means equal or greater than 0; their expected values are set to $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \tau^{H_1}_\delta$, where $\tau^{H_1}_\delta$ denotes the selected difference in means for endpoints under H_1. The other two endpoints under H_1 have means of $\mu_{1j} = 100 + \tau^{H_1}_\delta$ and $\mu_{2j} = 100$. $\tau^{H_1}_\delta$ takes the values from a) 0 to 50 and b) 400 to 460.

Each endpoint under H_0 receives a random true difference in means between δ_{lower} and δ_{upper} in steps of five units. The random differences in means per endpoint are uncorrelated. If not stated otherwise for all further simulation settings with tests on relevant differences the expected values for endpoints under H_0 are selected as described here.
If $\tau_\delta^{H_0} < 0$ then expected values are set to $\mu_{1j} = 100 + |\tau_\delta^{H_0}|$ and $\mu_{2j} = 100$. Otherwise the true mean values are $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \tau_\delta^{H_0}$. The relevance thresholds are set to a) $\delta_{lower} = \delta_{upper} = 0$ and b) $-\delta_{lower} = \delta_{upper} = 400$. For both scenarios the other parameters are set to $n_i = 5$, $\sigma_{ij} = 10$, $\rho_{ijj'} = 0.3$ and $\alpha = 5\%$. The proportional power of all five endpoints under H_1 is computed. Although this includes endpoints with a difference in means less than 0, for the sake of simplicity only the values of $\tau_\delta^{H_1}$ are plotted on the abscissae.

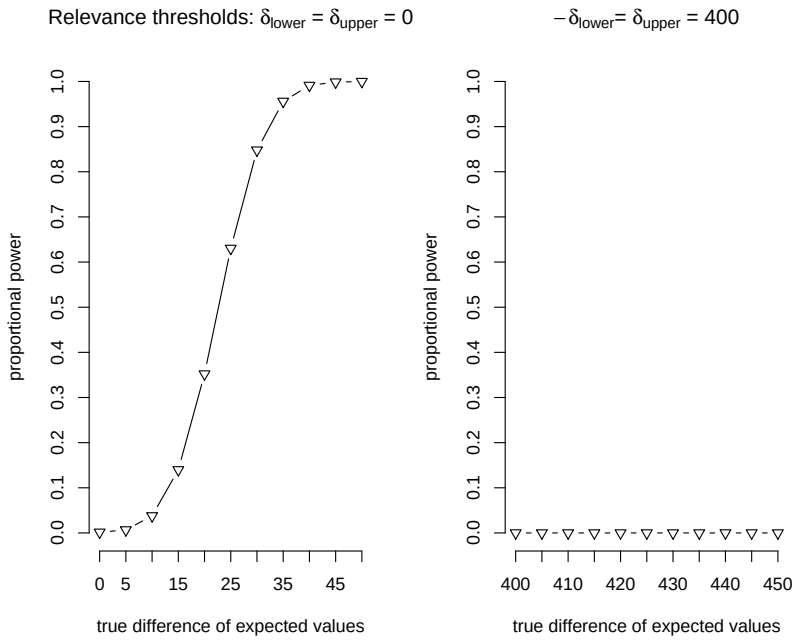

Figure 4.1: Power of the shift-selector procedure for varying differences in expected values and relevance thresholds

The procedure achieves a high power when it reduces to the procedure for point-zero hypotheses ($\delta_{lower} = \delta_{upper} = 0$). However with increasing relevance thresholds the approach

lacks power. The following graphic shows the reason for this problem. It depicts the value of the selector for increasing differences in means in a two-sided testing scenario. To construct this graphic for each point on the abscissae a data set consisting of two samples and five endpoints is created. The expected values are set to $\mu_{1j} = 10 + |\tau_\delta^{H_1}|$ and $\mu_{2j} = 10$ for $\tau_\delta^{H_1} < 0$ and $\mu_{1j} = 10$ and $\mu_{2j} = 10 + \tau_\delta^{H_1}$ for $\tau_\delta^{H_1} \geq 0$, with $\tau_\delta^{H_1}$ taking values from -30 to 30 in steps of one unit. The true standard deviation for each endpoint is set to $\sigma_{ij} = 1$ and the endpoints are uncorrelated. The sample size per group is $n_i = 20$. To reduce noise the mean of the five selectors per $\tau_\delta^{H_1}$ is computed and plotted against $\tau_\delta^{H_1}$. The relevance thresholds are set to $-\delta_{lower} = \delta_{upper} = 10$.

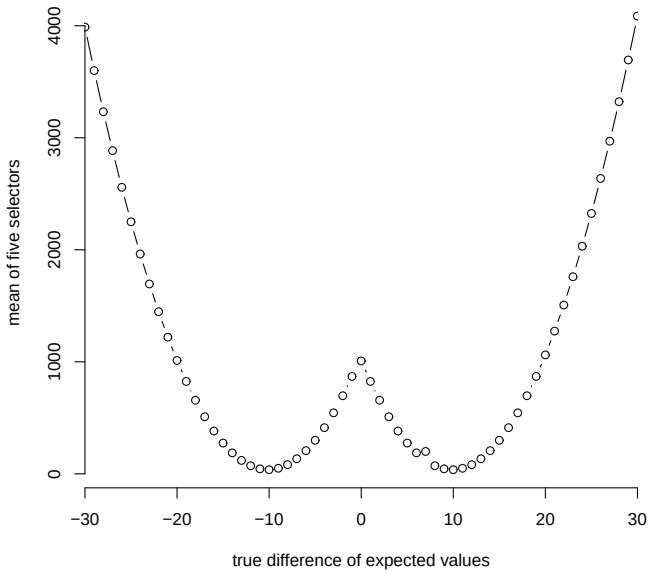

Figure 4.2: Shift-selector: minima at δ_{side} and local maximum at 0

The reason for the minima at the relevance thresholds and the local maximum at 0 is the shift of the first treatment group by the amount of δ_{side}. It results in a small variability between the two groups when the difference in means is approximately equal to the threshold. In particular, if the difference in means is equal to δ_{side}, then the selector takes the value of the numerator of the variance among both groups, because the difference in

means is reduced to 0 after the shift. To the left and to the right of these local minima the selector increases, because to the left of δ_{lower} and to the right of δ_{upper} the selector is a compound of the sample variance and the significant treatment effect minus a comparably smaller relevance shift. And for differences in means between δ_{lower} and δ_{upper} the selector increases when the difference in means changes towards 0, because then the selector is a compound of the variance and the non-relevant treatment effect minus a comparably larger relevance threshold. Finally the most extreme selector statistic for endpoints with non-relevant mean differences is achieved, when the difference is 0, because then the selector reflects the sample variance and a pseudo-treatment effect of δ_{side}.
This local maximum of the selector at the mean difference of 0 is the reason for the lack of the power. For example in this scenario endpoints with a difference in means of 0 have similar selectors as endpoints with a difference of -20 or 20. This means a significant endpoint needs a difference larger than $|\pm 20|$ to prove a relevant difference of at least $|\pm 10|$. The problem increases with increasing relevance thresholds $|\delta_{side}|$.

4.1.2 The δ-shift procedure

The former procedure lacks power due to endpoints under H_0 with $\delta_{lower} < \bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper}$. A data transformation to shift the difference in means equal to the relevance threshold(s) solves the problem. In particular, for testing one-sided for a relevant increase all endpoints with a difference in means less than δ_{upper} are transformed, such that their difference is equal to the threshold. Equivalently, for testing one-sided for a relevant decrease all endpoints with a difference greater than the lower threshold are shifted to δ_{lower}. Finally, for two-sided testing all endpoints with differences in means between δ_{lower} and 0 are set to δ_{lower} and differences between 0 and δ_{upper} are set to the value of the upper threshold. This data transformation is used for the computation of the selector only. The transformed endpoints cannot cause an α-error, because the p-value of the shifted t-test for endpoints with a difference in means not exceeding the threshold will always be larger than α.
In the following graphics and tables this procedure will be abbreviated as the 'δ-shift' pro-

cedure.

Compared to the two-sided testing procedure of the former section 4.1.1, 2. (c) changes to:

2. (c) For each endpoint with $\bar{x}_{2j} - \bar{x}_{1j} \geq 0$ transform the data independently to

$$x^*_{2jk} = \begin{cases} \delta_{upper} + \bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}; & \forall\, j \mid 0 \leq \bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper} \\ x_{2jk}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} \geq \delta_{upper}. \end{cases} \tag{4.4}$$

And for endpoints with $\bar{x}_{2j} - \bar{x}_{1j} < 0$ x_{2jk} changes to:

$$x^*_{2jk} = \begin{cases} \delta_{lower} + \bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}; & \forall\, j \mid 0 > \bar{x}_{2j} - \bar{x}_{1j} > \delta_{lower} \\ x_{2jk}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} \leq \delta_{lower}. \end{cases} \tag{4.5}$$

Calculate the selector statistic:

$$w_j = \sum_{k=1}^{n_1} (x^*_{1jk} - \bar{x}^*_j)^2 + \sum_{k=1}^{n_2} (x^*_{2jk} - \bar{x}^*_j)^2. \tag{4.6}$$

with the mean of the combined relevance-shifted first group and the δ-transformed second group per endpoint computed as $\bar{x}^*_j = (\sum_{i=1}^{n_1} x^*_{1jk} + \sum_{i=1}^{n_2} x^*_{2jk})/N$.

As well as for the two-sided procedure, for the one-sided testing problem a data transformation is included and the computation of the selector changes. In case of testing for a relevant increase x_{2jk} is transformed to $x^*_{2jk} = \delta_{upper} + \bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}$ if $\bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper}$, otherwise $x^*_{2jk} = x_{2jk}$. And for testing against a relevant decrease $x^*_{2jk} = \delta_{lower} + \bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}$ if $\bar{x}_{2j} - \bar{x}_{1j} > \delta_{lower}$, otherwise $x^*_{2jk} = x_{2jk}$. In both cases the selector statistic is computed by equation (4.6). In the former chapter, it was indicated that the procedures with a data-driven order of point-zero hypotheses may lack in power if one-sided hypotheses are tested. For this and all further procedures with a data-driven order of relevance-shifted hypotheses the problem does not occur. If tested one-sided on for example increase, then the data of an endpoint, which would be significant in the opposite direction, is transformed, such that the selector is comparably small.

The following graphic shows the effect of the transformation to a difference in means of δ_{side}. It is generated under the same conditions as graphic 4.2.

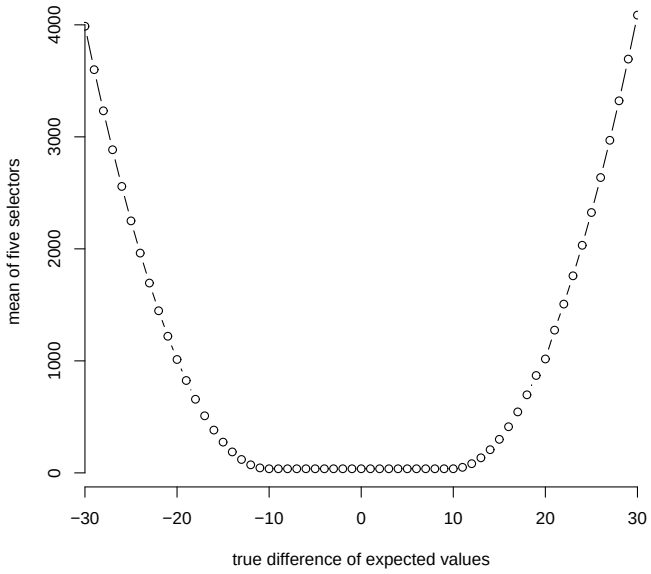

Figure 4.3: δ-shift with minimal selector between δ_{lower} and δ_{upper}

Note that all selector statistics of endpoints with the differences between δ_{lower} and δ_{upper} are exactly equal. After the two data transformations - δ-shift and the transformation by the relevance thresholds - all mean differences of these endpoints are 0. In this case the selector is a measurement of the numerator of the variance only. And due to the same starting value for each simulation run the variances are always equal. Thus in real data sets with approximately equal variances among the endpoints as well as in the further simulations with more than one endpoint the selector statistics of such endpoints are roughly but not exactly the same.

Further it should be noted, that the selector statistic still depends on δ_{side}. The following graphic shows this dependency of the selector statistic on the thresholds. To the former graphic two more curves are added; both are generated under the same conditions but with $-\delta_{lower} = \delta_{upper} = 50$ and $-\delta_{lower} = \delta_{upper} = 100$.

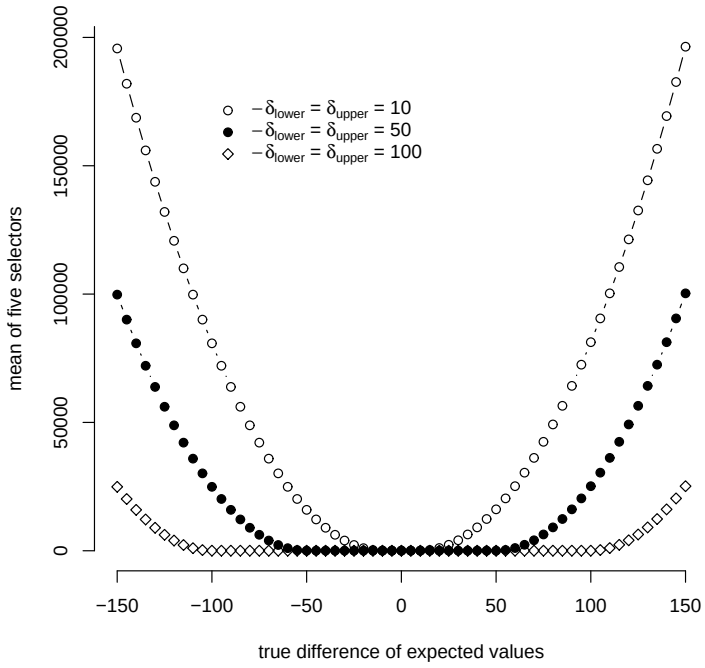

Figure 4.4: δ-shift selector for different δ_{side}

From the graphic it can be seen, that the selector statistics have a small value if the difference in means is between the relevance thresholds and the selectors increase when the difference exceeds one of the thresholds. Summarizing it can be said, that although the data is transformed by the amount of a relevance threshold, the selector is still dependent on the relevance criteria.

In section 4.3 and in more detail in the appendix it will be shown, that this procedures controls empirically the FWER in most cases. Slight exceeds may appear if one-sided hypotheses are tested.

Further with the inclusion of the data transformation the procedure with a data-driven order of hypotheses can achieve a high power. The behavior of the proportional power compared to other procedures is graphically shown in chapter 7.

4.1.3 The random$^{\delta}$ procedure

To ensure a control of the FWER for one-sided testing, another approach is presented. The δ-shift method transforms the data of the second treatment group such that endpoints with difference in means between 0 and δ_{side} receive exactly $\bar{x}_{2j} - \bar{x}_{1j} = \delta_{side}$. For the further method it is assumed that the FWER is exceeded because the selector statistics of the transformed endpoints lack variability and receive homogeneously small selectors. It is assumed that endpoints under H_0 with $\bar{x}_{2j} - \bar{x}_{1j} < \delta_{lower}$ or $\bar{x}_{2j} - \bar{x}_{1j} > \delta_{upper}$ have an advantage concerning the sorting by the selectors, because they have a larger selector as the transformed endpoints. Hence because of the construction of homogeneously small selectors the probability that the procedure stops before a false positive endpoint is declared as significant is decreased.

The following procedure transforms the second treatment group as the δ-shift method. However it generates random data from the normal distribution, such that the difference of means is not exactly equal but close to (one of) the relevance threshold(s) and the standard deviation is the same as the original pooled one $s_{pool,j}$. Summarizing, the transformed endpoints of the procedure have nearly the same difference in means and exactly the same pooled standard deviation as the δ-shift method. Therefore the variability of the selectors of the procedure is only slightly increased.

The following two histograms show the frequencies of the selector statistics of the two procedures. In each setting 1000 endpoints are generated with true means of $\mu_{1j} = 100$ and $\mu_{2j} = 200$. All endpoints have a sample size per group of 10, the correlation is set to 0.3, the standard deviation is 10 and the relevance thresholds are -100 and 100. Only selector statistics of transformed endpoints are plotted.

As in this plot and all further graphics and tables the procedure is abbreviated as 'random$^{\delta}$'.

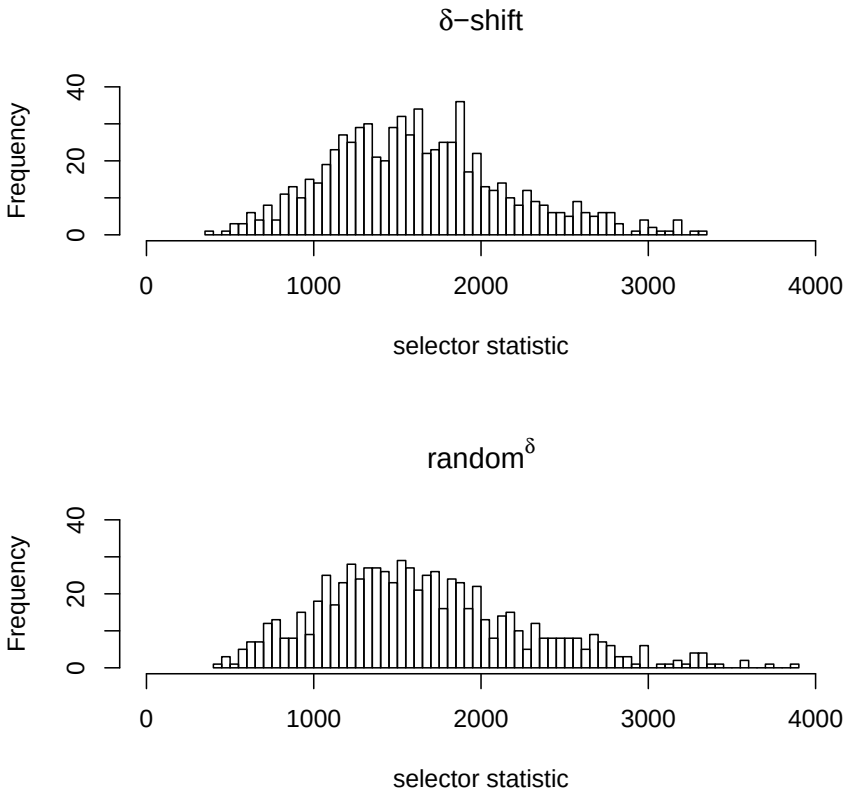

Figure 4.5: δ-shift and random$^{\delta}$: variation of the selector

The distribution of the selectors from the random$^{\delta}$ procedure is slightly wider compared to the one of the δ-shift method. With the inclusion of the small proportion of noise, the selectors become more variable.

The algorithm of the random$^{\delta}$ method is similar to the procedure of section 4.1.1, only 2. (c) changes to:

2. (c) Replace for each endpoint with $\delta_{lower} \leq \bar{x}_{2j} - \bar{x}_{1j} \leq \delta_{upper}$ separately x_{2jk} with Gaussian distributed random numbers $x^{z}_{2jk} \sim N(\mu = \bar{x}_{1j}, \sigma_{ij} = 1)$. Independently for each endpoint with $\bar{x}_{2j} - \bar{x}_{1j} \geq 0$ transform the data of the second

treatment group to

$$x^*_{2jk} = \begin{cases} \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + \delta_{upper}; & \forall\, j \mid 0 \leq \bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper} \\ x_{2jk}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} \geq \delta_{upper}. \end{cases} \tag{4.7}$$

And for endpoints with $\bar{x}_{2j} - \bar{x}_{1j} < 0$ the data of the second treatment group is

$$x^*_{2jk} = \begin{cases} \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + \delta_{lower}; & \forall\, j \mid 0 > \bar{x}_{2j} - \bar{x}_{1j} > \delta_{lower} \\ x_{2jk}; & \forall\, j \mid \bar{x}_{2j} - \bar{x}_{1j} \leq \delta_{lower}. \end{cases} \tag{4.8}$$

Calculate the selector statistic according to equation 4.6, which is the selector used in the former section.

The changes for the one-sided testing problem are: for a test on increase replace for each j with $\bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper}$ independently the x_{2jk} by $x^z_{2jk} \sim N(\mu = \bar{x}_{1j}, \sigma_{ij} = 1)$ and transform x^z_{2jk} to $x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + \delta_{upper}$. All other endpoints have $x^*_{2jk} = x_{2jk}$. Equivalently for testing on decrease endpoints with $\bar{x}_{2j} - \bar{x}_{1j} > \delta_{lower}$ transform x^z_{2jk} to $x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + \delta_{lower}$ and all other j have $x^*_{2jk} = x_{2jk}$. For both one-sided testing problems the selector according to equation (4.6) is computed for each endpoint separately.

As discussed in section 4.3 this procedure controls empirically the FWER in the weak and in the strong sense for more tested scenarios compared to the δ-shift procedure. Some further selected simulation results can be found in the appendix on page 156. The proportional power of the random$^\delta$ procedure is slightly less or in some cases equal as the δ-shift method. Again graphical results are shown in chapter 7 . For practical use this procedure has the disadvantage, that the result is dependent on the random number generator. That is, in extreme cases the number of significant endpoints can depend on the starting value.

4.2 Nonparametric procedure

Analog to the parametric test a nonparametric version can be constructed. As in the nonparametric procedure for point-zero hypotheses the interquartile range is used as selector. And in total analogy to the parametric procedures for relevant differences prior to

the computation of the selector statistic a relevance shift of the data is applied. Furthermore to achieve an appropriate power the additional data transformation is used. Hence the method presented in this section is similar to the δ-shift procedure. A nonparametric analog to the parametric random$^{\delta}$ method could not be found for two reasons. First, the generation of the random numbers needs an a-priori knowledge of the distribution of the data. And second, as only one type of distribution can be implemented in the data transformation, all data to be replaced would need to have the same distribution.

The procedure makes the same assumptions concerning the data as described in the former chapter for the nonparametric test. However as in this chapter a relevant difference is of interest, it is now tested, whether the estimate of Δ_j of endpoint j exceeds the a-priori selected relevance threshold(s) δ_{side}. In terms of two-sided hypotheses this is stated as $H_{0j} : \delta_{lower} \leq \Delta_j \leq \delta_{upper}$ versus $H_{1j} : \Delta_j < \delta_{lower}$ or $\Delta_j > \delta_{upper}$. As with the relevance-shifted nonparametric two-sample test for independent samples, the rank sum test according to Wilcoxon (1945) is used.

The following section presents this relevance-shifted test and afterwards the procedure is discussed.

The relevance-shifted rank sum test: The well-known rank sum test proposed by Wilcoxon (1945) and introduced in section 3.1.2 is usually used for point-zero hypotheses and not for testing against relevance criteria. Following Hollander and Wolfe (1999) the rank sum test can be applied to test against a specified relevance threshold δ_{side}. For this purpose a pseudosample is formed, such that $x^*_{1k} = x_{1k} + \delta_{side}$ and the rank sum test is computed from the samples x^*_{1k} and x_{2k}.

As for the test for the point-zero hypothesis both the exact and the asymptotic version are presented.

Exact nonparametric test for relevant differences: The exact two-sided relevance-shifted Wilcoxon rank sum test is computed as follows: Shift the first treatment group, such that $x^*_{1k} = x_{1k} + \delta_{lower}$, sort the combined samples x^*_{1k} and x_{2k} in an increasing order

and rank them. Denote the ranks of the second treatment group by r_{2k}. Calculate the sum of the r_{2k}:

$$W_{lower} = \sum_{k=1}^{n_2} r_{2k}. \tag{4.9}$$

Repeat the procedure with the ranks taken among $x^*_{1k} = x_{1k} + \delta_{upper}$ and x_{2k} and compute

$$W_{upper} = \sum_{k=1}^{n_2} r_{2k}. \tag{4.10}$$

The two-sided null hypothesis $H_0 : \delta_{lower} \leq \Delta \leq \delta_{upper}$ is rejected, if either $W_{upper} \geq w_{\alpha/2}$ or $W_{lower} \leq n_2(n_1 + n_2 + 1) - w_{\alpha/2}$. In the one-sided case to test for a relevant increase H_0 reduces to $H_0 : \Delta \leq \delta_{upper}$ and it is rejected with $W_{upper} \geq w_\alpha$. For testing against a relevant decrease $H_0 : \Delta \geq \delta_{lower}$ is declined if $W_{lower} \leq n_2(n_1 + n_2 + 1) - w_\alpha$.

Asymptotic nonparametric test for relevant differences: For the computation of the asymptotic test the expectation and the variance are required again. These are denoted by:

$$E(W_{side}) = \frac{n_2}{N} \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik} \tag{4.11}$$

and

$$Var(W_{side}) = \frac{n_1 n_2}{N^2(N-1)} \left\{ N \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik}^2 - \left(\sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik} \right)^2 \right\}. \tag{4.12}$$

Hence the large-sample approximation of the Wilcoxon test is

$$W_{side}^{approx} = \frac{W_{side} - E(W_{side})}{\{Var(W_{side})\}^{1/2}}. \tag{4.13}$$

The two-sided null hypothesis is rejected, if either $|W_{lower}^{approx}|$ or W_{upper}^{approx} is greater or equal to $z_{1-\alpha/2}$. In the one-sided case H_0 is declined if $W_{upper}^{approx} \geq z_{1-\alpha}$ (increase) and $W_{lower}^{approx} \leq z_\alpha$ (decrease).

The two-sided p-values of the exact and the asymptotic rank sum test are denoted $p^{np-\delta}$.

4.2.1 The np-δ-shift procedure

By use of the relevance-shifted data and the data transformation of definite non-significant endpoints the nonparametric procedure with a data-driven order of hypotheses to test for a relevant difference is defined as:

1. Select relevance thresholds $\delta_{lower} \leq 0, \delta_{upper} \geq 0$.

2. For each endpoint j:

   (a) Depending on the sample size compute the p-value of the relevance-shifted rank sum test $p_j^{np-\delta}$ either exact or asymptotic by use of the test statistics in section 4.2.

   (b) Order the observations of both samples separately, such that $x_{ijk}^{(1)} \leq \ldots \leq x_{ijk'}^{(n_i)}$, where $k \neq k'$. Calculate for both samples the median $\tilde{x}_{ij}$. If n_i is odd, then

$$\tilde{x}_{ij} = x_{ijk}^{((n_i-1)/2+1)} \tag{4.14}$$

   and if n_i is even, then

$$\tilde{x}_{ij} = \frac{x_{ijk}^{(n_i/2)} + x_{ijk}^{(n_i/2+1)}}{2}. \tag{4.15}$$

   For each endpoint with $\tilde{x}_{2j} - \tilde{x}_{1j} \geq 0$ transform independently the data of the second group to

$$x_{2jk}^* = \begin{cases} x_{2jk} - \tilde{x}_{2j} + \tilde{x}_{1j} + \delta_{upper}; & \forall\, j \mid 0 \leq \tilde{x}_{2j} - \tilde{x}_{1j} < \delta_{upper} \\ x_{2jk}; & \forall\, j \mid \tilde{x}_{2j} - \tilde{x}_{1j} \geq \delta_{upper} \end{cases} \tag{4.16}$$

   And for endpoints with $\tilde{x}_{2j} - \tilde{x}_{1j} < 0$ transform the x_{2jk} to

$$x_{2jk}^* = \begin{cases} x_{2jk} - \tilde{x}_{2j} + \tilde{x}_{1j} + \delta_{lower}; & \forall\, j \mid 0 > \tilde{x}_{2j} - \tilde{x}_{1j} > \delta_{lower} \\ x_{2jk}; & \forall\, j \mid \tilde{x}_{2j} - \tilde{x}_{1j} \leq \delta_{lower} \end{cases} \tag{4.17}$$

   If $\tilde{x}_{2j} - \tilde{x}_{1j} < 0$ compute the selector IQR_j among the pooled treatment groups $x_{1jk} + \delta_{lower}$ and x_{2jk}^*. Otherwise it is calculated among $x_{1jk} + \delta_{upper}$ and x_{2jk}^*.

3. Sort the m p-values for decreasing selectors IQR_j.

4. For each endpoint independently compare the jth ordered p-value with the unadjusted α. It is significant, if it is less than α and all previously tested null hypotheses are rejected as well.

5. Stop at the first non-significance and accept for all further endpoints the null hypothesis.

For one-sided testing the selectors are calculated from either $x_{1jk} + \delta_{lower}$ and x^*_{2jk} (test on decrease) or $x_{1jk} + \delta_{upper}$ and x^*_{2jk} (test on increase).
As for the parametric δ-shift method this procedure controls empirically the FWER in both the strong and the weak sense for two-sided testing. Detailed results of FWER-simulations start on the page 159. The graphical power simulation results start on page 109 in chapter 7. In the further work this procedure is abbreviated as 'np-δ-shift'.

4.3 Control of the FWER

In this section the main statements about the empirical control of the FWER by the procedures with a data-driven order of hypotheses presented in this and the further chapter are given. As noted above, more detailed information and results are listed in the appendix starting on page 155.
However before the behavior of the empirical power can be characterized, a decision rule is required, to decide whether an approach controls the FWER or not. For example as an empirical FWER of 5.23% is larger than the nominal level of 5%, it could be either concluded that the method does not control the FWER, or that the exceeding is a result of the simulation error. In the literature it is common to use the Wald interval as a decision rule. It is given by $\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/NSIM}$, where $\hat{\pi}$ denotes the empirical error rate and $NSIM$ is the number of simulation runs. This interval is advantageous because of its simplicity. However it is known to be liberal for $\hat{\pi}$ close to 0 or 1 combined with a small number of simulation runs $NSIM$. Hence instead of the Wald interval the Wilson interval is used throughout this thesis to decide, whether a procedure controls the FWER or not. The two-sided (1-α) Wilson score interval for π is denoted as:

$$\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2 \cdot NSIM} \pm z_{\alpha/2}\sqrt{\left[\hat{\pi}(1-\hat{\pi}) + \frac{z^2_{\alpha/2}}{4 \cdot NSIM}\right]/NSIM}\right)/(1 + z^2_{\alpha/2}/NSIM) \quad (4.18)$$

(Agresti and Coull (1998), Piegorsch (2004)). A detailed summary of confidence intervals for the proportion is proposed by Schaarschmidt (2005).
Here the lower one-sided confidence limit is used, to decide if the empirical FWER exceeds the nominal level. The limit itself is computed to an α-level of 5% and 10,000 simulation runs. For example, for $\alpha = 5\%$ the observed error rate has to be larger than 5.358%, such that the 5% are not included in the interval. In all further chapters an exceed of the nominal FWER is printed bold.
Detailed simulation studies on the weak and the strong control of the FWER are done for all new procedures. In all simulations 50 endpoints are tested. Further, for the procedures discussed in this chapter the nominal FWER is set to 5%. Tested are scenarios with varying relevance thresholds, correlation structures among endpoints, sample sizes and variances. For the parametric procedure the simulated data follow a normal distribution. In addition for the nonparametric procedure skewed distributed data is generated as well. Tested are both one- and two-sided hypotheses.
By use of the one-sided Wilson confidence limit, it can be said that all four procedures empirically control the FWER for two-sided testing. If it is tested one-sided and $\delta_{lower} = \delta_{upper} = 0$, then the δ-shift and the np-δ-shift procedure exceed slightly the FWER. Largest simulated empirical error rates are 5.89% for δ-shift and 6.24% for the np-δ-shift. In comparison to the δ-shift the largest empirical FWER of the random$^{\delta}$ approach in a one-sided testing scenario is 5.25%, which is not an exceed according to the Wilson confidence limit. However one-sided tests for high dimensional data sets are rarely used, as this requires either the a-priori knowledge of the direction of the treatment effect or a general interest in over- or under-expression of all genes.

4.4 Examples

In this section the application of the δ-shift, the random$^{\delta}$ and the np-δ-shift procedure is shown. As in the former chapter, the small data set including the morphological measurements of possums and the large microarray data sets are analyzed with two-sided tests and

an error rate of 0.05. However in contrast to the former analysis, relevance thresholds have to be selected.

4.4.1 Possum data set

For the possum data the relevance thresholds are set to $-\delta_{lower} = \delta_{upper} = 0.25$. As the absolute values of the differences are larger than 0.25, the data transformations of the procedures are not applied. Hence the δ-shift and the random$^{\delta}$ approaches lead to the same results.

Results of the δ-shift and the random$^{\delta}$ procedure:

endpoint	difference in means	selector statistic	test statistic	unadjusted p-value	adjusted p-value	Bonferroni adjusted p-value
totlngth	6.000	262.729	2.247	0.048	0.048	0.436
hdlngth	5.025	187.609	2.190	0.053	0.053	0.480
earconch	-7.463	163.597	-7.468	$2.14 \cdot 10^{-05}$	0.053	$1.92 \cdot 10^{-4}$
footlgth	-3.087	119.739	-1.478	0.170	0.170	1.000
skullw	3.950	114.357	2.165	0.056	0.170	0.500
belly	1.938	93.813	0.938	0.370	0.370	1.000
taill	2.250	41.729	1.853	0.094	0.370	0.842
chest	1.938	36.313	1.626	0.135	0.370	1.000
eye	1.075	9.123	1.576	0.146	0.370	1.000

The α-adjustment and the procedures with a data-driven order of hypotheses reject one hypothesis only. While the δ-shift and the random$^{\delta}$ method declare the endpoint 'totlngth' as significant, the Bonferroni correction results in a significant p-value for 'earconch'. The reason for the non-superior behavior of the procedures with a data-driven order of hypotheses is the abortion of the algorithm at the second endpoint ('hdlngth'), which has a larger selector as 'earconch' and a p-value of 0.053. If this p-value would have been significant, the procedures with a data-driven order of hypotheses would have lead to three significant endpoints.

In the following table results of the nonparametric relevance-shifted procedures are shown.

Results of the np-δ-shift procedure:

endpoint	difference in medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value	Bonferroni adjusted p-value
earconch	-7.350	7.100	10	0.004	0.004	0.036
totlngth	4.500	5.250	36	0.101	0.101	0.909
skullw	4.300	4.525	36	0.109	0.109	0.982
hdlngth	5.450	4.175	36	0.109	0.109	0.982
belly	1.75	3.938	29	0.683	0.683	1.000
footlgth	-3.950	3.925	15	0.073	0.683	0.655
chest	2.500	2.500	35	0.141	0.683	1.000
taill	2.000	1.813	35	0.137	0.683	1.000
eye	1.350	1.313	35	0.150	0.683	1.000

With the use of the relevance-shifted rank sum test all unadjusted p-values are increased compared to the ones resulting from the t-test. Only the endpoint 'earconch' has an unadjusted p-value less than the significance level of 0.05. As for the parametric analysis the Bonferroni adjustment declares this endpoint to be significant. And as 'earconch' has the largest selector statistic, the same result is given by the np-δ-shift method.

To give an example of the transformations, it is now assumed, that in site 3 for endpoint 'belly' data the following measurements are observed: 31.1, 33.1, 33.5 and 30.5. This results in a difference in treatment means of -0.013 and a difference in medians of -0.150.

Results of the data transformations:

procedure	difference in means / medians	selector statistic	test statistic	unadjusted p-value
δ-shift	-0.013	86.229	0.132	1.000
random$^{\delta}$	-0.013	105.756	0.132	1.000
np-δ-shift	-0.150	3.500	16	1.000

Initially the results do not change. The order of the hypotheses stay the same for the parametric procedure and in the nonparametric approach 'belly' and 'footlgth' switch their position. However the adjusted p-values of the endpoints with a smaller selector as 'belly' would change to 1.000.

The effect of the additional noise generated by the random$^{\delta}$ method can clearly be seen. While 'belly' achieves a selector of 86.229 if the δ-shift is used, the selector increases to 105.756 by use of the random$^{\delta}$ method. It has to be pointed out that the selector statistic of 'belly' depends on the random number generator and the initial seed. In this analysis the seed is set to 1010 and the final transformed data to compute the selector is 31.68397, 28.42102, 35.55216 and 32.59850.

4.4.2 TSHR mutation data set

As noted in the former chapter the range of mean expression among the genes is between 0 and 50,000. Here it is impossible to select a relevance threshold in terms of the difference, because to achieve a significant result the required treatment effect increases with an increasing mean expression. Hence individual relevance thresholds for thousands of genes would need to be specified. However an analysis is possible if the data is logarithmized. Besides the moderate variance stabilization the logarithmic transformation is advantageous, because by its application a multiplicative model changes into an additive one. That is here a test on ratio turns into a test on difference. By setting $-\delta_{lower} = \delta_{upper} = \log(1.5)$ and analyzing the logarithmized data it is tested for a relevant ratio of means of 1.5 based on the data on the original scale. Then without a multiplicity correction 65 significant endpoints can be found. In comparison with results of the former chapter ($-\delta_{lower} = \delta_{upper} = 0$), that is 1,176 fewer significant endpoints. As in the analysis in the former chapter no significant endpoints were found by use of the Bonferroni correction and the procedure with a data-driven order of hypotheses, none are found here either. The following two tables list the first ten endpoints with the highest selector statistics. As in this chapter a relevant ratio of means of 1.5 based on the original data is of interest, the second column values are transformed back on the original scale. Still this is equivalent to the difference in means of the logarithmized data as listed in the table in the former chapter. The first table gives the results of the δ-shift method.

Results of the δ-shift procedure:

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
11321	0.130	62.583	-1.466	0.166	0.166
6022	1.195	42.359	-0.230	1.000	1.000
8435	0.099	39.309	-2.415	0.031	1.000
12177	3.149	38.971	0.801	0.437	1.000
11876	5.865	35.058	1.671	0.119	1.000
4145	0.060	34.786	-4.011	0.001	1.000
3839	12.207	33.756	3.157	0.007	1.000
2148	3.869	33.711	1.125	0.281	1.000
7940	0.070	33.364	-3.668	0.003	1.000

The same ten endpoints can be found as in the analysis with the parametric procedure with a data-driven order of point-zero hypotheses applied on the logarithmized data. As expected all p-values enlarge, the selector statistics decrease and, no significant results can be found. Endpoint # 6022 is transformed by the δ-shift method, because the ratio of means is between 1/1.5 and 1.5. The selector statistic does not change marginally. However as for all other endpoints the selector decreases, # 6022 moves up one rank in the order of hypotheses.

Results of the random$^{\delta}$ procedure:

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
11321	0.130	62.583	-1.466	0.166	0.166
6022	1.195	40.049	-0.230	1.000	1.000
8435	0.099	39.309	-2.415	0.031	1.000
12177	3.149	38.971	0.801	0.437	1.000
1321	1.147	35.811	-0.332	1.000	1.000
11876	5.865	35.058	1.671	0.119	1.000
4145	0.060	34.786	-4.011	0.001	1.000
1681	0.908	34.672	0.407	1.000	1.000
3839	12.207	33.756	3.157	0.008	1.000
2148	3.869	33.711	1.125	0.281	1.000

By use of the random$^{\delta}$ procedure the overall result stays the same. However the list of the ten genes with the highest selectors is slightly different. Eight endpoints appear in both lists. The two genes, which are included in the list of the random$^{\delta}$ only, are transformed by both procedures. As the random$^{\delta}$ method adds a small noise term to the data, these two endpoints achieve a comparably large selector and appear in this list. Here it can be disadvantageous, as such endpoints may abort the procedure prematurely.
However the opposite effect is possible as well. With the data transformation the endpoint # 6022 achieved a smaller selector compared to results of the δ-shift procedure.

The application of the nonparametric tests is omitted, because with the logarithmic transformation the procedure with a data-driven order of hypotheses is equal to the nonparametric procedure to test for a relevant ratio. This approach will be discussed in chapter 6.

4.4.3 TNFα data set

In analogy to the analysis of the TSHR mutation data set the TNFα experiment is evaluated in terms of the selection of relevance thresholds for the difference in means based on

the logarithmic transformed data. These are set to $-\delta_{lower} = \delta_{upper} = \log(1.5)$ as well. And the nonparametric analysis is omitted, as it gives the same results as the nonparametric procedure with a data-driven order of hypotheses discussed in chapter 6.
Without the application of a multiple testing approach 502 significant genes can be found. By use of the Bonferroni correction 56 discriminatory endpoints are achieved.

Results of the δ-shift procedure: In comparison with the parametric procedures with data-driven order of point-zero hypotheses, instead of nine significant endpoints three are found by the δ-shift procedure here. As expected, the number of significant genes is less, because here relevance thresholds are included. However a second reason is the endpoint # 10284, which is the fourth endpoint in the order and has an unadjusted p-value of 0.073. Hence it stops the procedure, before the following four genes with small p-values are evaluated.

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
5979	5.789	6.310	27.140	$3.660 \cdot 10^{-09}$	$3.660 \cdot 10^{-09}$
13618	5.259	5.675	16.062	$2.265 \cdot 10^{-07}$	$2.265 \cdot 10^{-07}$
11600	5.202	5.453	36.420	$3.541 \cdot 10^{-10}$	$2.265 \cdot 10^{-07}$
10284	4.208	5.383	2.265	0.073	0.073
8563	5.013	5.268	19.322	$5.341 \cdot 10^{-08}$	0.073
13585	5.023	5.202	35.460	$4.379 \cdot 10^{-10}$	0.073
6629	4.973	5.163	25.392	$6.200 \cdot 10^{-09}$	0.073
13653	4.792	5.079	12.692	$1.397 \cdot 10^{-06}$	0.073
12453	0.506	4.505	-0.733	0.496	0.496
1689	2.951	4.469	1.868	0.111	0.496

Results of the random$^{\delta}$ procedure: A rather unexpected result is achieved by use of the random$^{\delta}$ procedure. Out of the ten endpoints with the highest selectors, only one has an unadjusted p-value less than 1. And its adjusted p-value is 1, because it is not on the top of the ordering. As the random$^{\delta}$ procedure includes additional noise to the transformed data of endpoints with a difference in means not exceeding a relevance threshold, the selectors

increase. In this example some of the transformed endpoints receive such a large selector, that they appear at the top of the ordering and stop the procedure prematurely.

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
10965	0.945	11.106	1.498	1.000	1.000
7162	0.938	9.376	2.385	1.000	1.000
13277	1.087	7.311	-0.356	1.000	1.000
2613	0.840	7.050	0.055	1.000	1.000
9027	0.949	6.465	1.608	1.000	1.000
5979	5.789	6.310	27.140	$3.660 \cdot 10^{-09}$	1.000
3792	0.969	6.173	1.876	1.000	1.000
1968	0.893	5.959	1.699	1.000	1.000
3357	0.842	5.895	0.062	1.000	1.000
13711	0.975	5.771	0.469	1.000	1.000

Chapter 5

Parametric testing procedures for relevant ratios

In the former chapter, procedures are presented which are used to detect relevant differences among the treatment groups. However, in high dimensional data with endpoints that have different scaled observations, it is hardly possible to define for each variable a relevant difference. Especially this is not possible for microarrays, because the individual genes for one experimental unit are expressed with differential intensity. Hence this would require the a-priori knowledge of the expression intensity of thousands of genes. Here a relevance criteria in terms of the ratio is more appropriate. In this chapter such parametric procedures are discussed. The null hypothesis to be tested is $H_{0j} : \theta_{lower} \leq \frac{\mu_{2j}}{\mu_{1j}} \leq \theta_{upper}$ against the alternative $H_{1j} : \frac{\mu_{2j}}{\mu_{1j}} < \theta_{lower}$ or $\frac{\mu_{2j}}{\mu_{1j}} > \theta_{upper}$, with the lower threshold $\theta_{lower} \leq 1$ and the upper one $\theta_{upper} \geq 1$. Throughout the entire thesis the thresholds are set to $\theta_{lower}^{-1} = \theta_{upper}$. As well as for the procedures in the former chapter it is not possible to use an appropriate two-sample test in combination with the selector statistic $w_j = \sum_{i=1}^{2} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_j)^2$ from the parametric procedure for point-zero hypotheses. Simulations of the FWER with a nominal level of 5% resulted in an empirical α around 40%.

A conceivable approach is a data transformation in terms of a reduction of a treatment group by the relevance threshold before computing the selector, as it was applied in the

former chapter. A simple multiplication or division of a treatment group by θ_{lower} or θ_{upper} is inappropriate because of the change of the variance. However as proposed by HAUSCHKE (1999) a hypothesis system to test for a relevant difference in means can be rewritten to a system for a relevant ratio of means and vice versa. For example the one-sided hypotheses to test on a relevant increase can be rewritten to:

$$H_{0,j}^{\delta} : \mu_{2j} - \mu_{1j} \leq \delta_{upper,j} = f\mu_{1j} \quad \Leftrightarrow \quad H_{0,j}^{\theta} : \frac{\mu_{2j}}{\mu_{1j}} \leq 1 + f = \theta_{upper}, \tag{5.1}$$

$$H_{1,j}^{\delta} : \mu_{2j} - \mu_{1j} > \delta_{upper,j} = f\mu_{1j} \quad \Leftrightarrow \quad H_{1,j}^{\theta} : \frac{\mu_{2j}}{\mu_{1j}} > 1 + f = \theta_{upper}. \tag{5.2}$$

By use of the estimates of μ_{ij} an approximate data transformation follows as

$$x_{2jk}^{*} = x_{2jk} - \delta_{upper,j} = x_{2jk} - f\bar{x}_{1j} = x_{2jk} - (\theta_{upper} - 1)\bar{x}_{1j}. \tag{5.3}$$

Such an approach was tested for various conditions as for one- and two-sided hypotheses, different sample sizes and relevance thresholds. The method achieved in the tested settings empirical error rates up to around 30%. The reason for the exceed is the use of the estimates instead of parameters for the data transformation.

A procedure which controls exactly the FWER in the weak and the strong sense can be derived from the theory of the stabilized tests. Such a procedure is used here. However, as mentioned in the introduction of chapter 4, the proposed procedure solves the problem of the control of the FWER, but lacks power. The reason for the lack of power is the same as discussed in chapter 4. But as will be presented in this chapter, two approximations of the procedure can be done by including a data transformation. The resulting approaches control empirically the FWER and are powerful compared to alternative multiple testing methods if certain data conditions hold.

5.1 Procedures

The Sasabuchi-test: In practice it is often difficult to define relevance thresholds for a difference of means. An alternative to the δ-shifted t-test is the parametric test on ratio according to SASABUCHI (1988). In the further work this test will be denoted as the

Sasabuchi-test.

In the univariate case for two-sided testing the test tests the null hypothesis $H_0 : \theta_{lower} \leq \frac{\mu_2}{\mu_1} \leq \theta_{upper}$ against the alternative $H_1 : \frac{\mu_2}{\mu_1} < \theta_{lower}$ or $\frac{\mu_2}{\mu_1} > \theta_{upper}$ The null hypothesis is rejected, if either the ratio of the means is significantly less than θ_{lower}:

$$t^{\theta}_{lower} = \frac{\bar{x}_2 - \theta_{lower}\bar{x}_1}{s_{pool}\sqrt{\frac{1}{n_2} + \frac{\theta^2_{lower}}{n_1}}} \tag{5.4}$$

or the ratio is significantly greater than θ_{upper} :

$$t^{\theta}_{upper} = \frac{\bar{x}_2 - \theta_{upper}\bar{x}_1}{s_{pool}\sqrt{\frac{1}{n_2} + \frac{\theta^2_{upper}}{n_1}}}. \tag{5.5}$$

The two-sided null hypothesis is rejected, if either $t^{\theta}_{lower} \geq t_{df=N-2,\alpha/2}$ or $t^{\theta}_{upper} \leq t_{df=N-2,1-\alpha/2}$.

For the one-sided test against a relevant decrease the null hypothesis and the test statistic reduce to $H_0 : \frac{\mu_2}{\mu_1} \geq \theta_{lower}$ and equation (5.4) with the critical value $t_{df=N-2,\alpha}$ and to test for a relevant increase $H_0 : \frac{\mu_2}{\mu_1} \leq \theta_{upper}$ and equation (5.5) with the critical value $t_{df=N-2,1-\alpha}$ are used. For testing multiple endpoints the equations are similar, only the index j for the endpoints is added. For example equation (5.4) changes to:

$$t^{\theta}_{lower,j} = \frac{\bar{x}_{2j} - \theta_{lower}\bar{x}_{1j}}{s_{pool,j}\sqrt{\frac{1}{n_2} + \frac{\theta^2_{lower}}{n_1}}}. \tag{5.6}$$

In the following multiple testing procedures the two-sided p-value of the Sasabuchi-test for the jth endpoint is denoted by p^{θ}_j.

Compared to the other two-sample tests used in this work the Sasabuchi-test has an exceptional behavior of the power. This is briefly discussed here, as it influences the simulation settings. Due to the asymmetry of the ratio, the power is dependent on the direction of the test. The dependency can be illustrated by use of the equation for the sample size calculation, which is proposed by Kieser and Hauschke (1999):

$$n \geq (1 + \theta^2_{side}) \cdot (t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2})^2 \cdot \left(\frac{CV_1}{\tau_\theta - \theta_{side}}\right)^2 \tag{5.7}$$

with τ_θ denoting the true ratio of means, $n_1 = n_2 = n$ and CV_1 is the coefficient of variation of the first treatment group. To test one-sided on increase against $\theta_{upper} = 2$, a sample

size of 15 per group is needed to prove a ratio of means of 2.15 as significant ($CV_1 = 0.1$, $\alpha = 5\%, \beta = 20\%$ and $\mu_1 = 100, \mu_2 = 215$, $\sigma = 10$.) Compared to this, for testing on decrease with $\theta_{lower} = 0.5$ and $\mu_2/\mu_1 = 2.15^{-1}$ with $\mu_1 = 100, \mu_2 = 46.5$ and all other parameters equal to the test on increase a sample size of 65 is required. The effect is based in the last term of the sample size equation: $\left(\frac{CV_1}{\tau^\theta - \theta_{side}}\right)^2$. Given that for testing on decrease the ratio and the relevance threshold are the inverse of the ones from testing on increase, the denominator of the fraction is much smaller for testing on decrease as for testing on increase because of the asymmetry of the ratio.

To achieve for the simulations a behavior of the power independent from the direction of the test, for both test on increase and decrease the same settings of expected values are used, but for testing on decreased the index of the group is switched. When testing on increase the non-centrality parameter of the Sasabuchi-test is

$$\nu_{upper} = \frac{\mu_2 - \mu_1 \cdot \theta_{upper}}{\sigma\sqrt{\frac{1}{n_2} + \frac{\theta_{upper}^2}{n_1}}} \tag{5.8}$$

(Hauschke (1999)). By dividing by μ_1 the equation changes to:

$$\nu_{upper} = \frac{\frac{\mu_2}{\mu_1} - \theta_{upper}}{\frac{\sigma}{\mu_1}\sqrt{\frac{1}{n_2} + \frac{\theta_{upper}^2}{n_1}}}. \tag{5.9}$$

And the Sasabuchi-test on increase can be written as:

$$t^\theta_{upper} = \frac{\bar{x}_2 - \theta_{upper}\bar{x}_1}{s_{pool}\sqrt{\frac{1}{n_2} + \frac{\theta_{upper}^2}{n_1}}} = -\frac{\bar{x}_1 - \theta_{upper}^{-1}\bar{x}_2}{s_{pool}\sqrt{\frac{\theta_{upper}^{-2}}{n_2} + \frac{1}{n_1}}}. \tag{5.10}$$

Hence the modified non-centrality parameter is:

$$\nu_{upper^{-1}} = -\frac{\mu_1 - \mu_2\theta_{upper}^{-1}}{\sigma\sqrt{\frac{\theta_{upper}^{-2}}{n_2} + \frac{1}{n_1}}} \overset{:\mu_2}{\Rightarrow} -\frac{\frac{\mu_1}{\mu_2} - \theta_{upper}^{-1}}{\frac{\sigma}{\mu_2}\sqrt{\frac{\theta_{upper}^{-2}}{n_2} + \frac{1}{n_1}}}. \tag{5.11}$$

Note that the coefficient of variation changed to $\frac{\sigma}{\mu_2}$, which is now dependent of group 2. Given that $\theta_{lower} = \theta_{upper}^{-1}$ it follows that the same power behavior (or sample size requirements) for both testing on increase and decrease are achieved in simulations, if the expected values of the groups are switched. The effect can be visualized with an estimation

of the sample size. In the equation the coefficient of variation is now taken from group 2. In the example above for testing on increase, a sample size of 15 is required. To test on decrease the same parameters are used, but $\mu_1 = 215, \mu_2 = 100$ and $\theta_{lower} = 0.5$. Then the coefficient of variation is $CV_2 = 0.0465$ and the required sample size is again 15.

5.1.1 The Sasabuchi selector procedure

In this section the derivation of a selector statistic for an exact parametric procedure with a data-driven order of hypotheses to test for a relevant ratio and a first approximation of it are shown. To construct the exact approach the factor-shifted null-hypothesis $H_0 : \mu_{\mathbf{2}} = \theta\mu_{\mathbf{1}}$ is considered, where the a-priori chosen shift parameter $\theta > 0$ denotes the relevance threshold. In the special case of $\theta = 1$ the hypothesis reduces to the classical unshifted test problem. For the construction of the selector the matrices $\mathbf{H}$ and $\mathbf{G}$ are needed. While the error matrix $\mathbf{G}$ remains unchanged with respect to (3.4)

$$\mathbf{G} = \sum_{i=1}^{2}\sum_{k=1}^{n_i}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)', \tag{5.12}$$

the matrix $\mathbf{H}$ representing the deviation from the null hypothesis changes to

$$\mathbf{H} = \frac{n_1 n_2}{n_1 + \theta^2 n_2}(\bar{\mathbf{x}}_2 - \theta\bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \theta\bar{\mathbf{x}}_1)'. \tag{5.13}$$

To derive the weight vectors and the scores the matrix of the total sums of squares and cross products $\mathbf{W}$ is needed. It is given by

$$\mathbf{W} = \frac{n_1 n_2}{n_1 + \theta^2 n_2}(\bar{\mathbf{x}}_2 - \theta\bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \theta\bar{\mathbf{x}}_1)' + \sum_{i=1}^{2}\sum_{k=1}^{n_i}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_i)'. \tag{5.14}$$

In analogy to the multiple testing procedure for point-zero hypotheses (Kropf and Läuter (2002)) the diagonal elements of $\mathbf{W}$ are used as selector statistic:

$$\begin{aligned} w_j &= \frac{n_1 n_2}{n_1 + \theta^2 n_2}(\bar{x}_{2j} - \theta\bar{x}_{1j})^2 + \sum_{i=1}^{2}\sum_{k=1}^{n_i}(x_{ijk} - \bar{x}_{ij})^2 && (5.15)\\ &= \frac{n_1 n_2}{n_1 + \theta^2 n_2}(\bar{x}_{2j} - \theta\bar{x}_{1j})^2 + \left(\sum_{i=1}^{2}\sum_{k=1}^{n_i} x_{ijk}^2\right) - n_1\bar{x}_{1j}^2 - n_2\bar{x}_{2j}^2 && (5.16)\\ &= \sum_{i=1}^{2}\sum_{k=1}^{n_i} x_{ijk}^2 - \frac{1}{n_1 + \theta^2 n_2}\cdot(n_1\bar{x}_{1j} + n_2\theta\bar{x}_{2j})^2. && (5.17) \end{aligned}$$

By use of scores instead of the variables the univariate terms corresponding to **H** and **G** result in

$$h_z = \frac{n_1 n_2}{n_1 + \theta^2 n_2}(\bar{z}_2 - \theta \bar{z}_1)^2 \tag{5.18}$$

and

$$g_z = (N-2)s_z^2. \tag{5.19}$$

Then the Sasabuchi test for the scores is given by $\sqrt{\frac{h_z}{g_z}}$. As in the procedure with a data-driven order of hypotheses the Sasabuchi test is used on the two samples per gene, the single variables are used as scores.

With the use of equation (5.17) combined with the Sasabuchi-test this procedure with a data-driven order of hypotheses is an exact α test. However as described in chapter 4 the equation allows the inclusion of only one relevance threshold - θ instead of θ_{lower} and θ_{upper}. Hence in this section a modified version of the selector is used. All endpoints indicating an over-expression, that is here a ratio of means greater or equal than 1, use θ_{upper} instead of θ. And for the remaining endpoints, the ones which indicate an under-expression, have θ_{lower} instead of θ.

The resulting algorithm is:

1. Select relevance thresholds $\theta_{lower} \leq 1, \theta_{upper} \geq 1$.
2. For each endpoint j:
   (a) Compute the two-sided p-value p_j^θ of the Sasabuchi-test by the use of equations 5.4 and 5.5.
   (b) Calculate the selector statistic

$$w_j = \begin{cases} \sum_{i=1}^{2}\sum_{k=1}^{n_i} x_{ijk}^2 - \frac{1}{n_1+\theta_{lower}^2 n_2} \cdot (n_1\bar{x}_{1j} + n_2\theta_{lower}\bar{x}_{2j})^2; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} < 1 \\ \sum_{i=1}^{2}\sum_{k=1}^{n_i} x_{ijk}^2 - \frac{1}{n_1+\theta_{upper}^2 n_2} \cdot (n_1\bar{x}_{1j} + n_2\theta_{upper}\bar{x}_{2j})^2; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} \geq 1 \end{cases} \tag{5.20}$$

3. Sort the m p-values for decreasing selectors w_j.
4. Compare the jth ordered p-value with the unadjusted α. It is significant, if $p_j^\theta <$ unadjusted α.

5. Stop at the first non-significance and accept for all further endpoints the null hypothesis.

For one-sided testing the one-sided p-values are used and the selector statistic stays the same. If tested on increase, for the selector statistic the lower relevance threshold is given by $\theta_{lower} = \theta_{upper}^{-1}$ and vice versa for one-sided testing on decrease.

However it lacks power for the same reason as the shift-selector method proposed in section 4.1.1: the selector statistic reaches a local maximum when the treatment effect is 0 or the ratio of means is 1, respectively.
An additional problem may arise when the mean levels of the endpoints differ. As could be seen from the procedures with a data-driven order of point-zero hypotheses the power depends on the degree of the variance homogeneity among the endpoints. This is true as well for the relevance-shifted ones, as it will be shown in chapter 7. For the analysis of microarrays this is a problem, as usually the individual genes per treatment group vary in the intensity of the expression and thus in the variance. However even if the variance is set to a constant value, while the mean level of the endpoints differ, the Sasabuchi-selector procedure shows a lack of power. The following graphic depicts this problem. Here the selector is plotted against the true ratio of means. The setting is equivalent to the corresponding graphic 4.2 for the shift-selector procedure. Only the expected values and the relevance thresholds change to $\mu_{1j} = (1/\tau_\theta^{H_1}) \cdot 10$ and $\mu_{2j} = 10$ for $\tau_\theta^{H_1} < 1$ and $\mu_{1j} = 10$ and $\mu_{2j} = \tau_\theta^{H_1} \cdot 10$ for $\tau_\theta^{H_1} \geq 1$, with $\tau_\theta^{H_1}$ taking values from 0.25 to 4 in steps of 0.1 units. The random ratio $\tau_\theta^{H_1}$ is constructed from two combined sets of values. One set includes all values between 1 and θ_{upper} in steps of 0.05. To receive an equal amount of ratios compared to the first set, for the second one all values between 1 and θ_{lower}^{-1} in steps of 0.05 are computed and the second set is the inverse of these values.
In this graphic two additional curves are included. For each of the three curves the underlaying data levels are different. The first one has a level of 50 ($\mu_{1j} = (1/\tau_\theta^{H_1}) \cdot 50$ and $\mu_{2j} = 50$ for $\tau_\theta^{H_1} < 1$, $\mu_{1j} = 50$ and $\mu_{2j} = \tau_\theta^{H_1} \cdot 50$ for $\tau_\theta^{H_1} \geq 1$), the second has a level of

75 and the third one 100. Again the scenario is a two-sided one and the thresholds are set to 0.5 and 2.

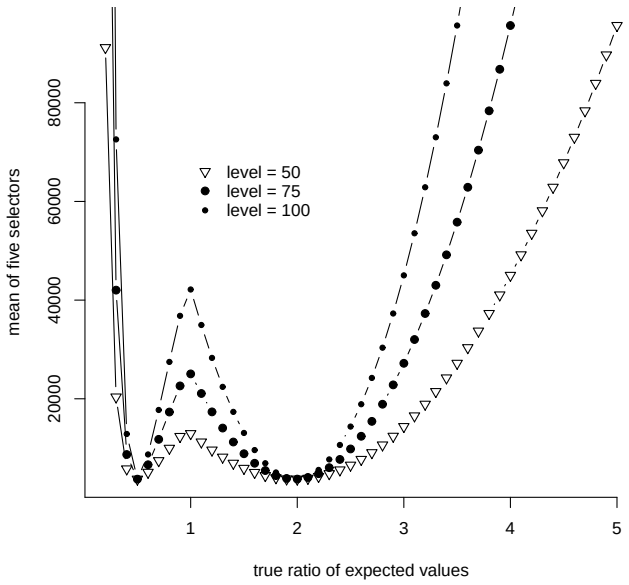

Figure 5.1: Sasabuchi selector: comparison of selectors for different data levels

If the ratio equals the thresholds, all lines show the same selector statistic, because none has a difference in means and all have the same standard deviation. For all other values of $\tau_\theta^{H_1}$ the selector depends extremely on the level of the data. The reason for this behavior is, that while the variance per endpoint and the true ratio of means stay constant, the difference in means increases with increasing data levels. The effect can be best explained by use of the test statistic of the Sasabuchi-test. For each of the simulated experiments the denominator of the test is the same. But the numerator changes; for example if the true ratio of means is set to 3, then the true difference in means for the data levels are $\mu_{2j} - \theta_{upper}\mu_{1j}$, these are $150 - 2 \cdot 50 = 50$, $225 - 2 \cdot 75 = 75$ and $300 - 2 \cdot 200 = 100$. Hence the test statistic increases with the increasing data levels. As the selector statistic reflects the test statistic, the same behavior can be seen. The effect is correct for endpoints, which exceed one of the thresholds. However the selector of endpoints under H_0 with a large

mean level can abort the testing procedure too early.
In the further work this approach is abbreviated as 'Sasabuchi-selector'.

5.1.2 The θ-shift procedure

As for the procedures with a data-driven order of hypotheses for relevant differences the low proportional power can be increased with an inclusion of a data shift for endpoints under H_0 with $\theta_{lower} < \bar{x}_{2j}/\bar{x}_{1j} < \theta_{upper}$. Furthermore the data shift solves the lack of power which occurs with different levels: With the data transformation the treatment effect of the shifted endpoints is dismissed. Thus the selector reflects the pooled variance of the two samples only.
The computation of this procedure with a data-driven order of hypotheses for two-sided testing against a relevant ratio is the same as in section 5.1.1, only 2. (b) changes to:

2. (b) For each endpoint with $\bar{x}_{2j}/\bar{x}_{1j} \geq 1$ transform the data independently to

$$x^*_{2jk} = \begin{cases} \theta_{upper}\bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}; & \forall\, j \mid 1 \leq \frac{\bar{x}_{2j}}{\bar{x}_{1j}} < \theta_{upper} \\ x_{2jk}; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} \geq \theta_{upper}. \end{cases} \tag{5.21}$$

For endpoints with $\bar{x}_{2j}/\bar{x}_{1j} < 1$ exchange the group indices, such that $x'_{1jk} = x_{2jk}$ and $x'_{2jk} = x_{1jk}$ and transform the switched second group to

$$x^*_{2jk} = \begin{cases} \theta^{-1}_{lower}\bar{x}'_{1j} - \bar{x}'_{2j} + x'_{2jk}; & \forall\, j \mid 1 > \frac{\bar{x}_{2j}}{\bar{x}_{1j}} > \theta_{lower} \\ x'_{2jk}; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} \leq \theta_{lower}. \end{cases} \tag{5.22}$$

Calculate the selector statistic

$$w_j = \begin{cases} \sum_{k=1}^{n'_1} x'^{\,2}_{1jk} + \sum_{k=1}^{n'_2} x^{*\,2}_{2jk} - \frac{1}{n'_1+\theta^{-2}_{lower}n'_2} \cdot (n'_1\bar{x}'_{1j} + n'_2\theta^{-1}_{lower}\bar{x}^*_{2j})^2; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} < 1, \\ \sum_{k=1}^{n_1} x^2_{1jk} + \sum_{k=1}^{n_2} x^{*\,2}_{2jk} - \frac{1}{n_1+\theta^2_{upper}n_2} \cdot (n_1\bar{x}_{1j} + n_2\theta_{upper}\bar{x}^*_{2j})^2; & \forall\, j \mid \frac{\bar{x}_{2j}}{\bar{x}_{1j}} \geq 1. \end{cases} \tag{5.23}$$

which is the selector statistic used in the former procedure, but including the transformed second treatment group and the exchange of the group index i.

When testing one-sided against a relevant increase for endpoints with $\bar{x}_{2j}/\bar{x}_{1j} < \theta_{upper}$ then x_{2jk} is transformed to $x^*_{2jk} = \theta_{upper}\bar{x}_{1j} - \bar{x}_{2j} + x_{2jk}$ and for j with $\bar{x}_{2j}/\bar{x}_{1j} \geq \theta_{upper}$,

$x^*_{2jk} = x_{2jk}$. The selector is computed as $w_j = \sum_{k=1}^{n_1} x_{1jk}^2 + \sum_{k=1}^{n_2} {x^*_{2jk}}^2 - \frac{1}{n_1+\theta^2_{upper}n_2} \cdot (n_1\bar{x}_{1j} + n_2\theta_{upper}\bar{x}^*_{2j})^2$.

Equivalently for one-sided testing against a relevant decrease for endpoints with $\bar{x}_{2j}/\bar{x}_{1j} > \theta_{lower}$, x'_{2jk} transforms to $x^*_{2jk} = \theta^{-1}_{lower}\bar{x}'_{1j} - \bar{x}'_{2j} + x'_{2jk}$ and for each j with $\bar{x}_{2j}/\bar{x}_{1j} \leq \theta_{lower}$ $x^*_{2jk} = x_{2jk}$. The computation of the selector is done by $w_j = \sum_{k=1}^{n'_1} {x'_{1jk}}^2 + \sum_{k=1}^{n'_2} {x^*_{2jk}}^2 - \frac{1}{n'_1+\theta^{-2}_{lower}n'_2} \cdot (n'_1\bar{x}'_{1j} + n'_2\theta^{-1}_{lower}\bar{x}^*_{2j})^2$.

With this data transformation the procedure solves both the problem occurring with different data levels and the problem of the selector statistic having a maximum at the ratio of means equal to 1. The next graphic shows the selector statistic for different ratio of means and data levels. It has the same setting as figure 5.1.

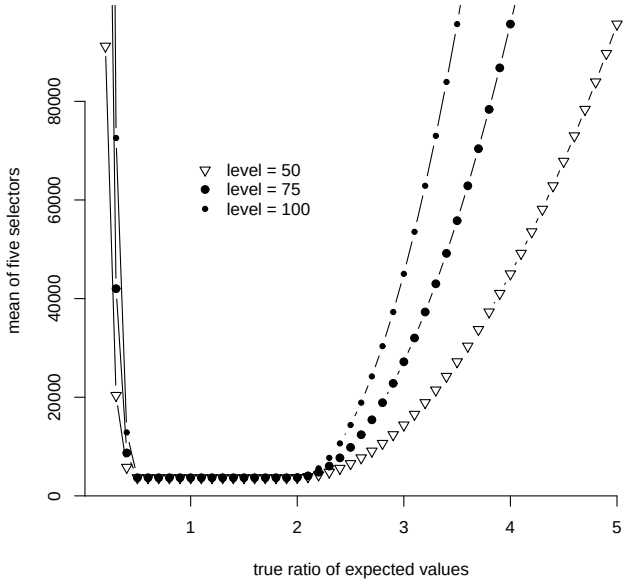

Figure 5.2: θ-shift procedure: comparison of selectors for different data levels

As for all scenarios with different data levels and ratio of means the variances are equal, the selectors do not vary between the relevance thresholds. If the ratio of means exceeds one of the thresholds, the selectors increase. Here one can see differences in the increase: with higher data levels the increase is stronger compared to lower data levels.

In the following graphics and tables this procedure will be abbreviated as the θ-shift procedure.

5.1.3 The random$^{\theta}$ procedure

The final procedure presented in this chapter is a procedure with a data-driven order of hypotheses for relevant ratios in analogy to the random$^{\delta}$ method for differences. For the test on relevant differences the data transformation for two-sided testing and endpoints with $0 \leq \bar{x}_{2j} - \bar{x}_{1j} < \delta_{upper}$ is

$$x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + \delta_{upper}. \tag{5.24}$$

To construct a transformation for the test on relevant ratios, the transformation of relevance threshold from difference to ratio according to Hauschke (1999) is applied. Following Hauschke (1999) δ_{upper} can be exchanged with $(\theta_{upper} - 1)\mu_{1j}$. By replacing the unknown μ_{1j} with it's estimate $\bar{x}_{1j}$, the data transformation for the two-sided testing and endpoints with $1 \leq \frac{\bar{x}_{2j}}{\bar{x}_{1j}} < \theta_{upper}$ changes to:

$$x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + (\theta_{upper} - 1) \cdot \bar{x}_{1j}. \tag{5.25}$$

Again, the procedure is based on the procedure of section 5.1.1 with 2. (b) changing to to:

2. (b) For each endpoint j with $1 < \frac{\bar{x}_{2j}}{\bar{x}_{1j}} < \theta_{upper}$ transform the data of the second group x_{2jk} to:

$$x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + (\theta_{upper} - 1) \cdot \bar{x}_{1j}, \tag{5.26}$$

where x^z_{2jk} denotes random numbers taken from the normal distribution $N(\mu = \bar{x}_{1g}, \sigma_{ij} = 1)$. For endpoints with $\bar{x}_{2j}/\bar{x}_{1j} < 1$ exchange the group indices, such that $x'_{1jk} = x_{2jk}$ and $x'_{2jk} = x_{1jk}$. Endpoints with $1 > \bar{x}_{2j}/\bar{x}_{1j} > \theta_{lower}$ have x'_{2jk} transformed to

$$x^*_{2jk} = \frac{x^z_{2jk} - \bar{x}^z_{2j}}{s^z_{2j}} \cdot s_{pool,j} + \bar{x}^z_{2j} + (\theta^{-1}_{lower} - 1) \cdot \bar{x}'_{1j}, \tag{5.27}$$

with Gaussian distributed random numbers $x^z_{2jk} \sim N(\mu = \bar{x}'_{1g}, \sigma_{ij} = 1)$. Endpoints with $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} \leq \theta_{lower}$ or $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} \geq \theta_{upper}$ have $x^*_{2jk} = x_{2jk}$.
Calculate the selector statistic (5.23), which is the selector statistic from the former section.

The changes for the one-sided procedures are as follows: to test against a relevant decrease, transform for each endpoint with $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} < \theta_{upper}$ independently x_{2jk} according to equation (5.26), with random numbers $x^z_{2jk} \sim N(\mu = \bar{x}_{1g}, \sigma_{ij} = 1)$. Endpoints with $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} \geq \theta_{upper}$ have $x^*_{2jk} = x_{2jk}$. Compute the selector statistic $w_j = \sum_{k=1}^{n_1} x^2_{1jk} + \sum_{k=1}^{n_2} {x^*_{2jk}}^2 - \frac{1}{n_1 + \theta^2_{upper} n_2} \cdot (n_1 \bar{x}_{1j} + n_2 \theta_{upper} \bar{x}^*_{2j})^2$.
For one-sided testing against a relevant decrease, transform x'_{2jk} of endpoints with $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} > \theta_{lower}$ as proposed in equation (5.27), with random numbers $x^z_{2jk} \sim N(\mu = \bar{x}'_{1g}, \sigma_{ij} = 1)$. Endpoints with $\frac{\bar{x}_{2j}}{\bar{x}_{1j}} \leq \theta_{lower}$ have $x^*_{2jk} = x'_{2jk}$. Calculate the selector by use of $w_j = \sum_{k=1}^{n'_1} {x'_{1jk}}^2 + \sum_{k=1}^{n'_2} {x^*_{2jk}}^2 - \frac{1}{n'_1 + \theta^{-2}_{lower} n'_2} \cdot (n'_1 \bar{x}'_{1j} + n'_2 \theta^{-1}_{lower} \bar{x}^*_{2j})^2$.

In the graphics at the end of this chapter, this procedure will be abbreviated as 'random$^\theta$'.

5.2 Control of the FWER

For all three procedures the weak and the strong control of the FWER is tested for varying sample sizes, variances, correlation structures among the endpoints and relevance thresholds. Both one- and two-sided hypotheses are concerned. Furthermore different α-levels are chosen. In general 50 endpoints are observed; in some scenarios the number of endpoints under H_0 varies. For most of the studies the mean level of the data is set to 100 and in some cases the individual levels are exponentially distributed (see section A.3 in the appendix). Unbalanced designs are observed as well. Finally for some scenarios the correlations vary among the endpoints; in general the correlation is constant.
The Sasabuchi selector keeps empirically the FWER in the weak as well as in the strong sense. In none of the scenarios is an exceeding observed. As with the δ-shift method the

θ-shift procedure controls the FWER empirically, if it is tested two-sided. For one-sided testing and in the case of $\theta_{lower} = \theta_{upper} = 1$ slight exceeds of the FWER occur; the largest empirical error rates for nominal levels of 1%, 5% and 10% are 1.21%, 5.92% and 11.67%. Compared to this the random$^{\delta}$ procedure shows only exceeds, if the levels of the endpoints are exponentially distributed and if tested one-sided. Here the highest empirical error rate of 5.49% is observed. Detailed simulation results of the three procedures are presented in the appendix starting on page 165.

5.3 Examples

In this chapter the possum data set and afterwards the TSHR mutation data set are analyzed only. An evaluation of the TNFα data set is omitted, as the data is already logarithmized and a test on ratio for such data is beyond the scope of this work. For both examples two-sided tests and $\alpha = 0.05$ are used.

5.3.1 Possum data set

The use of relevance-shifted tests on difference is problematic for the possum data set, because the nine endpoints vary in the scale of the measurements. Largest values are observed for the length of the head, it has measurements around 90. In contrast the distance from medial canthus to lateral canthus of right eye takes values around 15. Hence the selection of relevance thresholds based on the difference is hardly possible for this data. The appropriate analysis is done by the use of statistics which test a treatment effect against relative relevance criteria.

In the following tables the possum data set is analyzed by the θ-shift and the random$^{\theta}$ procedures. Additionally results of the Bonferroni correction are shown. The relevance thresholds are set to $\theta_{lower}^{-1} = \theta_{upper} = 1.02$. The following table gives the results for the θ-shift, the random$^{\delta}$ and the Bonferroni methods:

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value	Bonferroni adjusted p-value
totlngth	1.073	224.524	1.692	0.122	0.122	1.000
hdlngth	1.056	154.038	1.465	0.174	0.174	1.000
earconch	0.858	136.759	-6.706	$5.322 \cdot 10^{-5}$	0.174	$4.790 \cdot 10^{-4}$
footlgth	0.956	106.117	-0.894	0.392	0.392	1.000
skullw	1.071	99.043	1.650	0.130	0.392	1.000
belly	1.060	90.640	0.716	0.490	0.490	1.000
taill	1.065	37.404	1.429	0.184	0.490	1.000
chest	1.075	34.034	1.360	0.204	0.490	1.000
eye	1.072	8.900	1.476	0.171	0.490	1.000

In this analysis the Bonferroni adjustment is superior to the θ-shift and the random$^{\delta}$ approach, as it gives one significant endpoint ('earconch'), while the procedures with a data-driven order of hypotheses accept all null hypotheses. Compared to the tests on relevant differences the endpoint 'totlngth' is not significant. The reason for this non-significance in this analysis is the scale of the measurement. Observations of 'totlngth' are around 85 and the relevant difference is set to 0.25. This relevance criteria is marginal in comparison with the scale of the measurements. Hence for the tests on difference this endpoint is tested against a neglecting relevance criteria and it is therefore significant. In contrast to this, the use of percental thresholds is clearly more appropriate for the analysis of this data set.

As well as for the relevance-shifted tests on difference results for a modified endpoint 'belly' are computed.

Results of the data transformations:

procedure	ratio of means	selector statistic	test statistic	unadjusted p-value
θ-shift	1.000	86.229	-0.361	1.000
random$^{\theta}$	1.000	105.756	-0.361	1.000

Basically the overall results do not change. With the random$^{\theta}$ procedure the positions of 'belly' and 'skullw' exchange. Furthermore all adjusted p-values in the order after 'belly'

change to 1.

The observations to compute the selector statistic generated by random$^{\theta}$ are 32.57522, 29.31227, 36.44341 and 33.48975.

5.3.2 TSHR mutation data set

For sake of completeness in this section the microarray data is analyzed with the parametric procedures, which test for a relevant ratio. However for practical use this is not recommended for the analysis of microarrays.

By setting the relevance thresholds to $\theta_{lower}^{-1} = \theta_{upper} = 1.5$, 37 significant endpoints can be found if the Sasabuchi-tests are applied without a multiplicity correction. For comparison in chapter 3 1,266 significant genes were found. With the Bonferroni adjustment and the procedures with a data-driven order of hypotheses no significant endpoints can be found. The following two tables list the results of the ten endpoints with the highest selector statistics.

Results of the θ-shift procedure:

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
1702	2.165	5196498397	0.747	0.468	0.468
2071	1.429	4838853130	-0.190	1.000	1.000
12575	0.796	4719516117	0.637	1.000	1.000
12569	0.791	4598244393	0.395	1.000	1.000
1539	2.273	4054920369	0.767	0.457	1.000
7940	0.102	3527908282	-2.512	0.026	1.000
3304	0.106	3414275223	-1.842	0.088	1.000
3305	0.114	3397795663	-1.713	0.110	1.000
12573	0.800	3308209852	0.975	1.000	1.000
1637	1.141	2971993013	-1.459	1.000	1.000

Results of the random$^{\theta}$ procedure:

endpoint	ratio of means	selector statistic	test statistic	unadjusted p-value	adjusted p-value
2071	1.429	6626295535	-0.190	1.000	1.000
1702	2.165	5196498397	0.747	0.468	1.000
12575	0.796	4395406702	0.637	1.000	1.000
1539	2.273	4054920369	0.767	0.457	1.000
12569	0.791	3866838410	0.395	1.000	1.000
7940	0.102	3527908282	-2.512	0.026	1.000
3304	0.106	3414275223	-1.842	0.088	1.000
3305	0.114	3397795663	-1.713	0.110	1.000
1637	1.141	3363643455	-1.459	1.000	1.000
8273	0.137	2939211692	-2.244	0.043	1.000

Both procedures do not show any significant result. But the lists of the approaches are consistent: Nine of ten endpoints appear in both lists. Here the θ-shift method is preferable, as with the random$^{\theta}$ procedure the not discriminating endpoint # 2071 gains a high selector.

Chapter 6

Nonparametric testing procedures for relevant ratios

In this chapter two nonparametric multiple testing procedures for multivariate data are shown, which use tests on ratio for relevance-shifted hypotheses. First a data-driven order of relevance-shifted hypotheses is presented in analogy to the δ-shift and the θ-shift methods. As this procedure is the main aim of this work, in chapter 7 its behavior of the power will be compared with the powerful permutation algorithm for step-down minP adjusted p-values proposed by Westfall and Young (1993), because the application of the permutation algorithm is common in the analysis of microarray data. However the permutation method tests point-zero hypotheses. Hence a modification in terms of testing relevance-shifted hypotheses has to be created. This modified permutation algorithm is the second procedure discussed in this chapter.

Although the nonparametric procedure with a data driven-order of hypotheses proposed in this chapter is similar to the former nonparametric ones, it has one main difference. The assumptions about the data differ from the former methods. Up to now the nonparametric procedures assumed independent and continuous sample vectors with equal distribution functions except an additive treatment effect $\mathbf{\Delta}$. In this chapter for both methods it is assumed as well, that the independent and continuous sample vectors $\mathbf{x}_{1k} > 0$ and $\mathbf{x}_{2k} > 0$

have the distribution functions $F_m(\mathbf{x})$ and $G_m(\mathbf{x})$. But both distribution functions are assumed to be equal except for the treatment effect κ, such that $G_m(\mathbf{x}) = F_m(\mathbf{x}/\kappa)$ where '/' indicates a component-wise division of vectors and $\kappa = (\kappa_1, \ldots, \kappa_m)'$ denotes a scaling factor. This means, the treatment effect is not only a shift in the location; the scale changes as well. Hence by setting $\theta_{lower} = \theta_{upper} = 1$ the procedure with a data-driven order of hypotheses presented in this chapter does not reduce to the nonparametric analog for point-zero hypotheses (section 3.1.2), unless prior to the computation a logarithmic data transformation is applied. In particular, the unadjusted p-values are the same, however the order of the selector statistic may change.

However the hypotheses system used in this chapter is similar to the one used for the parametric tests on ratio. In this chapter the parameter of interest is κ_j, which denotes the true ratio of medians of the jth endpoint. And it will be tested, whether the two-sided null hypothesis $H_{0,j} : \theta_{lower} \leq \kappa_j \leq \theta_{upper}$ can be rejected in favor of the alternative $H_{1,j} : \kappa_j < \theta_{lower}$ or $\kappa_j > \theta_{upper}$.

The chapter begins with the introduction of the nonparametric test on ratio. For sake of simplicity it is presented for the univariate case. Afterwards the two new procedures are shown. Then some results of the simulation study on the control of the FWER are given for both methods. Finally the three example data sets are analyzed.

6.1 Procedures

The nonparametric test for a relevant ratio As introduced in section 4.2 the Wilcoxon rank sum procedure tests, whether the treatment effect Δ is different from 0: $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$. To construct a nonparametric test for a relevant ratio, the confidence interval for Δ based on the logarithmized samples is computed. The confidence limits are transformed back to the original scale and become limits for the treatment effect κ in terms of the ratio of medians. With these limits a test for a relevant ratio is constructed, which rejects in the univariate case the two-sided null hypothesis $H_0 : \theta_{lower} \leq \kappa \leq \theta_{upper}$ in favor of the alternative $H_1 : \kappa < \theta_{lower}$ or $\kappa > \theta_{upper}$.

HODGES AND LEHMANN (1963) propose an estimator of the difference in medians Δ, which is denoted here as $\hat{\Delta}$. To estimate the location shift all $n_1 n_2$ differences between the two samples are computed with $U_{kk'} = x_{2k} - x_{1k'}$. Afterwards the differences are sorted in an increasing order, such that $U_{(1)} \leq \ldots \leq U_{(n_1 n_2)}$. If $n_1 n_2$ is odd, then calculate $b = (n_1 n_2 - 1)/2$ and Δ is estimated with $\hat{\Delta} = U_{(b+1)}$. And if $n_1 n_2$ is even, then $b = n_1 n_2/2$ and $\hat{\Delta} = (U_{(b)} + U_{(b+1)})/2$.

To construct an exact $(1-\alpha)$ confidence interval for Δ, the Wilcoxon Mann Whitney test statistic of x_{1k} on the transformed observations $z_{2k}(d) = x_{2k} - d$ is computed for all possible values of d, $d \in \mathbb{R}$:

$$W^{WMW}(d) = \sum_{k=1}^{n_2} r_{2k}(d) - \frac{n_2(n_2+1)}{2}, \tag{6.1}$$

where $r_{2k}(d)$ is the rank of the transformed $z_{2k}(d)$ among both samples. The ranks $r_{2k}(d)$ change only in the points $d \in \Omega = \{U_{kk'} = x_{2k} - x_{1k'}\}$ and with increasing d the rank sum and with it $W^{WMW}(d)$ decrease. Thus the statistic $W^{WMW}(d)$ is a monotone decreasing step-function, which changes only at the points of the $n_1 n_2$ differences $U_{kk'}$ (BAUER (1972)). To receive the confidence interval for Δ the null distribution of the test statistic $W^{WMW}(d)$ is linked to the differences in Ω. Let $w^{WMW} = w^{WMW}_{\alpha/2,n_2,n_1}$ be the lower $\alpha/2$-quantile of the null-distribution of the WMW test and $n_1 n_2 - w^{WMW}$ be the upper $\alpha/2$-quantile. Then the two-sided interval is given by

$$CI = (U_{lower}; U_{upper}) = (U_{(w^{WMW})}; U_{(n_1 n_2 - w^{WMW}+1)}), \tag{6.2}$$

with the ordered differences $U_{(1)} \leq \ldots \leq U_{(n_1 n_2)}$.
The one-sided confidence limit is calculated with the critical value according to α instead of $\alpha/2$. If ties occur the critical values have to be taken from the conditional distribution of the WMW test.
To compute an asymptotic confidence interval for the Hodges-Lehmann estimator the non-relevance-shifted rank sum test $W = \sum_{k=1}^{n_2} r_{2k}$, it's expectation $E(W) = \frac{n_2}{N} \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik}$ and it's variance $Var(W) = \frac{n_1 n_2}{N^2(N-1)} \left\{ N \sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik}^2 - \left(\sum_{i=1}^{2} \sum_{k=1}^{n_i} r_{ik} \right)^2 \right\}$ are needed,

where r_{ik} denote the ranks among x_{ijk}. Besides the exclusion of the relevance thresholds, these equations are exactly the same as introduced in section 4.2 for the nonparametric relevance-shifted test on differences.
The confidence limit is given, when $(U_{lower}; U_{upper})$ are found, such that

$$z_{\alpha/2} = \frac{W(U_{lower}) - E(W(U_{lower}))}{\sqrt{Var(W(U_{lower}))}} \quad \text{and} \quad z_{1-\alpha/2} = \frac{W(U_{upper}) - E(W(U_{upper}))}{\sqrt{Var(W(U_{upper}))}} \tag{6.3}$$

(Hothorn and Hornik, 2002). For a one-sided limit z_α or $z_{1-\alpha}$ is used, respectively.

With the confidence limit for Δ the test for a relevant ratio is constructed. Note that this test makes other assumptions regarding the data than the Wilcoxon rank sum test and the confidence limit for Δ. For the univariate case let $x_{1k} > 0$ and $x_{2k} > 0$ be independent and continuous observations from the distribution functions $F(x)$ and $G(x) = F(x/\kappa)$, with $k = 1, \ldots, n_i$ as the index of the repetition and $i = 1, 2$ as the index of the group.
To test whether κ is more extreme than a specific relevance threshold θ_{side}, the nonparametric confidence interval for the ratio of medians proposed by Hothorn & Munzel (2002) is used. The two-sided procedure is as follows:

1. Select relevance thresholds $\theta_{lower} \leq 1$ and $\theta_{upper} \geq 1$.

2. Logarithmize the data $y_{ik} = \log_e(x_{ik})$, where here the natural logarithm is used.

3. Dependent on the sample size calculate the nonparametric confidence interval $CI = (U_{lower}; U_{upper})$ for the Hodges-Lehmann estimate Δ by the use of y_{ik} either exact or asymptotic as defined in the former subsections.

4. Transform the limits of the confidence interval back into the original scale: $(\hat{\kappa}_{lower}; \hat{\kappa}_{upper}) = (e^{U_{lower}}; e^{U_{upper}})$.

5. The null hypothesis is rejected if either $\hat{\kappa}_{lower} > \theta_{upper}$ or $\hat{\kappa}_{upper} < \theta_{lower}$.

One-sided tests are computed by the use of the according one-sided confidence limit.
An alternative construction of the nonparametric test for relevant ratios is the use of the relevance-shifted rank sum test as introduced in section 4.2 with the use of both the logarithmized data and the logarithmized relevance thresholds (Pflüger and Hothorn (2002)). This alternative method results in the same FWER and power as the computation of the confidence limit.

6.1.1 The np-θ-shift procedure

With the use of the logarithm the multiplicative model switches to an additive model. Thus the confidence interval for the difference of medians can be used as a confidence for the ratio of medians. To compute the selectors the logarithmized data is used as well. As selector the interquartile range IQR is used. However compared to the nonparametric procedure with a data-driven order of hypotheses of Kropf *et al.* (2004), the data is logarithmized and the first treatment group is shifted by the amount of the logarithm of the relevance threshold. In particular for one-sided testing against a relevant increase to the logarithmized data of the first treatment group the logarithm of the upper relevance threshold is added and for one-sided testing on decrease the logarithm of the lower threshold is added. For two-sided testing the data is split in two parts: endpoints with a ratio of medians greater or equal to 1 follow the data shift as introduced for one-sided testing on increase and the other endpoints have a shift equal to the test on decrease.
Furthermore a data transformation similar to the one of the parametric θ-shift procedure from section 5.1.2 is embedded: all endpoints with a ratio of medians neither equal to nor exceeding the relevance threshold(s) receive a transformation, such that their transformed data has a ratio of medians equal to the threshold(s).
The algorithm of the 'np-θ-shift' procedure is:

1. Choose relevance thresholds $\theta_{lower} \leq 1$ and $\theta_{upper} \geq 1$.

2. Logarithmize the data, such that $y_{ijk} = \log_e x_{ijk}$, where $\log_e$ is the natural logarithm.

3. For each endpoint j:

   (a) Depending on the sample size compute the limits $\hat{\kappa}_{lower}$ and $\hat{\kappa}_{upper}$ of the confidence interval for the Hodges-Lehmann estimate, either exact or asymptotic as introduced in section 6.1 by the use of y_{ijk}. Use the unadjusted $\alpha/2$ for the computation of the confidence limit.

   (b) Transform the data of the second group to

$$y^*_{2jk} = \begin{cases} y_{2jk} - \tilde{y}_{2j} + \tilde{y}_{1j} + log_e(\theta_{upper}); & \forall\, j \mid 0 \leq \tilde{y}_{2j} - \tilde{y}_{1j} < log_e(\theta_{upper}) \\ y_{2jk} - \tilde{y}_{2j} + \tilde{y}_{1j} + log_e(\theta_{lower}); & \forall\, j \mid log_e(\theta_{lower}) < \tilde{y}_{2j} - \tilde{y}_{1j} < 0 \\ y_{2jk}; & \text{else.} \end{cases} \quad (6.4)$$

   If $\tilde{y}_{2j} - \tilde{y}_{1j} < 0$ compute the interquartile range $IQR_j = q_{75,j} - q_{25,j}$ as selector statistic among the combined samples $y^*_{1jk} = y_{1jk} + log_e(\theta_{lower})$ and y^*_{2jk}. Otherwise the IQR_j is calculated from the pooled samples $y^*_{1jk} = y_{1jk} + log_e(\theta_{upper})$ and y^*_{2jk}.

4. Sort the m confidence intervals $CI = (\hat{\kappa}_{j,lower}; \hat{\kappa}_{j,upper})$ for decreasing selectors IQR_j.

5. For each endpoint independently compare the j^{th} ordered $\hat{\kappa}_{j,lower}$ with θ_{upper} and $\hat{\kappa}_{j,upper}$ with θ_{lower}. It is significant, if either $\hat{\kappa}_{j,lower} \geq \theta_{upper}$ or $\hat{\kappa}_{j,upper} \leq \theta_{lower}$.

6. Stop at the first non-significance and accept for all further endpoints the null hypothesis.

In the one-sided case, the endpoints with ratios not exceeding the relevance criteria are transformed by the equation with the corresponding relevance threshold only. The selectors are computed by the use of $y_{1jk} + \log \theta_{upper}$ and y^*_{2jk} (increase) and $y_{1jk} + \log \theta_{lower}$ and y^*_{2jk} (decrease).

To illustrate that the procedure solves both the problems of the maximum of the selector statistic at 1 and the lack of power due to different data levels, two graphics are shown next. The left graphic is generated under similar conditions as 5.1 with one exception. The nonparametric test on ratio assumes, that the population of the second sample is the

κ-fold of the population corresponding to the first sample of endpoint j. Hence instead of setting the true means to $\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot \kappa^{H_1}$ only, the standard deviation is multiplied by κ as well: $\sigma_{1j} = 10$ and $\sigma_{2j} = 10 \cdot \kappa^{H_1}$, where this is the setting for endpoints with a true ratio of medians greater or equal than 1. For endpoints with a ratio less than 1 the expected values are set to $\mu_{1j} = 100 \cdot (1/\kappa^{H_1})$ and $\mu_{2j} = 100$ and the true standard deviations are $\sigma_{1j} = 10 \cdot (1/\kappa^{H_1})$ and $\sigma_{2j} = 10$. The right figure has a different setting of σ_{ij}. While in the left one for all three mean levels σ_{ij} was set to 10, it depends now on the mean level with a coefficient of variation of 10% - which is far more realistic than the former setting.

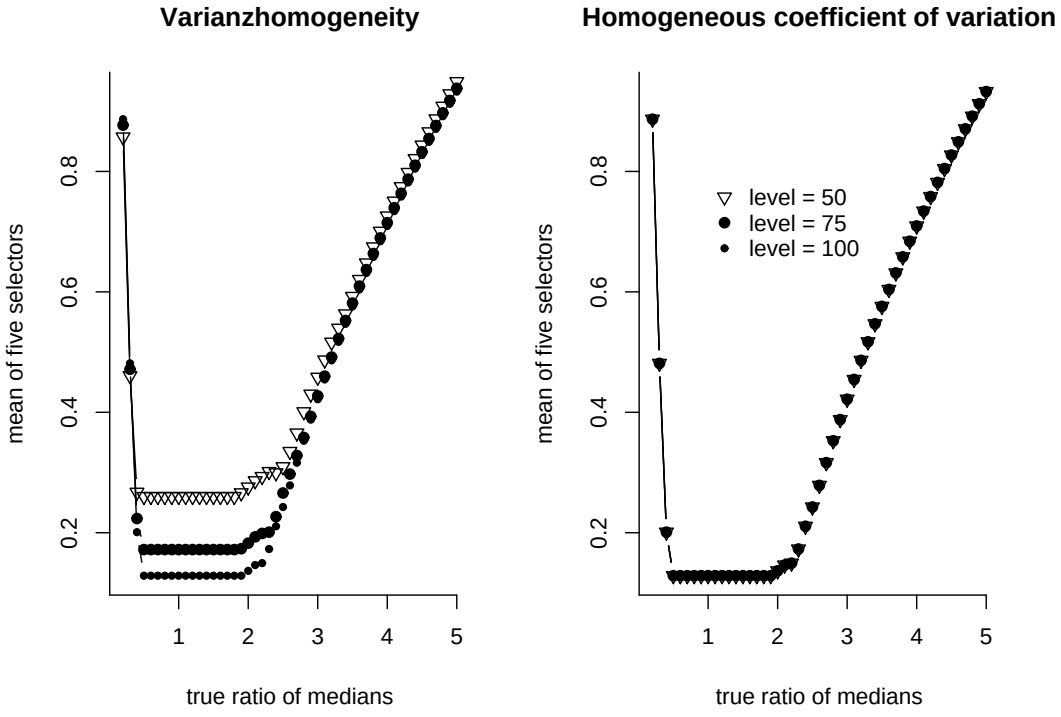

Figure 6.1: np-θ-shift procedure: comparison of selectors for different data levels

In the left figure the selector for endpoints with $\theta_{lower} \leq \frac{\tilde{x}_{2j}}{\tilde{x}_{1j}} \leq \theta_{upper}$ increases for decreasing data levels. Due to the logarithmic transformation large values move together much more than small values. Because the variances are homogeneous among all three curves and only

a constant is added, the IQR for samples with larger mean levels is smaller.
This effect is negated with a constant coefficient of variation, as it can be seen in the right graphic. With the logarithmic algorithm $log(a \cdot b) = log(a) + log(b)$, one can see that the mean level between the curves change, but the variances stay constant. Hence this method is appropriate for the analysis of microarray data, given the logarithmic transformation stabilizes the variances among the genes sufficiently.

6.1.2 The relevance-shifted permutation algorithm for step-down minP adjusted *p*-values

In many experimental questions test statistics and therefore p-values are correlated. For microarray data groups of genes are co-regulated and correlations between them and hence between their p-values occur. The Bonferroni adjustment, which is used for comparison as well, does not consider this correlation structure. Among others this is the reason for its conservativeness. The permutation algorithm for step-down minP adjusted p-values according to Westfall and Young (1993) regards the dependence structure of the test statistics, and hence can achieve a much higher power. Thus the nonparametric procedure with a data-driven order of hypotheses to test for relevant ratios shall be compared with the permutation algorithm, as it is the most important procedure for the analysis of microarrays proposed here. The comparison with the other methods is abandoned, because the permutation algorithm is very time consuming.
The traditional double permutation algorithm as proposed by Westfall and Young (1993) is split in two parts. First permuted raw (unadjusted) p-values are computed for each endpoint and afterwards the raw p-values are adjusted with a second resampling algorithm. Due to the two resampling runs, this "double permutation algorithm" is very time consuming. A faster algorithm computing exactly the same results is presented by Ge *et al.* (2003). This algorithm needs the permutation process only once, because for the second permutation algorithm the permutations of the first one are used.
However the permutation algorithm according to Westfall and Young (1993) does

not consider relevance-shifted tests. Hence a modification in terms of a relevance-shifted algorithm is used here for comparison. No mathematical proof exists for this modification and the empirical control of the FWER is shown by simulations only.

The modified permutation algorithm splits in two parts: first the multiplicity adjusted p-values for one-sided testing against a relevant decrease; and in the second step the adjusted p-values for a relevant increase are computed. Afterwards for each endpoint the two corresponding one-sided p-values are combined to a two-sided p-value. For the computation the fast algorithm of Ge *et al.* (2003) is used.

Computation of the modified permutation algorithm: As explained above first the permutation raw p-values p_j^* have to be computed.
Create the pseudosamples $y_{1jk} = log_e(x_{1jk}) + log_e(\theta_{lower})$ and $y_{2jk} = log_e(x_{2jk})$.

Permutation algorithm for raw p-values: For each gene compute the one-sided asymptotic Wilcoxon rank sum test W_j on decrease as introduced in section 3.1.2.
For the b^{th} permutation, $b = 1, \ldots, B$:

1. Permute the $n_1 + n_2 = N$ columns of the $m \times N$ dimensional data matrix $\mathbf{X}$.

2. Compute the test statistics $W_{1b}, \ldots, W_{mb}$ for each hypothesis.

After the B permutations are done the one-sided raw p-value for hypothesis $H_{0,j} : \kappa \geq \theta_{lower}$ is

$$p_{j,lower}^* = \frac{\#\{b : W_{j,b} \leq W_j\}}{B} \quad \text{for} \quad j = 1, \ldots, m. \tag{6.5}$$

Permutation algorithm for multiplicity correction: Instead of doing additional B permutations to adjust p-values for multiplicity, the already existing B permutations are used for further computations. In the following algorithm three $m \times B$ matrices are computed.
The first one will include the test statistics:

$$\mathbf{T} = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1b} & \cdots & W_{1B} \\ \vdots & \vdots & & \vdots & & \vdots \\ W_{j1} & W_{j2} & \cdots & W_{jb} & \cdots & W_{jB} \\ \vdots & \vdots & & \vdots & & \vdots \\ W_{m1} & W_{m2} & \cdots & W_{mb} & \cdots & W_{mB} \end{pmatrix}. \tag{6.6}$$

In the second matrix the raw p-values will be filled in:

$$\mathbf{P} = \begin{pmatrix} p_{jb} \end{pmatrix} \tag{6.7}$$

and the matrix $\mathbf{Q}$ will contain the minima of raw p-values:

$$\mathbf{Q} = \begin{pmatrix} q_{jb} \end{pmatrix} \tag{6.8}$$

as explained below in the algorithm.

For $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{Q}$ the b^{th} column corresponds to the data matrix $\mathbf{X}$ with permuted columns. Next follows the algorithm by use of these three matrices.

1. Assume that the raw permuted p-values computed as described above are $p^*_{1,lower} \leq p^*_{2,lower} \leq \ldots \leq p^*_{m,lower}$, otherwise sort the rows of the data matrix $\mathbf{X}$ according to the ordered $p^*_{j,lower}$. Set $j = m$ and $q_{m+1,b} = 1$ for $b = 1, \ldots, B$.

2. For hypothesis $H_{0,j}$ (row j), use the B permutation test statistics $W_{j,1}, \ldots, W_{j,B}$ and get the B one-sided raw p-values $p_{j,1}, \ldots, p_{j,B}$ with

$$p_{j,b} = \frac{\#\{b' : W_{j,b'} \leq W_{j,b}\}}{B}, \tag{6.9}$$

   which is in row j for each $W_{j,b}$ the percentage of equal or smaller test statistics $W_{j,b'}$.

3. Update the successive minima $q_{j,b}$

$$q_{j,b} \leftarrow \min(q_{j+1,b}, p_{j,b}),\ b = 1, \ldots, B. \tag{6.10}$$

4. Compute the adjusted p-value for hypothesis $H_{0,j} : \kappa \geq \theta_{lower}$:

$$\tilde{p}^*_{j,lower} = \frac{\#\{b : q_{j,b} \leq p^*_{j,lower}\}}{B}. \tag{6.11}$$

5. Move up one row, i.e. $j \leftarrow j - 1$. If $j = 0$, go to step 7, otherwise go to step 2.

6. Enforce monotonicity of $\tilde{p}^*_{j,lower}$:

$$\tilde{p}^*_{1,lower} \leftarrow \tilde{p}^*_{1,lower}, \qquad \tilde{p}^*_{j,lower} \leftarrow \max(\tilde{p}^*_{j-1,lower}, \tilde{p}^*_{j,lower}) \qquad \text{for} \quad j = 2, \ldots, m. \tag{6.12}$$

Repeat the entire procedure with the pseudosamples $y_{1jk} = log_e(x_{1jk}) + log_e(\theta_{upper})$ and $y_{2jk} = log_e(x_{2jk})$ and the asymptotic rank sum test on increase to achieve the one-sided multiplicity adjusted p-values on increase $\tilde{p}^*_{j,upper}$. Then the two-sided adjusted p-values are given by $\tilde{p}^*_j = \min(2 \cdot \tilde{p}^*_{j,lower}, 2 \cdot \tilde{p}^*_{j,upper})$.

The R package `multtest` by Pollard *et al.* (year not specified by authors) includes the new permutation algorithm of Ge *et al.* (2003). In this package the asymptotic Wilcoxon rank sum test can be used as the two-sample test. And as the relevance shift of the data is prior the computation of the algorithm this package is used for the simulation of the power and the FWER. In these simulations the relevance-shifted permutation algorithm is abbreviated as 'minP'. Simulations results of the proportional power are given in chapter 7.

6.2 Control of the FWER

For the np-θ-shift procedure the control of the FWER is tested in detail. Due to the high duration of the simulations for the minP method, a limited set of scenarios is observed. For both methods 50 endpoints are tested with various conditions are tested with varying sample sizes (unbalanced designs as well), variances, correlations structures and relevance criteria. Most of the simulations include multivariate Gaussian distributed data; however

deviations from this assumption are tested as well.

Only for the np-θ-shift procedure is additionally the asymptotic method tested. Furthermore, for the exact approach scenarios are tested, where the correlation structure among the endpoints is not constant; in particular three of 50 endpoints have a different correlation. Usually in all scenarios the mean level of the data is fixed. For the np-θ-shift method scenarios are included with mean levels following the exponential distribution. In these scenarios either the variance or the coefficient of variation is constant. Finally only for this method and not for the minP one-sided tests are observed as well.

The np-θ-shift procedure shows an equivalent behavior of the simulated FWER as for all three θ- or δ-shifted methods as well. For one-sided testing the empirical FWER exceeds in some situations with $\theta_{lower} = \theta_{upper} = 1$ slightly the nominal level. The largest exceed is 6.30% (nominal level: 5%). But in the more relevant case of two-sided testing the FWER is controlled asymptotically. In the case of two-sided testing no exceeds are found. For the minP algorithm only two-sided scenarios are tested. Here no exceeds occurred.

6.3 Examples

The three example data sets are analyzed in this section by use of the nonparametric tests on ratio. As in all former chapters two-sided tests are used and the significance level is set to 0.05.

6.3.1 Possum data set

For the possum data set the relevance thresholds are set to $\theta_{lower}^{-1} = \theta_{upper} = 1.02$. The following table shows the result of the np-θ-shift procedure, the Bonferroni correction and the minP algorithm:

endpoint	ratio of medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value	minP adjusted p-value	Bonferroni adjusted p-value
earconch	0.859	0.130	10	0.004	0.004	0.024	0.036
belly	1.054	0.113	28	0.808	0.808	1.000	1.000
chest	1.097	0.088	34	0.198	0.808	0.812	1.000
eye	1.093	0.084	35	0.149	0.808	0.723	1.000
skullw	1.077	0.072	34	0.214	0.808	0.800	1.000
totlngth	1.055	0.061	33	0.267	0.808	0.800	1.000
footlgth	0.945	0.049	17	0.154	0.808	0.610	1.000
taill	1.057	0.045	34	0.194	0.808	0.800	1.000
hdlngth	1.060	0.040	35	0.154	0.808	0.747	1.000

All three methods reject the null hypothesis of endpoint 'earconch'. All other endpoints have unadjusted p-values larger than 0.05, hence none of the procedures could have led to more significant results.

It has to be noted that due to the logarithmic transformation the ordering of the endpoints differ from all previous analysis. Especially the endpoint 'belly' has a comparatively larger selector.

To demonstrate the data transformation the statistics of the modified endpoint 'belly' are computed as well.

Results of the data transformations:

procedure	ratio of medians	selector statistic	test statistic	unadjusted p-value
np-θ-shift	0.995	0.109	27	1.000

Apart from the statistics for 'belly' the result does not change.

6.3.2 TSHR mutation data set

In this section the microarray data set is analyzed by the nonparametric tests for relevant ratios. As well as in the former chapters the relevance thresholds are set to $\theta_{lower}^{-1} = \theta_{upper} =$

1.5. The analysis begins with the unadjusted exact rank sum test for relevant ratios. Here 48 significant genes are found. As could be expected, no significant endpoints can be found with the Bonferroni adjustment. In particular all adjusted p-values are 1. The same result is achieved by the procedure with a data-driven order of hypotheses. The next table lists the ten endpoints with the highest selector statistic:

endpoint	ratio of medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value
2808	2.568	2.709	77	1.000	1.000
6022	3.671	2.654	77	1.000	1.000
2148	3.404	2.635	89	0.310	1.000
9207	0.897	2.625	83	1.000	1.000
12337	0.902	2.606	86	1.000	1.000
3568	1.300	2.600	75	1.000	1.000
7940	0.069	2.592	59	0.008	1.000
1383	0.264	2.474	72	0.371	1.000
6168	7.881	2.462	91	0.206	1.000
12462	0.577	2.413	82	1.000	1.000

Only the seventh endpoint (# 7940) has an unadjusted p-value less than 0.05. Finally the modified permutation algorithm is applied on the data. It achieves the same result as the Bonferroni adjustment: all adjusted p-values are 1.

Regarding the results of all procedures applied on the microarray data set, the procedures with a data-driven order of hypotheses did not result in more significant endpoints compared to the alternative multiple testing approaches. Beside the single significant endpoint found by the Bonferroni adjustment with a non-relevance-shifted t-test on the original scaled data, no further significant results were achieved with the alternative methods.

It can however be summarized, that although the sample sizes are not extremely small for microarray data, the alternative methods lack power. With an even smaller number of observations these approaches are unlikely to find any significant result because both methods suffer from the discreteness of the p-values. But the multiplicity correction of the procedure with a data-driven order of hypotheses is based on an order of the endpoints.

Hence irrespective of the number of endpoints and sample sizes (unless smaller than 4 per group), it can achieve significant results.

6.3.3 TNFα data set

Finally the TNFα data set is analyzed with the nonparametric procedures to test for a relevant ratio. As well as in the other analysis the thresholds are set to $\theta_{lower}^{-1} = \theta_{upper} = 1.5$. For the application of the np-θ-shift procedure the a-priori logarithmic transformation has to be taken into account; the transformation has to be omitted and instead of exponentiating the confidence limits, a base of 10 has to be used.

Without any multiplicity correction 370 significant genes can be found. As described in chapter 3 it is not possible to find significant endpoints with the Bonferroni adjustment, because the sample sizes are too small. The minP algorithm lacks power for the same reasons. With a sample size of 5 per group resulting in 252 permutations, the procedure becomes too discrete to find any significant endpoints. However by use of the np-θ-shift procedure six significant endpoints are achieved:

endpoint	ratio of medians	selector statistic	test statistic	unadjusted p-value	adjusted p-value
5979	53.432	1.520	40	0.008	0.008
11600	46.206	1.477	40	0.008	0.008
13585	43.662	1.447	40	0.008	0.008
13618	45.730	1.417	40	0.008	0.008
6629	36.737	1.370	40	0.008	0.008
8563	35.596	1.344	40	0.008	0.008
7018	38.869	1.334	15	0.100	0.100
13653	37.085	1.315	40	0.008	0.100
12453	0.129	1.311	10	0.629	0.629
7599	28.184	1.263	40	0.008	0.629

Chapter 7

Power simulations for relevance-shifted tests

In this chapter simulation results of the power for all relevance-shifted procedures discussed in this work are presented. For simulation the same scenarios as in chapter 3 are applied and for each scenario - as for example increasing sample size or disturbance - the results of all four types of methods are printed. As it will be seen in the graphics the behavior of the power of the relevance-shifted procedures with a data-driven order of hypotheses is similar to the methods with point-zero hypotheses.

As in chapter 3 two-sided testing scenarios with 50 endpoints are observed, where five have relevant differences between the treatment groups and 45 are under H_0. However in this chapter the true differences or ratios for endpoints under H_0 are not any longer 0 or 1. Here they are random values and are greater than the lower and less than the upper threshold. These settings are introduced in section 4.1.1 for tests on relevant differences and in section 5.1.1 for the parametric testing against a relevant ratio. For power simulations of the nonparametric testing methods the random treatment effect κ^{H_0} is derived as explained for the treatment effect $\tau_\theta^{H_0}$ for the parametric analog. The expected values and standard deviations are set to $\mu_{1j} = 100 \cdot \kappa^{H_0}$, $\mu_{2j} = 100$ and $\sigma_{1j} = \sigma_{2j} \cdot \kappa^{H_0}$ for tests on decrease and $\mu_{1j} = 100$, $\mu_{2j} = 100 \cdot \kappa^{H_0}$ and $\sigma_{2j} = \sigma_{1j} \cdot \kappa^{H_0}$ for tests on increase.

The expected values of the five endpoints under H_1 are set as follows:

procedure	treatment effect
parametric & nonparametric test on difference	$\mu_{1j} = 100 + \tau_\delta^{H_1}$ and $\mu_{2j} = 100$ (two endpoints)
	$\mu_{1j} = 100$ and $\mu_{2j} = 100 + \tau_\delta^{H_1}$ (three endpoints)
parametric test on ratio	$\mu_{1j} = 100 \cdot \tau_\theta^{H_1}$ and $\mu_{2j} = 100$ (two endpoints)
	$\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot \tau_\theta^{H_1}$ (three endpoints)
nonparametric test on ratio	$\mu_{1j} = 100 \cdot \kappa^{H_1}$, $\mu_{2j} = 100$ and $\sigma_{1j} = \sigma_{2j} \cdot \kappa^{H_1}$ (two endpoints)
	$\mu_{1j} = 100$, $\mu_{2j} = 100 \cdot \kappa^{H_1}$ and $\sigma_{2j} = \sigma_{1j} \cdot \kappa^{H_1}$(three endpoints)

If not stated otherwise, multivariate normal distributed random numbers are used. Furthermore the FWER is set to 5%, the standard deviation is $\sigma_{ij} = 10$ and the individual simulation results are calculated with 10,000 simulation runs each. The settings of all further parameters are given in the individual sections. As well as in chapter 3 the new procedures are compared to standard methods. In all graphics the power results of local tests with the corresponding two-sample test are printed and these local tests are abbreviated as 'local'. Except for the ones of the nonparametric tests on relevant ratios, in all figures the power results of the corresponding two-sample tests with the α-adjustment of Bonferroni are given. These are abbreviated as 'Bonferroni'. The nonparametric test for relevant ratios with a data-driven order of hypotheses is compared with the modified permutation algorithm for step-down minP adjusted p-values.
In the first graphics (section 7.2) results of an additional definition of the error rate to control the multiplicity problem is plotted for comparison. This error is the k-FWER, which shall be introduced here.

7.1 The *k*-FWER

As discussed in the introduction the FWER is a stringent error rate to control the multiplicity problem. For microarray data a researcher may tolerate a larger number of false rejections to prevent a miss in possible candidate genes. Then, instead of controlling the

probability of at least one false rejection (FWER), the probability of rejection of at least k true null hypotheses can be used. This so-called k-FWER is proposed by Lehmann and Romano (2005) and it is formerly defined as:

$$k - \text{FWER} = P(\text{reject at least } k \text{ of } H_{0,1}, H_{0,2}, \ldots, H_{0,m'} \mid H_{0,1}, H_{0,2}, \ldots, H_{0,m'} \text{ are true}). \tag{7.1}$$

By setting $k = 1$ the k-FWER reduces to the common FWER.
In their article Lehmann and Romano (2005) introduce a Bonferroni variant for the k-FWER, which will be used for comparison in section 7.2. The algorithm is initially the same as for the control of the FWER, but the individual p-values are compared with $k\alpha/m$ instead of α/m.

7.2 Power for varying treatment effect, relevance thresholds and correlation

In total analogy to chapter 3 the first graphics show the proportional power for varying treatment effects and correlations among the endpoints. In addition the behavior of the power is shown for varying relevance thresholds as well. The simulation settings are:

Tests on relevant differences

figure	$\tau_\delta^{H_1}$	$-\delta_{lower} = \delta_{upper}$	$\rho_{ijj'}$
a)	100 to 150 in steps of 5 units	100	0.01
b)	100 to 150 in steps of 5 units	100	0.5
c)	100 to 150 in steps of 5 units	100	0.999

Tests on relevant ratios

figure	$\tau_\theta^{H_1}$ and κ^{H_1}	$\theta_{lower}^{-1} = \theta_{upper}$	$\rho_{ijj'}$
a)	1 to 1.5 in steps of 0.05 units	1	0.01
b)	2 to 2.8 in steps of 0.05 units	2	0.01
c)	5 to 7 in steps of 0.1 units	5	0.01
a)	1 to 1.5 in steps of 0.05 units	1	0.5
b)	2 to 2.8 in steps of 0.05 units	2	0.5
c)	5 to 7 in steps of 0.1 units	5	0.5
a)	1 to 1.5 in steps of 0.05 units	1	0.999
b)	2 to 2.8 in steps of 0.05 units	2	0.999
c)	5 to 7 in steps of 0.1 units	5	0.999

For the tests on difference the sample size per group is set to 7 and the test on ratios have $n_i = 5$. The reason for the larger sample size for the tests on difference is the discreteness of nonparametric test with the adjustment of Bonferroni controlling the FWER. It requires a sample size of at least 7 to compute a significant p-value. Instead of the Bonferroni adjustment the permutation algorithm is used for comparison with the nonparametric test on ratio with a data-driven order of hypotheses. Here a sample size of 5 is sufficient.

Only in this section are results from the Bonferroni correction by controlling the k-FWER shown. These are abbreviated in the legend as 'k-FWER'.

Further, in the graphics for the relevance-shifted nonparametric tests on difference, the power curve of the corresponding parametric δ-shift method is plotted also for comparison.

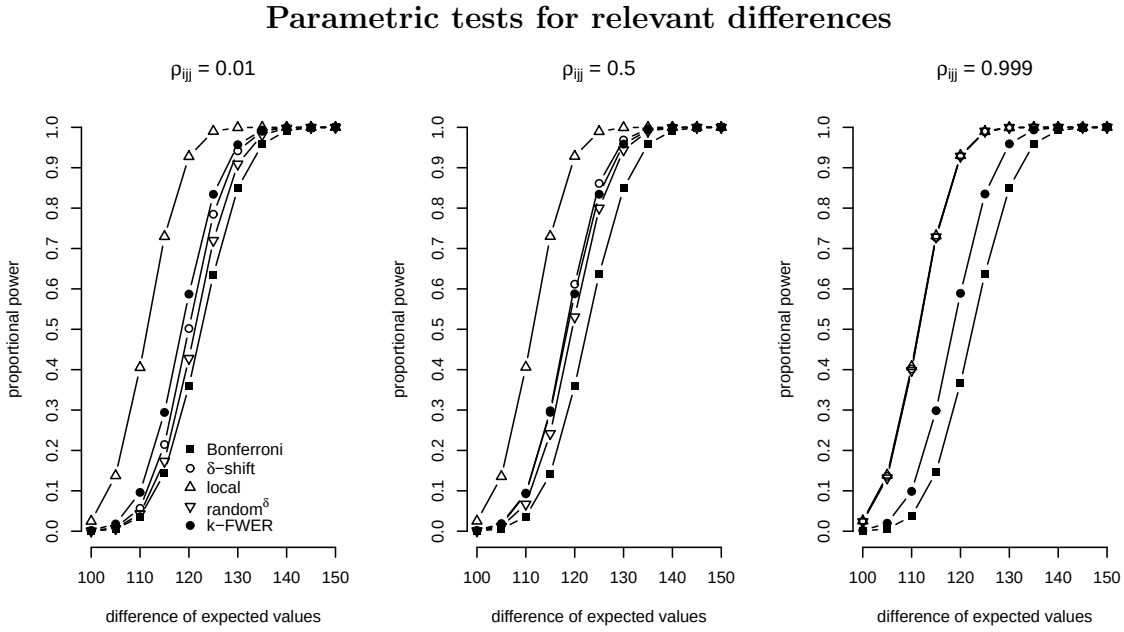

Figure 7.1: Parametric test for rel. diff.: Power for different correlation structures

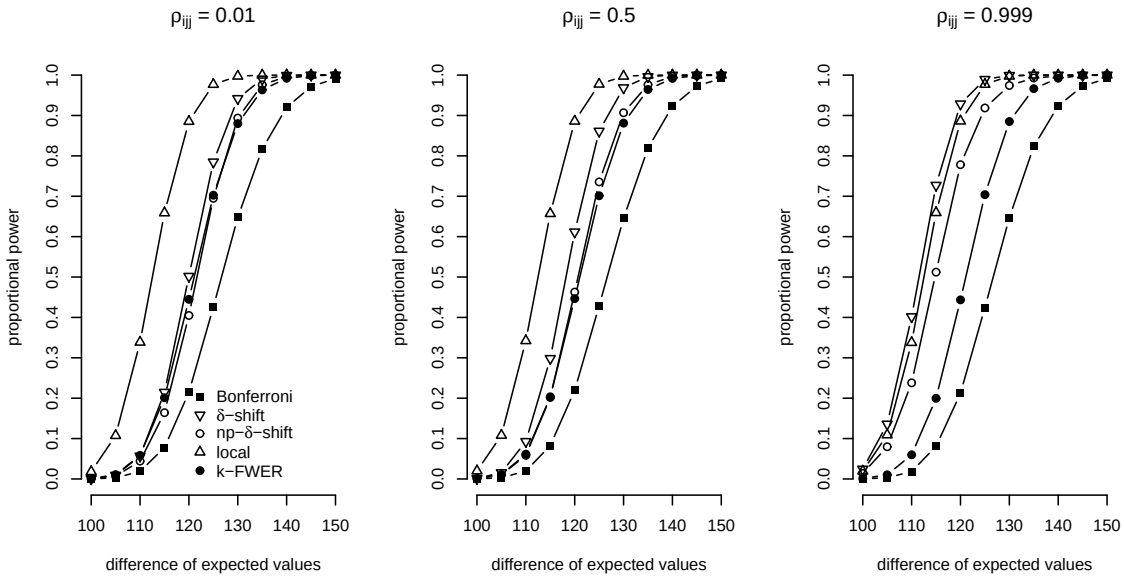

Figure 7.2: Nonparametric test for rel. diff.: Power for different correlation structures

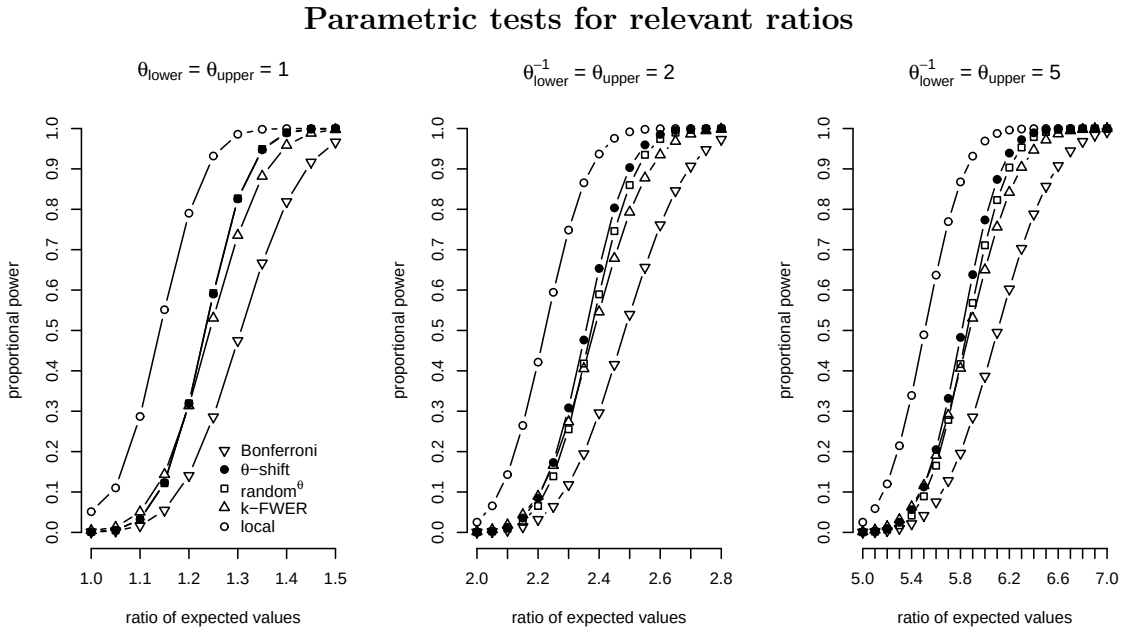

Figure 7.3: Parametric test for rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.01$

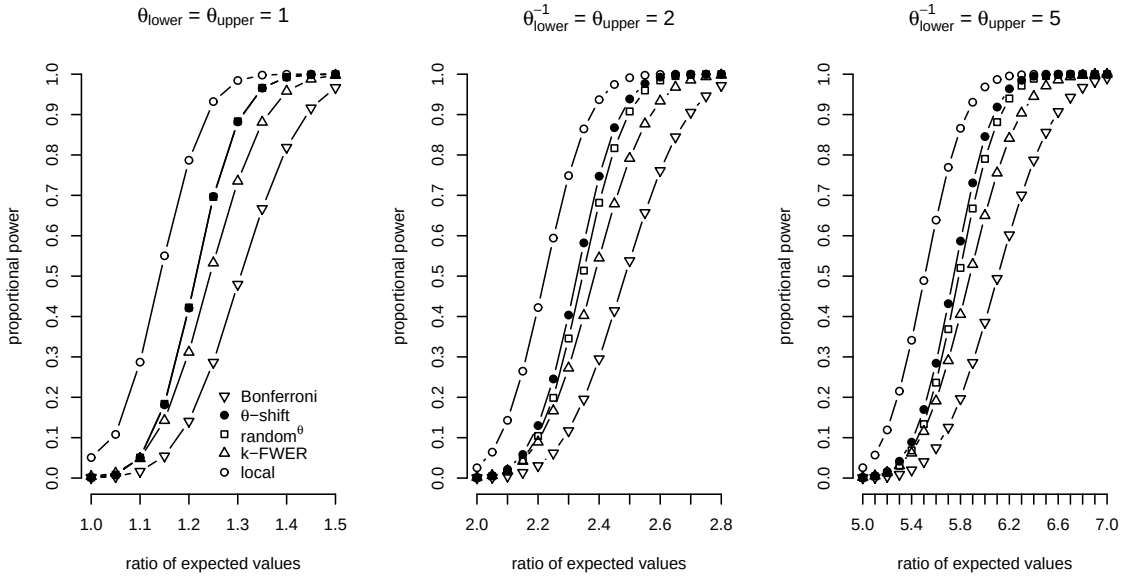

Figure 7.4: Parametric test for rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.5$

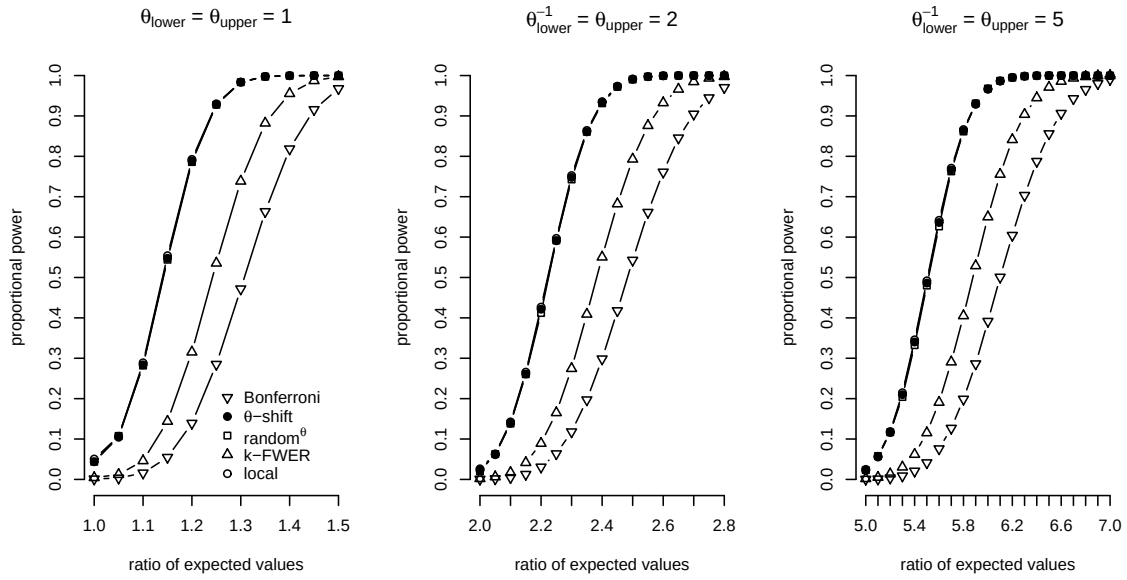

Figure 7.5: Parametric test for rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.999$

Nonparametric tests for relevant ratios

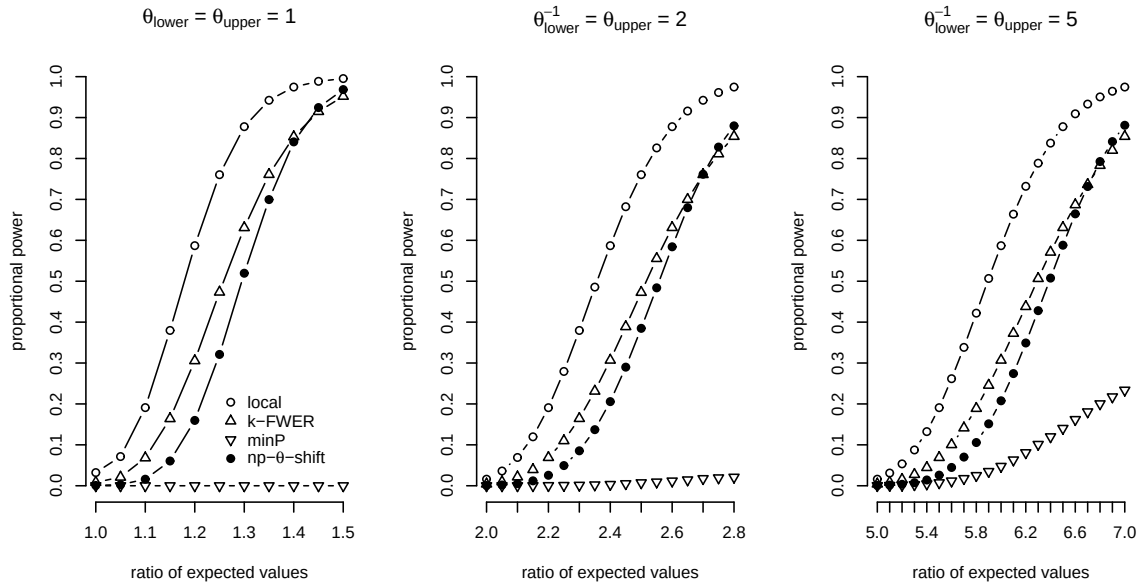

Figure 7.6: Nonparametric test on rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.01$

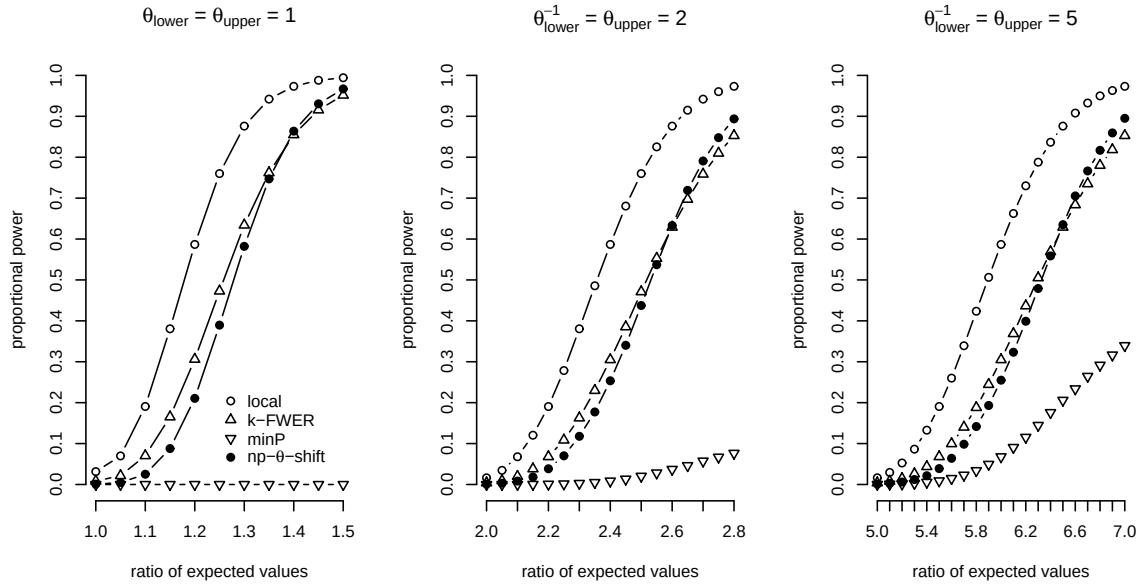

Figure 7.7: Nonparametric test on rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.5$

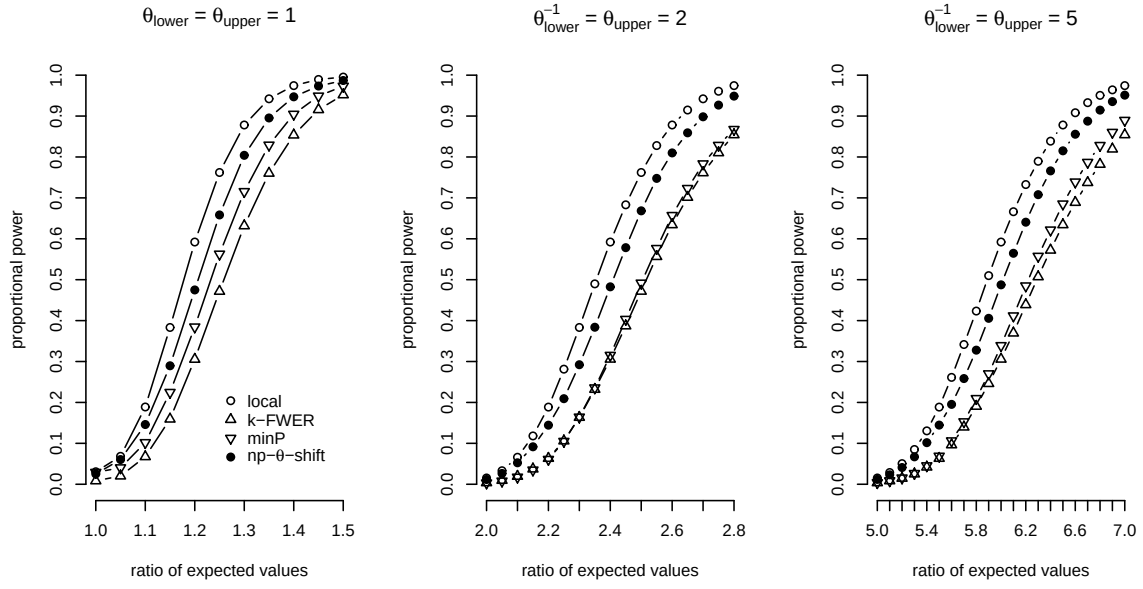

Figure 7.8: Nonparametric test on rel. ratios: Power for different θ_{side}; with $\rho_{ijj'} = 0.999$

In general, for all procedures with a data-driven order of hypotheses the power increases with an increasing correlation among the endpoints. If the correlation is set to $\rho_{ijj'} = 0.999$, the power of the parametric ones equals the power of the parametric local tests. The nonparametric procedures do not reach such a high power.

As the random numbers are generated from a normal distribution, the parametric procedure with a data-driven order of hypotheses for relevant differences is more powerful compared to the nonparametric one. This can be seen for the procedures with a data-driven order to test for a relevant ratio as well. However they cannot be compared as both have different underlaying data conditions.

For the tests on relevant differences all scenarios have the same relevance thresholds. The reason for this is, that the power is not dependent on them. There is only one exception: the power is slightly increased with thresholds different from 0 compared to the procedure for point-zero hypotheses. This increase is based on the generation of the endpoints under H_0. For tests with point-zero hypotheses ($\delta_{lower} = \delta_{upper} = 0$) all empirical deviations of $\mu_{2j} - \mu_{1j} = 0$ result in an increased selector. For the relevance-shifted procedures all endpoints under H_0 have a true random difference in means of $\delta_{lower} \leq \mu_{2j} - \mu_{1j} \leq \delta_{upper}$. Most of the endpoints receive a data transformation and thus a small selector, which is unlikely to abort the procedure before false negatives can occur. Largest difference in power are around 5% (see for comparison the results in chapter 3). It has to be noted that for testing against a relevant difference the power difference occurs only for the procedures with a data-driven order. The local tests and the Bonferroni adjusted ones have no data-driven order of hypotheses and are thus not influenced by a change of the thresholds. Furthermore the differences vanish, when the endpoints under H_0 receive a true difference in means equal to one of the thresholds.

The dependency of the power on the selection of the thresholds is more important for tests on a relevant ratio. Except for the permutation algorithm all procedures lose power with an increase of the relevance criteria. For the parametric procedures this effect is clarified

with the equation for sample size estimation of the Sasabuchi-test:

$$n \geq (1 + \theta_{side}^2) \cdot (t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2})^2 \cdot \left(\frac{CV_1}{\tau_\theta - \theta_{side}}\right)^2. \tag{7.2}$$

Let $\tau_\theta - \theta_{side}$, CV_1, n, α and β be a constant value for different θ_{side}. Then $(t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2})^2 \cdot \left(\frac{CV_1}{\tau^\theta - \theta_{side}}\right)^2$ can be omitted. Only $1 + \theta_{side}^2$ is left of the equation, which is the reason of the decrease in power for increasing θ_{side}: $(t_{1-\alpha/2,2n-2} + t_{1-\beta/2,2n-2})^2 \cdot \left(\frac{CV_c}{\tau^\theta - \theta_{side}}\right)^2$ is among others multiplied by the squared relevance threshold. For the analyzed examples the behavior power of the nonparametric tests except of the permutation algorithm is multiplicative. That is, for a k-fold treatment effect against the threshold the power stays constant for varying θ_{side}. For example in the scenario with $\rho_{ijj'} = 0.5$ and $\theta_{lower}^{-1} = \theta_{upper} = 2$ the true treatment effect is the 1.2fold of the threshold and the true ratio of means is set to 2.4. In this scenario the nonparametric test with a data-driven order of hypotheses achieves a power around 25.35%. Likewise for $\theta_{lower}^{-1} = \theta_{upper} = 5$ the true ratio is set to $1.2 \cdot 5 = 6$ and the power is around 25.498%, which is the same assuming the small differences result from the simulation error.

The permutation algorithm lacks power for small sample sizes, because a small number of repetitions results in a limited number of possible permutations. Hence already the unadjusted p-values are discrete and with the small number of permutations for the multiplicity correction (here: 252) the discreteness becomes even more severe. Even for extremely high correlated endpoints it shows an inferior power behavior compared to the procedure with a data-driven order of hypotheses. In addition a correlation of nearly 1 among the endpoints does not reflect the empirical correlation among the genes in a microarray experiment. The following figure shows a histogram of the empirical pairwise correlation coefficients of the TSHR example data set. Correlations are computed from the first 6,000 genes of the patients with a mutated TSH receptor.

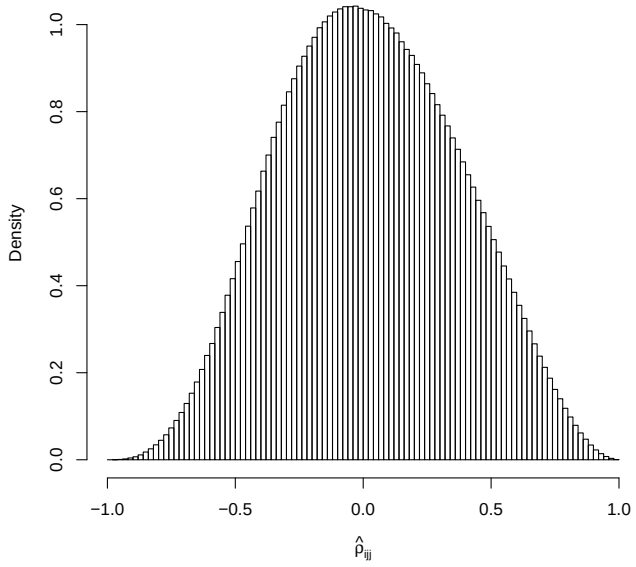

Figure 7.9: Histogram of empirical pairwise correlations among the endpoints

90.2% of the empirical correlation coefficients are less than $|\pm 0.5|$ and 76.8% are less than $|\pm 0.3|$. Shedden (2004) presents similar results with even a higher proportion of small absolute coefficients of correlation.

Finally the power results of the Bonferroni approach controlling the k-FWER are discussed. In comparison with the procedures with a data-driven order of hypotheses, the k-FWER method is in all scenarios except for the parametric tests on ratio superior, if the correlation is is near 0. This superiority decreases with an increasing correlation among the endpoints. Hence it can be said, that the FWER approaches are not necessarily less powerful compared to methods controlling less stringent error rates. Even if only the k-FWER has to be controlled, the conservative Bonferroni correction achieves a smaller power.

7.3 Power for varying sample sizes and α

The following figures show the behavior of the power for different sample sizes per group and four different levels of α. The treatment effects and relevance thresholds are set as listed in the following table:

procedure	treatment effect	relevance thresholds
parametric & nonparametric on difference	$\tau_\delta^{H_1} = 115$	$-\delta_{lower} = \delta_{upper} = 100$
parametric test on ratio	$\tau_\theta^{H_1} = 2.35$	$\theta_{lower}^{-1} = \theta_{upper} = 2$
nonparametric test on ratio	$\kappa^{H_1} = 2.35$	$\theta_{lower}^{-1} = \theta_{upper} = 2$

For all simulated experiments the standard deviation is set to $\sigma_{ij} = 10$ and the correlation is set to $\rho_{ijj'} = 0.3$. As stated above the number of permutations for the minP approach is set to 10,000, except for $\alpha = 0.01$. For $\alpha = 1\%$ up to 20,000 simulation runs are used, because with 10,000 permutations the null distribution is too discrete.

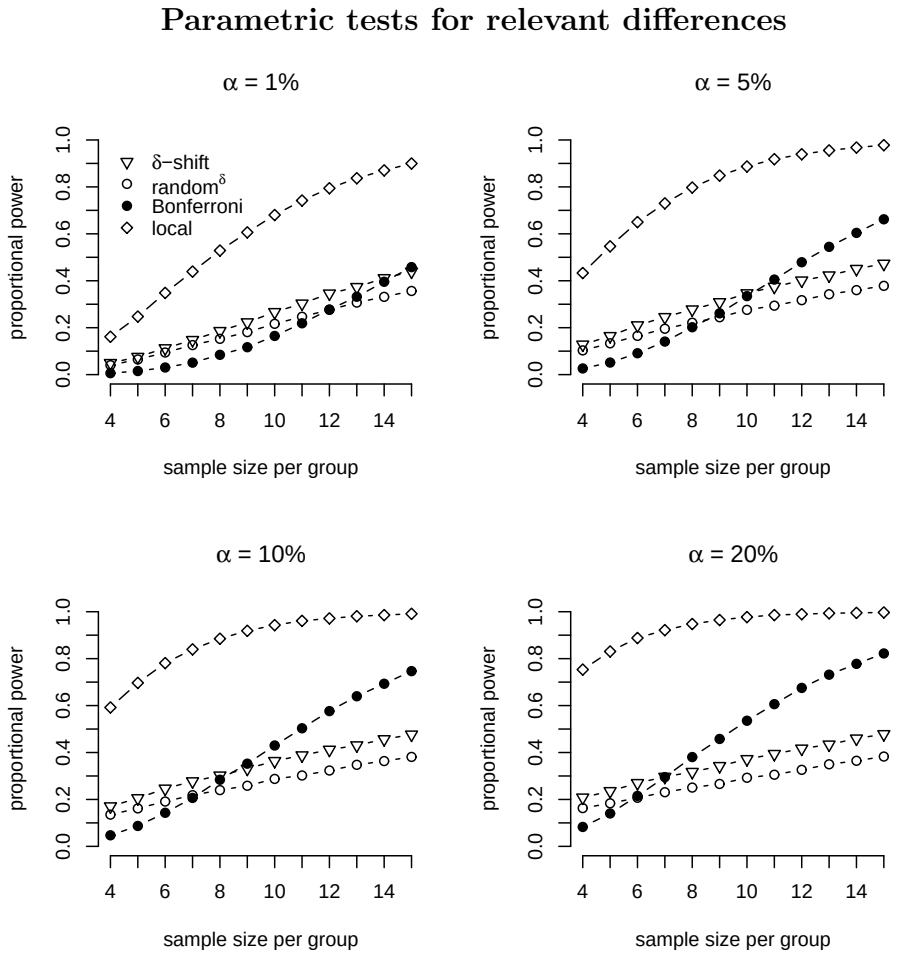

Figure 7.10: Parametric test for rel. diff.: Power for different n_i

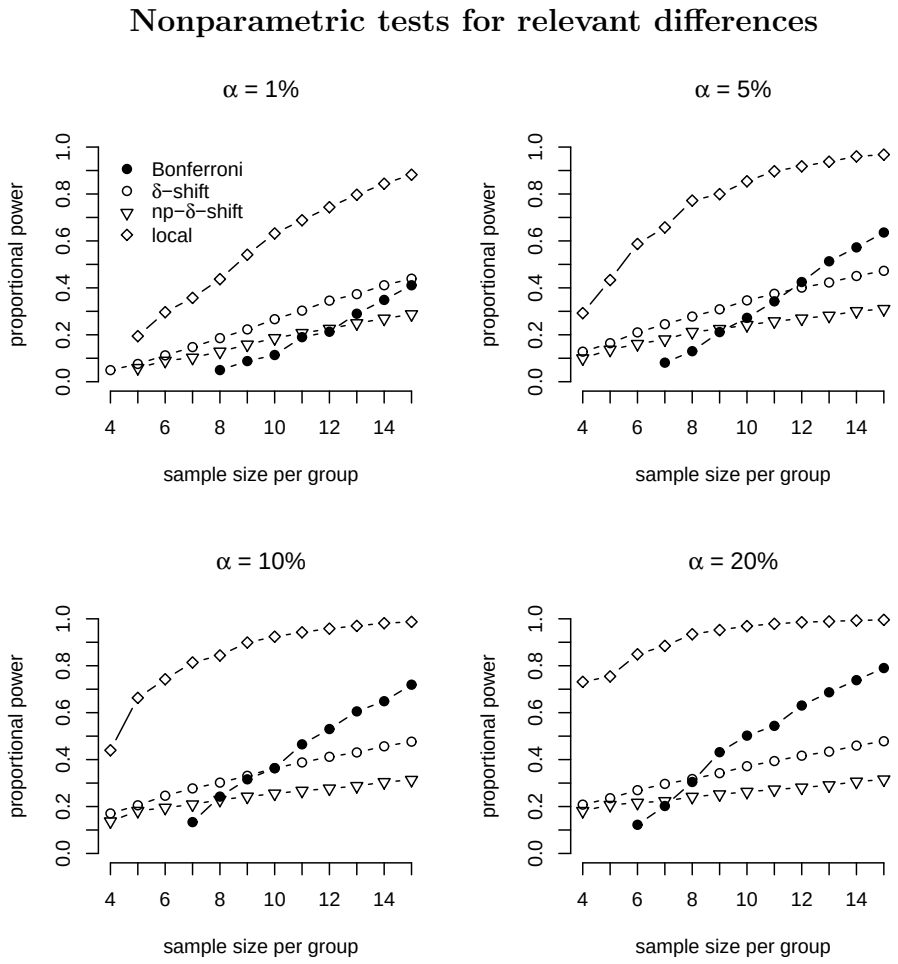

Figure 7.11: Nonparametric test for rel. diff.: Power for different n_i

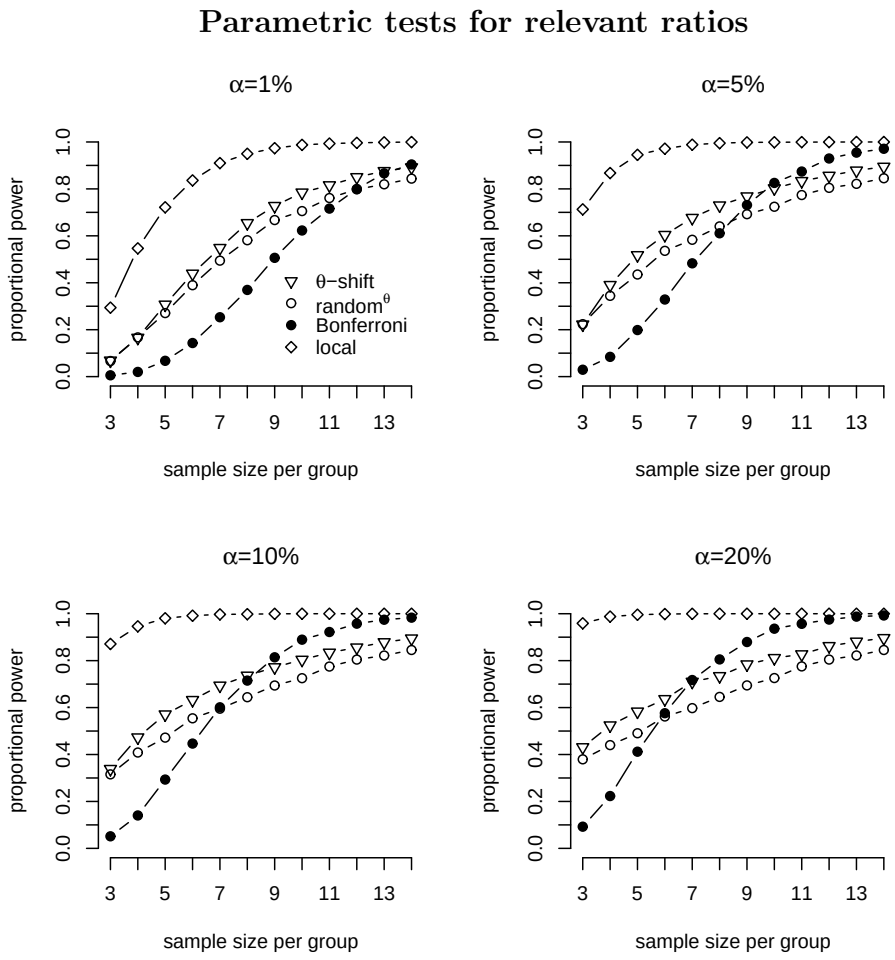

Figure 7.12: Parametric test for rel. ratios: Power for different n_i

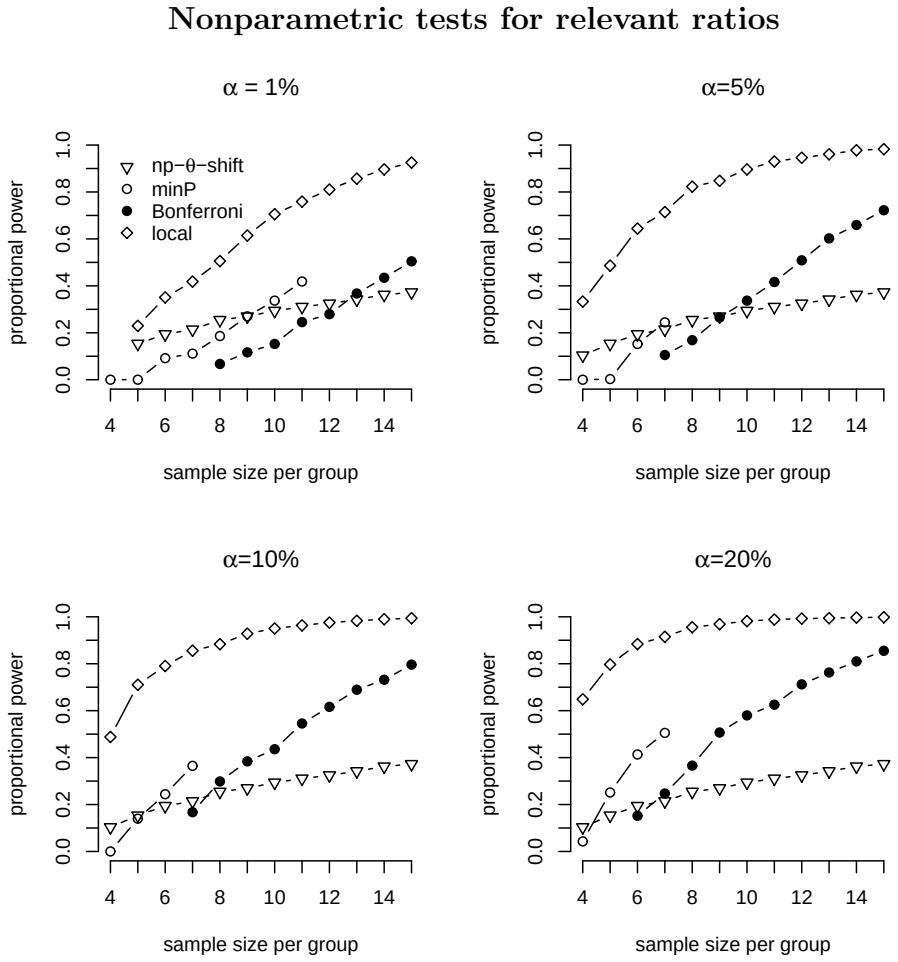

Figure 7.13: Nonparametric test for rel. ratios: Power for different n_i

The four graphics show exactly the same behavior of the power as the tests for point-zero hypotheses. All procedures with a data-driven order of hypotheses are superior to the alternative approaches with multiplicity correction when the sample sizes are small. From the graphics proposals for the application of these procedures can be made:

- The procedures to test for a relevant difference should be used, if $n_i \leq 9$ at $\alpha = 5\%$.
- The parametric procedure to test for a relevant ratio is superior, if $n_i \leq 8$ at $\alpha = 5\%$.
- The nonparametric procedure to test for a relevant ratio is the most powerful approach, if $n_i \leq 6$ at $\alpha = 5\%$.

However these guidelines depend on α, the treatment effect, the variance per endpoint, the amount of variance heterogeneity and correlation between the endpoints, and for the procedures for relevant ratios on the choice of θ_{side}. For larger sample sizes and α the permutation algorithm loses its discreteness and is therefore more powerful. Likewise the Bonferroni correction is more advantageous with a higher number of observations and especially with a large FWER.

7.4 Adapted expected treatment effect and varying sample size

As discussed in chapter 3, it is interesting to visualize the influence of the increasing sample size while the test statistic stays constant. The following figures show the dependency of the power on the sample size and the adapted treatment effect. For the tests for relevant differences and ratios an adapted difference and an adapted ratio are required. Both are computed by use of the relevance-shifted t-test and the Sasabuchi-test. For sake of simplicity the nonparametric tests use the adapted treatment effects of their parametric analog. Both adapted treatment effects are chosen, such that the local test has a power of around 80%.

The adapted difference: In a simulation setting with $-\delta_{lower} = \delta_{upper} = 100$, $n_1 = n_2 = n = 5$, $\sigma_{ij} = 10$ and $\rho_{ijj'} = 0.3$ a local relevance-shifted t-test requires a true difference in means of 120 to achieve a power of around 80%. To compute the adapted expected true difference in means the non-centrality parameter of the relevance-shifted t-test to test against a relevance threshold of δ_{upper} is used:

$$\nu = \frac{\mu_2 - \mu_1 - \delta_{upper}}{\sigma\sqrt{\frac{2}{n}}} \Leftrightarrow \mu_2 - \mu_1 - \delta_{upper} = \nu \cdot \sigma \cdot \sqrt{\frac{2}{n}}. \tag{7.3}$$

With the input of the above proposed simulation parameters the adapted expected difference in means of 120 is:

$$220 - 100 - 100 = 3.162278 \cdot 10 \cdot \sqrt{\frac{2}{5}}. \tag{7.4}$$

For any sample size n the difference becomes:

$$\mu_2 - \mu_1 - \delta_{upper} = \sqrt{3.162278^2} \cdot 10 \cdot \sqrt{\frac{2}{n}} = \sqrt{10} \cdot \sqrt{\frac{200}{n}} = \sqrt{\frac{2000}{n}}. \tag{7.5}$$

By the use of the latter equation the true means for endpoints under H_1 are set to $\mu_{1j} = 100$ and $\mu_{2j} = 200 + \sqrt{2000/n}$.

The adapted ratio: By setting $\theta_{lower}^{-1} = \theta_{upper} = 2$, $n_1 = n_2 = n = 5$, $\sigma_{ij} = 10$ and $\rho_{ijj'} = 0.3$ the local Sasabuchi-test requires a ratio of means of 2.325 to obtain a power of around 80%. The adapted expected true ratio of means is computed by use of the non-centrality parameter of the Sasabuchi-test. The equation to test against a relevance threshold of θ_{upper} is:

$$\nu = \frac{\mu_2 - \mu_1 \cdot \theta_{upper}}{\sigma\sqrt{\frac{1+\theta_{upper}^2}{n}}} \Leftrightarrow \mu_2 - \mu_1 \cdot \theta_{upper} = \nu \cdot \sigma \cdot \sqrt{\frac{1 + \theta_{upper}^2}{n}}. \tag{7.6}$$

By including the parameters as defined above the adapted expected true difference in means of 32.5 is:

$$232.5 - 100 \cdot 2 = 3.25 \cdot 10 \cdot \sqrt{1}. \tag{7.7}$$

For an arbitrary sample size n the true difference becomes:

$$\mu_2 - \mu_1 \cdot \theta_{upper} = \sqrt{3.25^2} \cdot 10 \cdot \sqrt{\frac{5}{n}} = \sqrt{10.625} \cdot \sqrt{\frac{500}{n}} = \sqrt{\frac{5281.25}{n}}. \tag{7.8}$$

Hence the expected values for endpoints under H_1 are set to $\mu_{1j} = 100$ and $\mu_{2j} = 200 + \sqrt{5281.25/n}$. This is the adapted treatment effect for the parametric tests. If this adapted effect were to be used for the nonparametric tests on relevant ratios, then the resulting power curves would be hardly comparable because of the lack of power. As an approximation the non-centrality parameter from the Sasabuchi-test is used as well, but with the input of a true ratio of means of 2.475. Then the expected values are set to $\mu_{1j} = 100$ and $\mu_{2j} = 200 + \sqrt{11281.25/n}$.

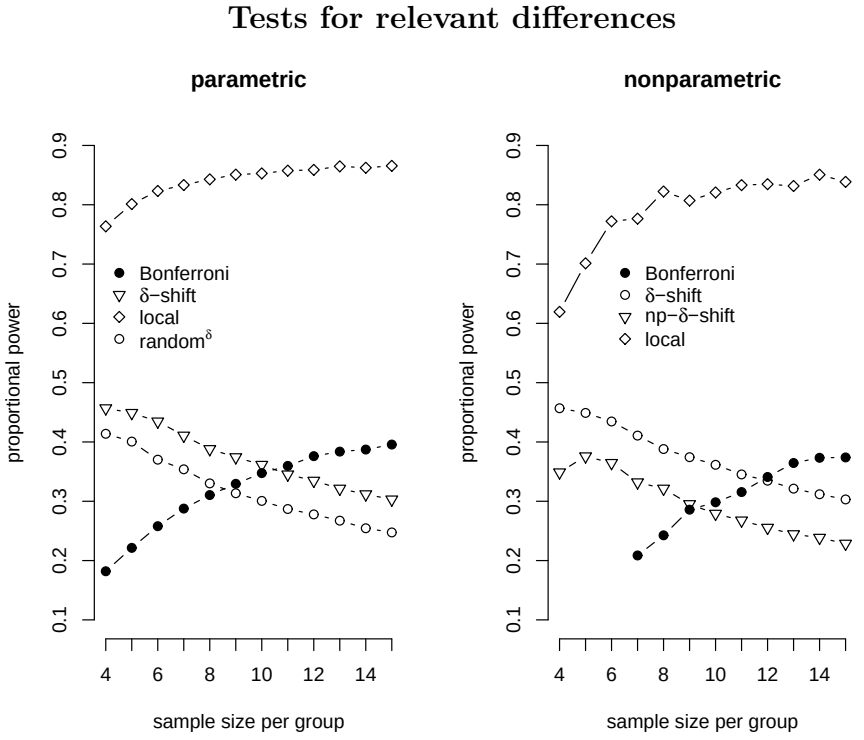

Figure 7.14: Test for rel. diff.: Power for different n_i with adapted difference

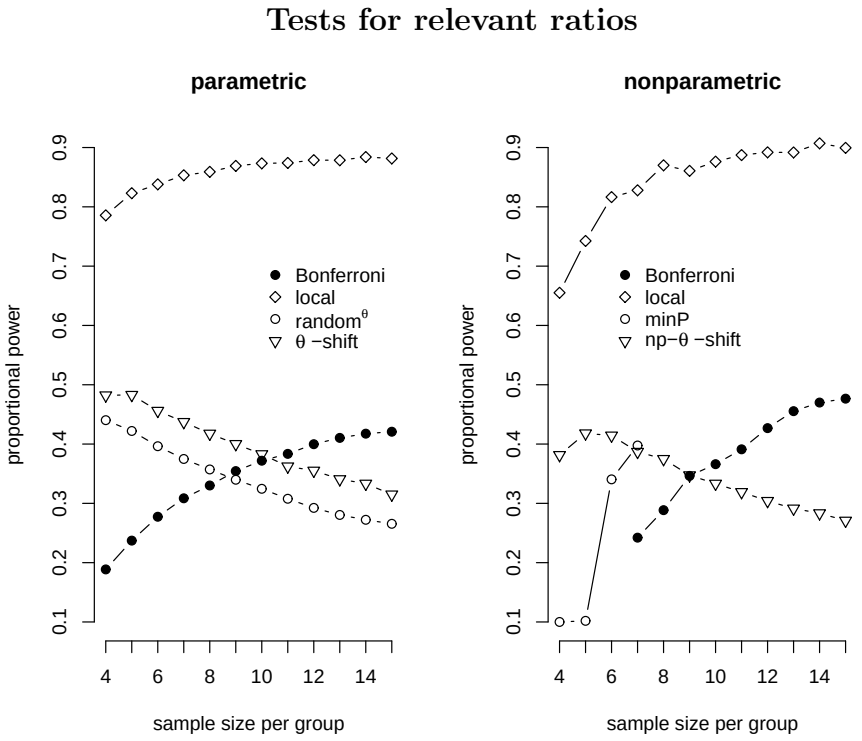

Figure 7.15: Test for rel. ratios: Power for different n_i with adapted ratios

These four graphics confirm the results of the former graphics: the procedures with a data-driven order of hypotheses achieve a higher power compared to the Bonferroni adjustment and the modified permutation algorithm when the sample sizes are small. Both the α-adjustment according to Bonferroni and the permutation algorithm benefit from the increase of the sample size - as for both the increase of the degrees of freedom has a direct influence on the p-values and - for the nonparametric tests - the discreteness of the exact rank sum test (concerning Bonferroni) and of the numbers of permutations (resampling algorithm) vanishes with increasing sample sizes. And as already discussed in chapter 3 the procedures with a data-driven order of hypotheses do not benefit from the increase in sample size, because the selector statistics of the significant endpoints decrease. They depend on the treatment effect and not on the sample size.

7.5 Simulations with increasing disturbance

Compared to the procedures with a data-driven order of point-zero hypotheses the new procedures show exactly the same lack of power when the variances among the endpoints are heterogeneous. The following figures depict this lack of power for all four types of relevance-shifted procedures. As in chapter 3 the true standard deviation per endpoint is computed as $\sigma_{ij} = 10 + u \cdot d$ ($u \sim U(-5, 5)$), where u takes values from 0 to 2 in steps of 0.1 units. The correlation among the endpoints per group is set to 0.3. And the remaining parameters are

procedure	treatment effect	relevance thresholds	sample size per group
tests on difference	$\tau_\delta^{H_1} = 220$	$-\delta_{lower} = \delta_{upper} = 200$	7
parametric test on ratio	$\tau_\theta^{H_1} = 3.6$	$\theta_{lower}^{-1} = \theta_{upper} = 3$	5
nonparametric test on ratio	$\kappa^{H_1} = 3.6$	$\theta_{lower}^{-1} = \theta_{upper} = 3$	7

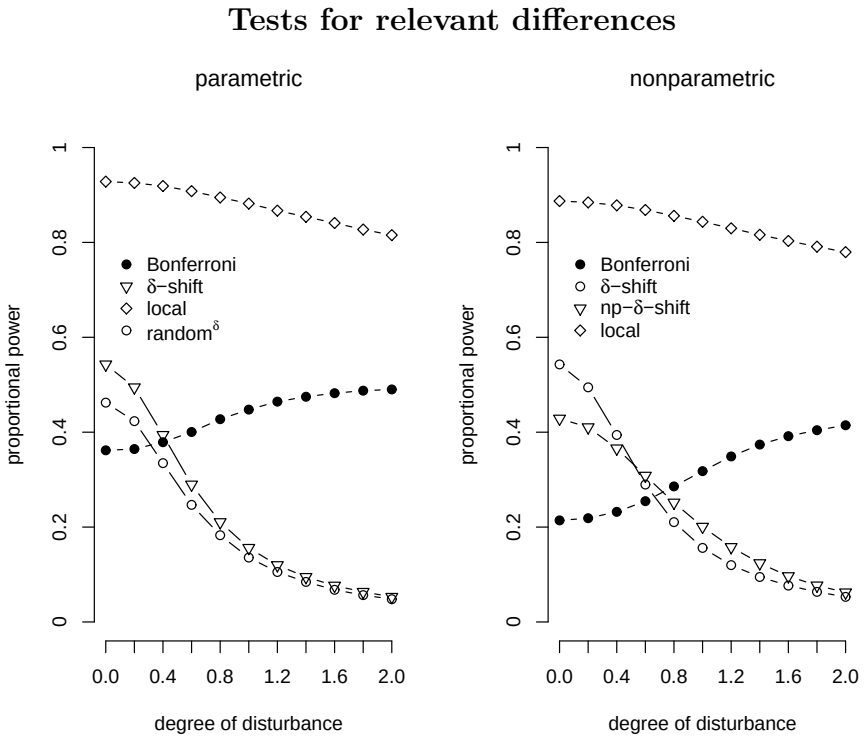

Figure 7.16: Test for rel. diff.: Power for increasing disturbance

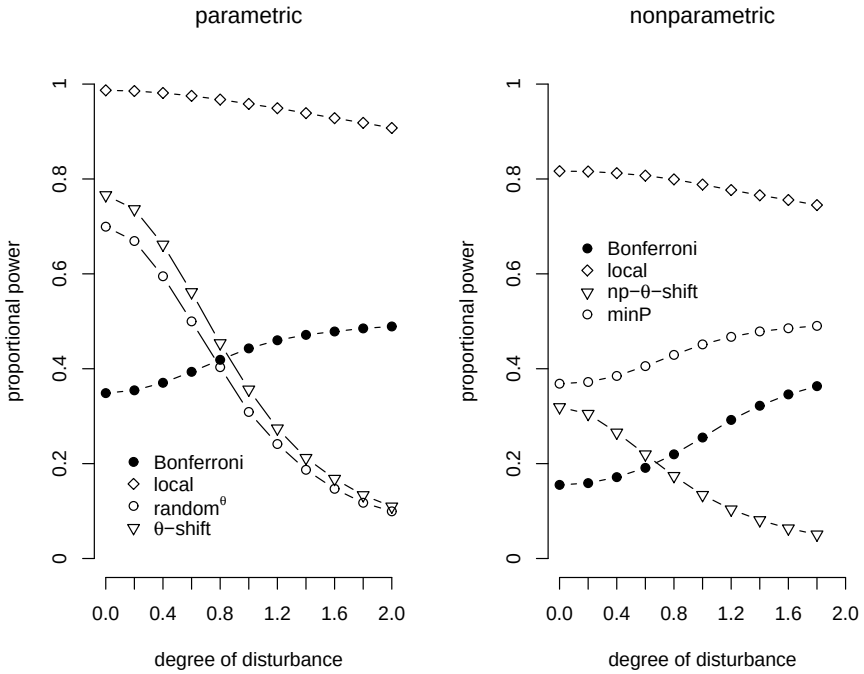

Figure 7.17: Test for rel. ratios: Power for increasing disturbance

The increasing variance heterogeneity among the endpoints has the same influence on all relevance-shifted procedures compared to the ones for point-zero hypotheses discussed in chapter 3. The modified permutation algorithm benefits in this setting from the increasing heterogeneity as well.

7.6 Simulations with varying mean levels

The next graphics show the proportional power for varying treatment effects and relevance thresholds. Basically they are the same graphics as 7.2 for $\rho_{ijj'} = 0.3$, but the simulation setting is more realistic to microarray data. Attoor *et al.* (2004) proposed a simulation setting with different expression-intensity means of genes. While so far all endpoints had a basic mean of 100, an exponential distributed level is proposed. It is characterized as $I_j \sim \phi + \text{Exp}(\pi)$, where ϕ is the minimal detectable expression level above the background noise and $\text{Exp}(\pi)$ denotes the exponential distribution with expectation $1/\pi$ and variance $1/\pi^2$. Following the proposal of Attoor *et al.* (2004) the parameters are set to $\pi = 100$ and $\phi = 2000$.

As well as the mean level of the endpoints differ, the variances are changed as well. For each endpoint and group the data is generated as $x_{ijk} \sim \text{N}(I_j, \gamma I_j)$, with the coefficient of variation γ. Following Attoor *et al.* (2004) γ is set to 0.1 (proposed: 0.05-0.15). The sample size per group is set to 5 and $\rho_{ijj'} = 0.3$.

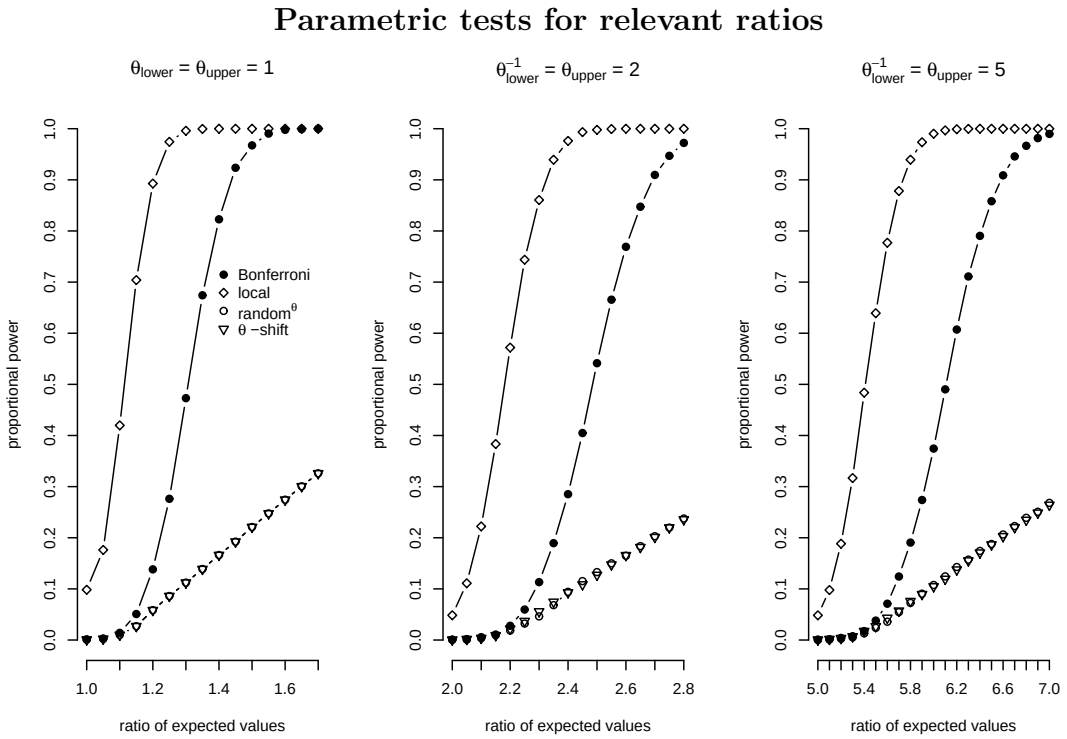

Figure 7.18: Param. test for rel. ratios: Power with data model of Attoor *et al.* (2004)

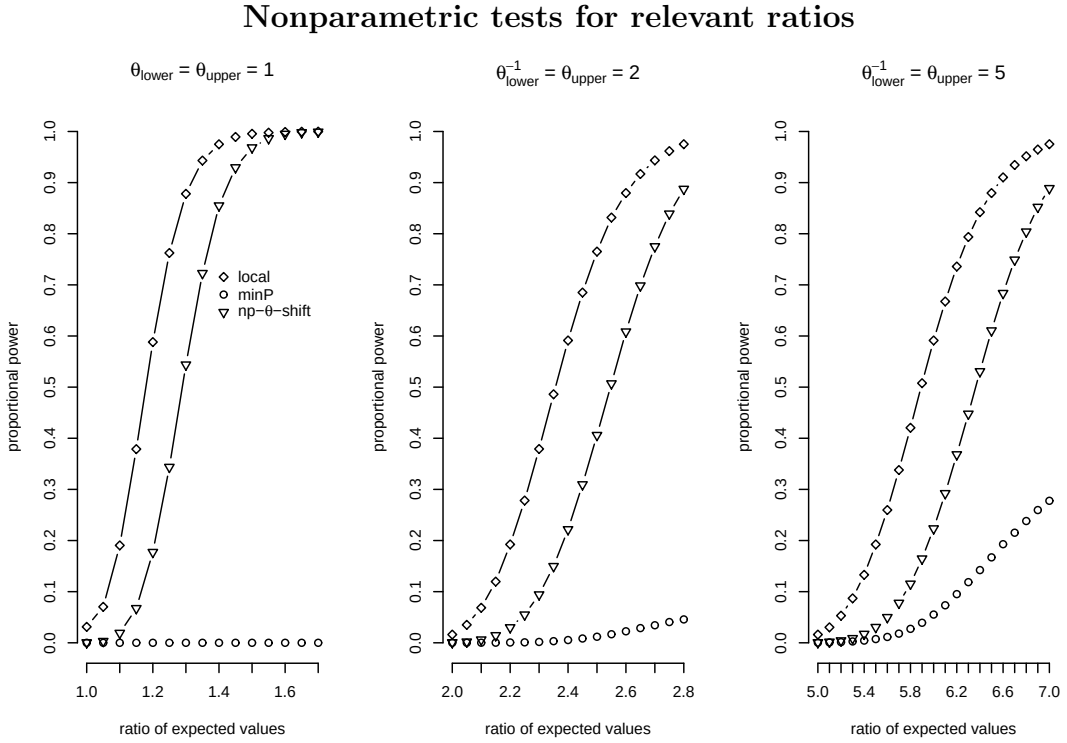

Figure 7.19: Non-param. test for rel. ratios: Power with data model of Attoor *et al.* (2004)

The parametric tests with a data-driven order of hypothesis to test for a relevant ratio lack power compared to the α-adjustment of Bonferroni. This decrease in power is expected, as due to the different levels of means and the constant coefficient of variation a variance heterogeneity among the endpoints appears.
This variance-heterogeneity vanishes with the logarithmic transformation included in the nonparametric test for relevant ratios. Hence the nonparametric test with a data-driven order of hypotheses is superior compared to Bonferroni's adjustment. However this superiority depends on the constant coefficient of variation. If the coefficient varies among the endpoints, the power of the nonparametric test using the selector statistic decreases.

7.7 Simulations with non-normal distributed data

For all former simulations of the power multivariate normal distributed data is used. However as in this work nonparametric procedures are discussed and furthermore microarray data tends to a skewed (or unknown) data distribution, the power is observed for non-normal distributed data as well. To construct univariate non-normal distributed samples with a priori selected expected value, standard deviation, skewness and kurtosis a polynomial data transformation proposed by Fleishman (1978) is used. If $X \sim N(0,1)$ is a random variate and transformed by $Y = a + bX + cX^2 + dX^3$, then Y has a distribution depending on the constants a, b, c and d with skewness $\gamma_1 = \frac{E((Y-\mu)^3)}{\sigma^3(Y)}$ and kurtosis $\gamma_2 = \frac{E((Y-\mu)^4)}{\sigma^4(Y)}$. In his article Fleishman (1978) lists the corresponding values for the constants for $-0.25 \leq \gamma_1 \leq 1.75$ and $-1 \leq \gamma_2 \leq 3.75$. Nürnberg (1982) extends the listing and gives constants for combinations of $0 \leq \gamma_1 \leq 2$ and $0 \leq \gamma_2 \leq 7$. To generate multiple variables with a specific variance and correlation structure this polynomial transformation is used as well. Let $\mathbf{Y}$ denote a $n \times m$ matrix with m univariate variables following a distribution with γ_1 and γ_2. Then the variables of the $n \times m$ matrix $\mathbf{Z}$ are non-normal multivariate distributed with variance-covariance matrix $\mathbf{\Sigma}$, where $\mathbf{Z}$ is given by the matrix multiplication of $\mathbf{Y}$ and $\mathbf{R}$:

$$\mathbf{Z} = \mathbf{YR}. \tag{7.9}$$

The $m \times m$ matrix $\mathbf{R}$ is given by the Cholesky decomposition, which factorizes $\mathbf{\Sigma}$ into the upper triangular matrix $\mathbf{R}$, such that $\mathbf{\Sigma} = \mathbf{R}'\mathbf{R}$.

In the following table the empirical estimates of a-priori selected parameters for two examples are presented. For both examples five endpoints are created with 1,000 observations each. Due to the estimation of the kurtosis such a high sample size is required. In the first example the correlation among the endpoints is set to 0 and in the second it is $\rho_{ijj'} = 0.5$.

example	parameter	estimates of endpoint				
		1	2	3	4	5
first	$\mu_j = 0$	0.010	0.019	-0.020	-0.002	-0.008
	$\sigma_j = 1$	0.989	1.042	1.023	1.056	1.016
	$\gamma_{1,j} = 0.5$	0.502	0.397	0.602	0.729	0.399
	$\gamma_{2,j} = 1$	1.017	1.001	1.732	1.565	0.630
second	$\mu_j = 0$	0.143	0.336	0.384	0.470	0.529
	$\sigma_j = 1.5$	1.486	1.534	1.528	1.570	1.512
	$\gamma_{1,j} = 1.5$	1.478	1.032	1.176	1.092	0.915
	$\gamma_{2,j} = 3$	3.262	1.448	2.383	1.498	1.060

And the estimated correlation matrices $\hat{\rho}$ are:

$$\hat{\rho} = \begin{pmatrix} 1 & 0.035 & 0.033 & -0.002 & 0.012 \\ 0.035 & 1 & -0.004 & -0.041 & 0.007 \\ 0.033 & -0.005 & 1 & 0.010 & 0.013 \\ -0.002 & -0.041 & 0.010 & 1 & -0.039 \\ 0.012 & 0.007 & 0.013 & -0.039 & 1 \end{pmatrix} \tag{7.10}$$

and

$$\hat{\rho} = \begin{pmatrix} 1 & 0.497 & 0.512 & 0.488 & 0.513 \\ 0.497 & 1 & 0.494 & 0.472 & 0.500 \\ 0.512 & 0.494 & 1 & 0.473 & 0.508 \\ 0.488 & 0.472 & 0.473 & 1 & 0.477 \\ 0.513 & 0.500 & 0.508 & 0.477 & 1 \end{pmatrix}. \tag{7.11}$$

The final graphics in this chapter are initially the same as in section 7.2. However instead of using Gaussian distributed data, the observation vectors follow a skewed distribution with a skewness of 2 and a kurtosis of 7 and the correlation among the endpoints is set to 0.3. To show the superiority of the np-θ-shift compared to a parametric analog, power results of the parametric procedure with a data-driven order of relevance-shifted hypotheses including the relevance-shifted t-test on difference are show as well (abbreviated as 'parametric'). This procedure equals the δ-shift method applied on the logarithmized data. However, to compare a parametric procedure with the np-θ-shift method, as described in the beginning of chapter 6, the assumptions on the data and hence the simulation settings are different. Before the results of the power are shown, a histogram is plotted for illustration of the distribution of 1,000 random numbers taken from N(0,1) and transformed by the Fleishman algorithm, such that $\gamma_1 = 2$ and $\gamma_2 = 7$. For a better comparability the results of the parametric and the nonparametric tests on difference are plotted in one figure. To avoid confusion resulting of too many curves, results of the local tests are omitted.

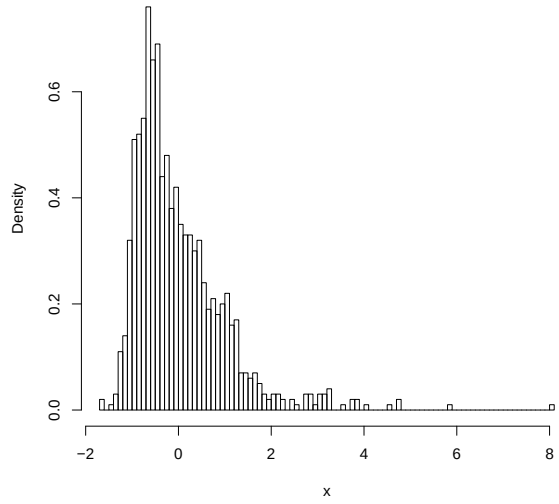

Figure 7.20: Histogram of 1000 Fleishman-transformed random numbers ($\gamma_1 = 2$, $\gamma_2 = 7$)

And in the following graphics the power results of the procedures are shown.

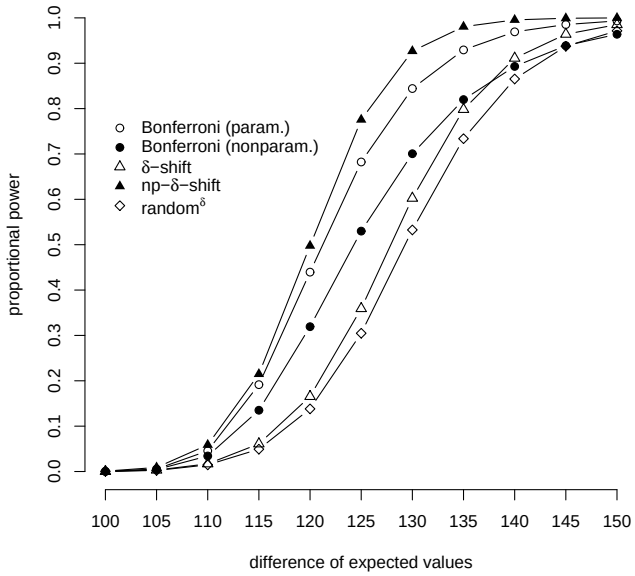

Figure 7.21: Test for rel. diff.: Samples taken from a non-normal distribution

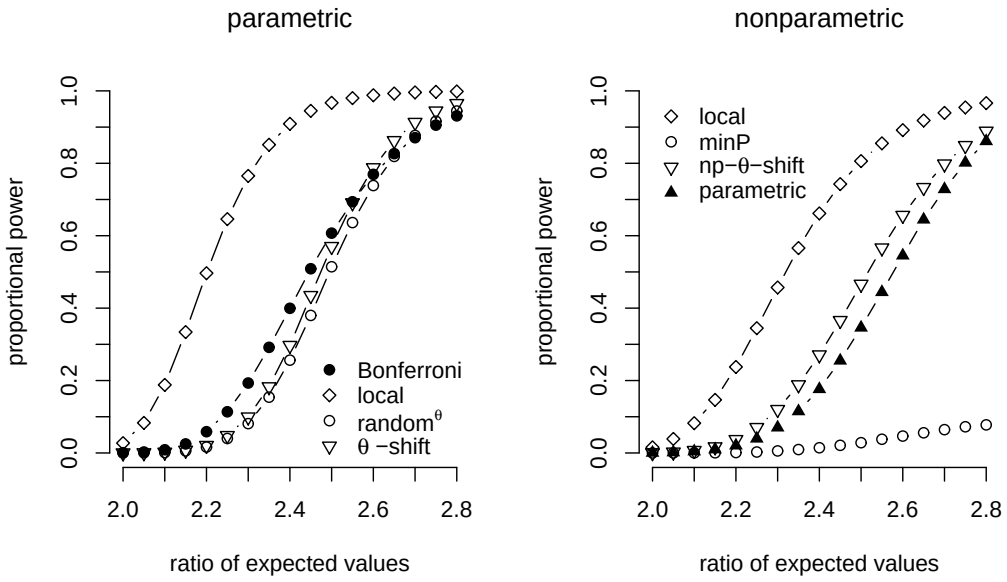

Figure 7.22: Test for rel. ratios: Samples taken from a non-normal distribution

The graphic corresponding to the tests on difference show that the np-δ-shift method is more powerful as the δ-shift and the random procedure when the data is not Gaussian distributed. The violation of the normal assumption biases both the test statistic and the selector statistic of the parametric versions. An oppositional behavior of the power can be seen for the parametric and the nonparametric tests with the Bonferroni correction: the parametric version achieves more significant results. The reason for this result is the discreteness of the rank sum test. As already described in the univariate example in the introduction (figures 1.1 and 1.2) the number of permutations of the exact rank sum test is dependent on the sample size. With small sample sizes this number is limited and due to the Bonferroni correction this effect becomes even more severe.

With deviations from the Gaussian distribution the parametric tests on ratio lose power. This is especially true for the two procedures with a data-driven order of hypotheses. The reason for this breakdown in power is, that not only is the test statistic influenced by the skewed distributed data, but the selector statistic as well. Hence the parametric Bonferroni adjustment is more powerful in this setting. It is only excelled by the local tests, which do not control the FWER.

The presence of non-normal distributed samples has no impact on the ranking of the nonparametric tests on ratio in terms of power. As expected the np-θ-shift procedure is more powerful than the modified permutation algorithm, because the sample sizes are very small. Furthermore it can be seen, that the np-θ-shift method is more powerful as the parametric analog, as the non-gaussian distributed data bias the selector statistics of the latter procedure. And the local tests achieve the highest power, because they do not consider the multiplicity. Clearly, the non-parametric procedures with a data-driven order of relevance-shifted hypotheses are advantages compared to the parametric methods.

Chapter 8

Final remarks and conclusions

Various problems occur in the analysis of microarray data. While the data is high-dimensional, the sample size per treatment group and endpoint is small. By testing thousands of endpoints simultaneously a multiplicity correction has to be applied. As sample sizes are small, multiple testing procedures lack power. Due to a possible discreteness, this problem even increases if the nonparametric two-sample tests to analyze the individual endpoints are used. Various methods to overcome these problems and test for differences between the treatment groups among the genes exist. However in many articles the authors interest is to find not only a general difference, but a ratio in expression intensities exceeding a specific k-fold. Here, instead of testing the point-zero hypothesis, a relevance-shifted test is appropriate. The goal of this thesis is to develop a nonparametric testing procedure, which analyzes such relevance-shifted hypothesis in terms of the ratio. Furthermore this procedure has to use the technique of the data-driven order of hypotheses as multiplicity correction, because this class of tests has been proven to be superior to the other FWER-controlling methods for the analysis of microarray data when the sample sizes are extremely small.

In total the thesis includes five new types of procedures; both parametric and nonparametric methods including relevance thresholds in terms of the difference or the ratio are presented. In addition a nonparametric relevance-shifted modification of the permutation

algorithm according to Westfall and Young (1993) is presented, which can be used to test for relevant ratios.
In general the new procedures can be appropriate for the analysis of high-dimensional data. The procedures to test for a relevant difference are however not appropriate to use for microarray data, as long as the data is not logarithmized. Without this transformation the overall distribution of the expression intensities is skewed, and the measurements range between 0 and several thousand. Hence the selection of a relevant difference is impossible.

In addition the variances among the endpoints are highly heterogeneous. With the logarithmic transformation the multiplicative model switches to an additive one. If the tests on difference are used on the logarithmized data and the relevance thresholds are adapted, then these tests become tests on relevant ratios. In this case these tests are appropriate for the analysis of microarray data. However the changes in the data assumptions have to be regarded. When used on the logarithmized data, the parametric test on difference assumes lognormal distributed samples on the original scale and as the nonparametric test on difference becomes exactly the nonparametric test on ratio proposed in this thesis, it achieves the same assumptions as this test.
Whereas the tests on difference can be used on the logarithmized microarray data, the parametric test on ratio with a data-driven order of hypotheses can not be recommended for the analysis of this type of data. If the data is not logarithmized then the procedure lacks power because of the variance heterogeneity among the endpoints. And by use of this transformation on the data the parametric procedures seeks for a relevant k-fold on the logarithmic scale, which is usually not the aim of a microarray experiment.
Recommended for the analysis of microarray data are the the nonparametric test on relevant ratios with a data-driven order of hypothesis and the modified permutation algorithm, because both do not have as stringent data assumptions as the parametric test. And as the procedure with a data-driven order of hypotheses incorporates the logarithmic transformation, the variance heterogeneity among the endpoints is highly decreased.

Detailed power studies show, that the power of the new procedures has a similar behavior compared to standard multiple testing methods as the methods such for point-zero hypotheses proposed by KROPF AND LÄUTER (2002) and KROPF *et al.* (2004). All relevance-shifted procedures with a data-driven order of hypotheses are superior to the α-adjustment of Bonferroni and the nonparametric test on ratio is even more powerful than the modified permutation algorithm if certain data conditions hold. Due to the high number of endpoints and the small sample sizes the Bonferroni correction is very conservative as well. The loss in power is even worse if the nonparametric rank sum test is used as the two-sample test, because due to the small sample sizes the number of permutations for the exact test is limited and therefore this test becomes discrete. The more powerful modified permutation algorithm suffers from the discreteness resulting from the small number of observations as well.
Based on the simulation results the nonparametric testing procedure to test for a ratio can be recommended for experiments with samples of sizes per gene and group less than 7. For larger sample sizes the modified permutation algorithm achieves a higher power. All other relevance-shifted procedures with a data-driven order of hypotheses are superior to the Bonferroni correction if the sample sizes are less than 10. This recommendation depends however on various factors, such as the significance level, the variance among the endpoints and the treatment groups, the number of endpoints and the correlation among the genes. For example in the simulation study only 50 endpoints are observed. It can be expected that the Bonferroni correction needs a larger sample size to be superior to the new procedures when the number of endpoints is higher. On the other hand, this simple multiple testing method can be recommended, if the sample sizes are not too small and the significance level is higher. In the simulation results the Bonferroni adjustment achieved a higher power compared to the new procedures if $\alpha = 0.2$ and the sample size is 7 or more.

However, to be superior compared to these alternative methods, the procedures with a data-driven order of hypotheses require moderately homogenous variances among the endpoints. As denoted above for microarray data, this can be achieved with the logarithmic

transformation, because moderately large expression intensities have an approximately constant coefficient of variation. The stabilization of the variance could even be improved with, for example, the algorithm of HUBER *et al.* (2002), but as this method uses the arsin-transformation, a different type of nonparametric two-sample test as used here would be required. An interesting further research topic is the variance-stabilization algorithm of DURBIN *et al.* (2002), who applies the natural logarithm on an additive data transformation to achieve a constant variance among the endpoints.

An alternative method is to implement the relevance-shifted tests in a weighted procedure with a data-driven order of hypotheses. Such an approach is proposed by WESTFALL *et al.* (2004), where an a-priori defined weight $\eta \geq 0$ is included in the computation of the selector statistics and the p-values. In the extreme cases of $\eta = 0$ the weighted procedure equals the α-adjustment according to Bonferroni-Holm (see HOLM (1979) for details). And if $\eta = \infty$ the procedure converges to the non-weighted procedure with a data-driven order of hypotheses as proposed for example in chapter 3. Hence if the weight is set to a value close to 0, then the procedure is more robust against deviations of the variance heterogeneity among the endpoints. The choice of the optimal weight η and a further testing procedure are proposed by KROPF AND HOTHORN (2004).

Other extensions of the procedures with a data-driven order of hypotheses exist, and an embedding of the relevance-shifted tests in such a method could be worthwhile. For example, if a procedure proposed in this work is applied, it may abort prematurely because of a non-significant endpoint with a high selector; although the following endpoints in the order could have been significant. To overcome the problem of a premature ending of the procedure, HOMMEL AND KROPF (2005) proposed to test each hypothesis against α/s, where s defines an a-priori chosen integer. The value of $s-1$ is the number of accepted null hypotheses which can be ignored before the procedure has to be stopped. Hence if $s = 1$ the algorithm reduces to the original procedures with a data-driven order of hypotheses.

An important research topic may be the derivation of a mathematical proof, as for all proposed procedures the control of the FWER is shown empirically with simulations of the error rate only.

Bibliography

Abruzzo, L.V., J. Wang, M. Kapoor, L.J. Medeiros, M.J. Keating, W.E. Highsmith, L.L. Barron, C.C. Cromwell and K.R. Coombes (2005): Biological validation of differentially expressed genes in chronic lymphocytic leukemia identified by applying multiple statistical methods to oligonucleotide microarrays. *Journal of Molecular Diagnostics*, 7 (3), 337–345.

Agresti, A. and B.A. Coull (1998): Approximate is better than 'exact' for interval estimation of binominal proportions. *The American Statistician*, 52 (2), 119–126.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and S.J. Korsmeyer (2002): MLL translocations specifiy a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30, 41–47.

Attoor, S., Dougherty, E.R., Chen, Y., Bittner, M.L. and J.M. Trent (2004): Which is better for cDNA-microarray-based classification: ratios or direct intensities. *Bioinformatics*, 20 (16), 2513–2520.

Bauer, D.F. (1972): Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, 67 (339), 687–690.

Beasley, T.M., Page, G.P. and J.P.L. Brand (2004): Chebyshev's inequality for nonparametric testing with small N and α in microarray research. *Appl. Statist.*, 53 (1), 95–108.

Benjamini, Y. and Y. Hochberg (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, series B, 57 (1), 289–300.

Bolstad, B.M., Irizarry, R.A., Åstrand, M. and T.P. Speed (2003): A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19 (2), 185–193.

Büning, H. and G. Trenkler (1994): *Nichtparametrische statistische Methoden.* Walter de Gruyter, Berlin.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P. and E.M. Rubin (2001): Microarray gene expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 10: 2022–2029.

DeRisi, J.L., Iyer, V.R. and P.O. Brown (1997): Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680–686.

Dudoit, S., Yang, Y.H., Callow, M.J. and and T.P. Speed (2002): Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12, 111–140.

Dudoit, S., Shaffer, J.P. and J.C. Boldrick (2003): Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18 (1), 71–103.

Durbin, B.P., Hardin, J.S., Hawkins, D.M. and D.M. Rocke (2002): A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1, S105–S110.

Eszlinger, M., Krohn, K., Frenzel, R., Kropf, S., Tönjes, A. and R. Paschke (2004): Gene expression analysis reveals evidence for inactivation of the TGF-β signaling cascade in autonomously functioning thyroid nodules. *Oncogene*, 23, 795–804.

Fleishman, A.I. (1978): A method for simulating non-normal distributions. *Psychometrika*, 43 (4): 521–532.

Ge, Y., Dudoit, S. and T.P. Speed (2003): Resampling-based multiple testing for microarray data hypothesis. *Test*, 12 (1): 1-44 (with discussions on 44–77).

Geller, S.C., Gregg, J.P., Hagerman, P. and D.M. Rocke (2003): Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19 (14), 1817–1823.

Giegé, P., Sweetlove, L.J., Cognat, V. and C.J. Leaver (2005): Coordination of nuclear and mitochondrial genome expression during mitochondrial biogenesis in Arabidopsis. *The Plant Cell*, 17, 1497–1512.

Giles, P.J. and D. Kipling (2003): Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19 (17), 2254–2262.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and E.S. Lander (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.

Halitschke, R., Gase, K., Hui, D., Schmidt, D.D. and I.T. Baldwin (2003): Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata.* VI. Microarray analysis reveals that most herbivore-specific transcriptional changes are mediated by fatty acid-amino acid conjugates. *Plant Physiology*, 131, 1894–1902.

Hartung, J. and B. Elpelt (1999): *Multivariate Statistik*, 6^{th} edition, Oldenbourg Wissenschaftsverlag GmbH, München.

Hauschke, D. (1999): *Biometrische Methoden zur Planung und Auswertung von Sicherheitsstudien*, Habilitationsschrift, Universität Dortmund.

Heidel, A.J. and I.T. Baldwin (2004): Microarray analysis of salicylic acid- and jasmonic adcid-signalling in responses of *Nicotiana attenuata* to attack by insects from multiple feeding guilds. *Plant, Cell and Environment*, 27 (11), 1362–1373.

Hochberg, Y. and A.C. Tamhane (1987): *Multiple Comparison Procedures*, John Wiley & Sons, Inc., New York.

Hodges, J.L., Jr., and E.L. Lehmann (1963): Estimates of location based on rank tests. *Annals of Mathematical Statistics*, 34: 598–611.

Hollander, M. and D.A. Wolfe (1999): *Nonparametric Statistical Methods.* John Wiley & Sons, New York.

Holm, S. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Hommel, G. and T. Hoffman (1988): Controlled uncertainty, in *Multiple Hypotheses Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.), Springer, Heidelberg, 154–161.

Hommel, G. and S. Kropf (2005): Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal*, 47 (4), 554–562.

Hothorn, T. and K. Hornik (2002): Exact Nonparametric Inference in R, in *Proceedings in Computational Statistics: COMPSTAT2002* (W. Härdle, B. Rönz, eds.), Physica-Verlag Heidelberg, 355–360.

Hothorn, T. and K. Hornik (2004): *exactRankTests: Exact Distributions for Rank and Permutation Tests.* R package version 0.8–9.

Hothorn, T. and U. Munzel (2002): Exact nonparametric confidence intervals for the ratio of medians. *Technical report,* Universität Erlangen-Nürnberg.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and M. Vingron (2002): Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 1 (1), 1–9.

Hunter, L., Taylor, R.C., Leach, S.M. and R. Simon (2001): GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, 17 Suppl 1, S115–S122.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barcklay, Y.D., Antonellis, K.J., Scherf, U., Speed, T. (2003): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4 (2), 249–264.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, Jr.,J., Boguski, M.S. , Lashkari, D., Shalon, D., Botstein, D. and P.O. Brown (1999): The transcriptional program in the response of human fibroblasts to serum. *Sciene*, 283, 83–87.

Kieser, M. and D. Hauschke (1999): Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics*, 9 (4), 641–650.

Kooperberg, C., Aragaki, A., Strand, A.D. and J.M. Olson (2005): Significance testing for small microarray experiments. *Statistics in Medicine*, 24, 2281–2298.

Kropf, S. (2000): Hochdimensionale multivariate Verfahren in der medizinischen Statistik. Shaker Verlag GmBH, Aachen.

Kropf, S. and L.A. Hothorn (2004): Multiple test procedures with multiple weights. *Proceedings in Computational Statistics: COMPSTAT2004* (Antoch, J., ed.), 16th symposium held in Prague, Czech Republic, 2004. Heidelberg, Physica-Verlag, 1353–1360.

Kropf, S. and J. Läuter (2002): Multiple Tests for Different Sets of Variables Using a Data-Driven Ordering of Hypotheses, with an Application to Gene Expression Data. *Biometrical Journal*, 44, 789–800.

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K. and R. Paschke (2004): Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference*, 125 (1), 31–48.

Kunz, M., Ibrahim, S.M., Koczan, D., Scheid, S., Thiesen, H.J. and G. Gross (2004): DNA microarray technology and its applications in dermatology. *Experimental Dermatology*, 13, 593–606.

Läuter, J., Glimm, E. and S. Kropf (1996): New multivariate tests for data with an inherent structure. *Biometrical Journal* 38, 5–23.

Läuter, J. (1996): Exact t and F Tests for Analyszing Studies with Multiple Endpoints. *Biometrics*, 52, 964–970.

Lehmann, E.L. and J.P. Romano (2005): Generalizations of the familiywise error rate. *The Annals of Statistic*, 33 (3): 1138–1154.

Li, C. and W.H. Wong (2001): Model-based analysis of oligonucleotide arrays: model validation, design issue and standard error application. *Genome Biology*, 2 (8): research0032.1–0032.11.

Lindenmayer, D.B., Viggers, K.L., Cunningham, R.B. and C.F. Donnelly (1995): Morphological variation among columns of the mountain brushtail possum, *Trichosurus caninus Ogilby* (Phalangeridae: Marsupiala). *Australian Journal of Zoology*, 43, 449–458.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and E.L. Brown (1996): Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14, 1675–1680.

Maindonald, J. and W.J. Braun (2004): *DAAG: Data Analysis And Graphics*. R package version 0.46, http://www.stats.uwo.ca/DAAG.

Miller, R.A., Galecki, A. and R.J. Shmookler-Reis (2001): Interpretation, Design, and Analysis of Gene Array Expression Experiments. *Journal of Gerontology: Biological Sciences*, **56a**, no. 2, B52–B57.

Neuhäuser, M. and R. Senske (2004): The Baumgartner-Weiß-Schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 20 (18), 3553–3564.

Nürnberg, G. (1982): Beiträge zur Versuchsplanung für die Schätzung von Varianzkomponenten und Robustheitsuntersuchungen zum Vergleich zweier Varianzen. *Probleme der angewandten Statistik.* (6), Dummerstorf–Rostock.

Parzen, E. (2004): Quantile probability and statistical data modeling. *Statistical Science*, 19 (4), 652–662.

Pflüger, R. and T. Hothorn (2002): Assessing Equivalence Tests with Respect to their Expected p-Value. *Biometrical Journal,* 44 (8), 1015–1027.

Piegorsch, W.W. (2004): Sample sizes for improved binominal confidence intervals. *Computational Statistics & Data Analysis*, 46, 309–316.

Polacek, D.C., Passerini, A.G., Shi, C., Francesco, N.M., Manduchi, E., Grant, G.R., Powell, S., Bischof, H., Winkler, H., Stoeckert, C.J. Jr. and P.F. Davies (2003): Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. **Physiol Genomics**, 13, 147–156.

Pollard, K.S., Ge, Y. and S. Dudoit: *multtest: Resampling-based multiple hypothesis testing*. R package version 1.5.2.

R Development Core Team (2004): *R: A language and and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reiner, A., Yekutieli, D. and Y. Benjamini (2003): Identifying differentially expressed genes using the false discovery rate controlling procedures. *Bioinformatics*, 19 (3), 368–375.

Sachs, L. (1997): *Angewandte Statistik*. 7. edition, Springer Verlag, Berlin.

Sasabuchi, S. (1988): A multivariate one-sided test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor, *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, 42, 9–19.

Schaarschmidt, F. (2005): *Binomial group testing - Design and Analysis*. Diplomarbeit, Universität Hannover.

Schmidt, D.D., Voelckel, C., Hartl, M., Schmidt, S. and I.T. Baldwin (2005): Specificity in ecological interactions. Attack from the same lepidopteran herbivore results in species-specific transcriptional responses in two solanaceous host plants. *Plant Physiology*, 138, 1763–1773.

Shaffer, J.P. (1995): Multiple hypothesis testing. *Annual Rev. Psychology*, 561–584.

Speed, T. (2005): Always log spot intensities and ratios. *Speed Group Microarray Page*, http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html.

Shedden, K. (2004): Confidence levels for the comparison of microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 (1), article 32.

'Student' [Gossett, W.S.] (1908). *Biometrika*, 6, 1–25.

Troyanskaya, O.G., Garber, M.E., Brown, P.O, Bostein, D. and R.B. Altman (2002): Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18 (11), 1454–1461.

Tusher, V., R. Tibshirani and G. Chu (2001): Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 98: 5116–5121.

Uppalapati, S.R., Ayoubi, P., Weng, H., Palmer, D.A., Mitchell, R.E., Jones, W. and C.L. Bender (2005): The phytotoxin coronatine and methyl jasmonate impact multiple phytohormone pathways in tomato. *The Plant Journal*, 42, 201–217.

Victor, N. (1982): Exploratory data analysis and clinical research. *Methods of Information in Medicine*, 21, 53–54.

Westfall, P.H. and A. Krishen (2001): Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planing and Inference*, 99, 25–40.

Westfall, P.H., Kropf, S. and L. Finos (2004): Weighted FWE-controlling methods in high-dimensional situations. *Recent Developments in Multiple Comparison Procedures* (Benjamini, Y., Bretz, F. and S.K. Sarkar, eds.), IMS Lecture Notes and Monograph Series, 47, 143–154.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. and Y. Hochberg (1999): *Multiple Comparisons and Multiple Tests Using the SAS System*, Cary, NC: SAS Institute Inc.

Westfall, P.H. and S.S. Young (1993): *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons, New York.

Wilcoxon, F. (1945): Individual comparisons by ranking methods. *Biometrics* 1, 80–83.

Wright, S.P. (1992): Adjusted p-values for simultaneous inference. *Biometrics* 48, 1005–1013.

Appendix A

Simulations

In this chapter a selection of simulation results of the FWER is presented. Irrespective of the power properties, those methods are listed which do not clearly exceed the FWER. This chapter is divided into four parts: First results from the parametric procedures to test for relevant differences are given, then the nonparametric methods to test for a relevant difference follow. Afterwards the simulation results of the parametric methods for relevant ratios and last for the nonparametric procedures are listed.

For all four parts the weak and the strong control of the FWER is analyzed. Only tables for one-sided testing against a relevant increase and for two-sided testing are listed because testing one-sided against a relevant decrease gives equivalent results as one-sided against a relevant increase. Simulation results exceeding the FWER according to the Wilson confidence limit (section 4.3) are printed bold. If not stated otherwise, a simulation scenario includes 50 endpoints and a FWER of 5% is chosen. Every simulation result is computed with the same starting value and 10,000 simulation runs. Likewise the number of permutations of the minP-algorithm is restricted to 10,000.

A.1 Parametric procedures to test for relevant differences

The first section presents a table with simulation results of the FWER for parametric procedures with a data-driven order of hypotheses to test for relevant differences. These are the shift-selector, the δ-shift and the random$^{\delta}$ procedure.

Most of the parameters, as for example sample size per group, vary between the scenarios. However depending on one- or two-sided testing all scenarios have the same computation of the true mean values. For one-sided testing the expected values for endpoints under the null hypothesis are $\mu_{1j} = 100$ and $\mu_{2j} = \tau_{\delta}^{H_0} + 100$, where for each j $\tau_{\delta}^{H_0}$ is assigned randomly a value between 0 and δ_{upper} in steps of 5 units. Endpoints under H_1 have the true mean values $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \delta_{upper} + 50$. For two-sided testing $\tau_{\delta}^{H_0}$ is a random value between δ_{lower} and δ_{upper} in 5 steps. If $\tau_{\delta}^{H_0} \geq 0$ endpoints under H_0 have the true mean values $\mu_{1j} = 100$ and $\mu_{2j} = \tau_{\delta}^{H_0} + 100$ and for $\tau_{\delta}^{H_0} < 0$ $\mu_{1j} = |\tau_{\delta}^{H_0}| + 100$ and $\mu_{2j} = 100$. If the control of the strong FWER is tested, three endpoints under H_1 receive the expected values $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \delta_{upper} + 50$ and two have $\mu_{1j} = 100 + |\delta_{lower}| + 50$ and $\mu_{2j} = 100$.

Table A1: Parametric for relevant differences

side	$-\delta_{lower} = \delta_{upper}$	n_i	σ	$\rho_{ijj'}$	shift-selector		δ-shift		random$^{\delta}$	
					weak	strong	weak	strong	weak	strong
one-sided	0	5	10	0.1	5.09	5.14	**5.84**	**5.89**	4.90	4.93
	50	5	10	0.1	0	0	0.11	0.12	0.07	0.06
	400	5	10	0.1	0	0	0.01	0.01	0.01	0.02
	0	30	10	0.1	5.17	5.09	**5.49**	**5.39**	4.54	4.42
	50	30	10	0.1	0	0	0	0	0	0
	400	30	10	0.1	0	0	0	0	0	0
	0	5	10	0.9	4.89	4.90	4.89	4.90	5.08	5.25
	50	5	10	0.9	0	0	0.81	0.84	0.71	0.067
	400	5	10	0.9	0	0	0.22	0.24	0.17	0.19
	0	5	50	0.1	5.09	4.45	**5.84**	5.03	4.90	4.30
	50	5	50	0.1	0.45	0.59	1.30	1.10	0.99	0.79
	400	5	50	0.1	0	0	0.16	0.14	0.10	0.13
two-sided	0	5	10	0.1	5.11	5.14	5.11	5.14	5.11	5.14
	50	5	10	0.1	0	0	0.38	0.48	0.26	0.22
	400	5	10	0.1	0	0	0.06	0.08	0.05	0.03
	0	30	10	0.1	5.09	5.00	5.09	5.00	5.09	5.00
	50	30	10	0.1	0	0	0.23	0.31	0.19	0.18
	400	30	10	0.1	0	0	0.04	0.05	0.02	0.02
	0	5	10	0.9	5.02	4.98	5.02	4.98	5.02	4.98
	50	5	10	0.9	0	0	3.23	3.01	2.50	2.44
	400	5	10	0.9	0	0	0.81	0.76	0.039	0.58
	0	5	50	0.1	5.11	4.24	5.11	4.24	5.11	4.24
	50	5	50	0.1	0.65	0.58	0.88	0.72	0.65	0.59
	400	5	50	0.1	0	0	0.12	0.12	0.08	0.06

All characteristics of the behavior of the FWER will appear in the following tables for all other methods (test for relevant means/ratios and parametric/nonparametric) as well. Hence these effects are discussed exemplary for these procedures and not further mentioned for the further ones.

The δ-shift exceeds slightly the FWER in the case of one-sided testing in combination with the setting of $\delta_{lower} = \delta_{upper} = 0$. But more important this method controls empirically the FWER for two-sided testing.

For the shift-selector procedure the empirical FWER is often 0, because the selector has the local maximum if the difference in means of an endpoint is 0. To receive a larger selector statistic than those variables, endpoints have to have a high treatment effect, which in unlikely to appear in these simulation settings.

With increasing distance of $\delta_{\mathbf{side}}$ from 0 the empirical FWER decreases in all scenarios, because the probability to achieve a false positive endpoint decreases as the endpoint under H_0 achieve a random difference in means larger than δ_{lower} and less than δ_{upper}.

In the special case of two-sided testing and $\delta_{\mathbf{lower}} = \delta_{\mathbf{upper}} = \mathbf{0}$ all three methods receive the same error rates because in the two-sided case the δ-shift and the random$^{\delta}$ method transform endpoints with $\delta_{lower} < \bar{x}_{2jk} - \bar{x}_{1jk} < \delta_{upper}$, that is $0 < \bar{x}_{2jk} - \bar{x}_{1jk} < 0$. Therefore no endpoints are transformed by both methods and all three procedures have the same number of false positives in the simulation.

A.2 Nonparametric procedures to test for relevant differences

This section gives the empirical results of the FWER for the nonparametric procedure with a data-driven order of hypotheses to test for relevant differences, which is the np-δ-shift method. If not stated otherwise the settings of the expected values are the same as for the parametric test for relevant differences.

Table B1: Nonparametric for relevant differences, FWER dependent on n_i

The first table shows the empirical FWER for exact and asymptotic testing when the sample size per group varies. In this setting a more unrealistic selection of the expected values is used as it gives the highest empirical FWER for scenarios with $\delta_{lower}, \delta_{upper} \neq 0$. Here all endpoints under H_0 have a difference in means equal to either δ_{lower} or δ_{upper}. For the weak control 25 endpoints have $\mu_{1j} = 100 + |\delta_{lower}|$ and $\mu_{2j} = 100$ and the other 25 variables have $\mu_{1j} = 100$ and $\mu_{2j} = 100 + \delta_{upper}$. In scenarios analyzing the strong control the true means of endpoint under H_0 are set to the same values, but for 22 variables on decrease and 23 on increase.

All scenarios include two-sided tests.

assumption	$-\delta_{lower} = \delta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak	strong
exact	0	4	10	0.1	2.96	2.88
	400	4	10	0.1	2.20	2.07
	0	7	10	0.1	3.76	3.89
	400	7	10	0.1	2.42	2.54
	0	15	10	0.1	4.70	4.62
	400	15	10	0.1	2.79	2.61
asymptotic	0	4	10	0.1	**5.90**	**5.69**
	400	4	10	0.1	4.38	4.11
	0	7	10	0.1	**5.36**	**5.43**
	400	7	10	0.1	3.49	3.48
	0	15	10	0.1	5.21	5.09
	400	15	10	0.1	3.09	2.92

By use of the exact tests the FWER is protected empirically in all analyzed settings. The asymptotic procedures exceed the FWER when the sample sizes are too small and $\delta_{lower} = \delta_{upper} = 0$. The reason why the FWER is not exceeded for the other setting of the relevance threshold is, that the FWER generally decreases in scenarios with $\delta_{lower}, \delta_{upper} \neq 0$:
For the exact methods the empirical FWER is a value roughly equal to the nominal FWE/2. The reason for the dependence of the FWER on the thresholds may be that while for $\delta_{lower}, \delta_{upper} \neq 0$ half of the data show a difference of means > 0 and the other half has a difference < 0 and to receive significant result, the simulated data of an endpoint has to be in the same direction as the endpoint's true treatment effect is. Either the simulated data for endpoints testing against δ_{lower} have to have an empirical difference $< \delta_{lower}$ or the endpoints against δ_{upper} an empirical difference $> \delta_{upper}$. For $\delta_{lower} = \delta_{upper} = 0$ the empirical differences have to be only significantly different from 0.
For example in the third row ($\delta_{lower} = \delta_{upper} = 0$), for the weak control and the exact procedure an endpoint has an empirical difference in means of 21.5175 and receives an unadjusted p-value of 3.79%. The same endpoint can be observed with $-\delta_{lower} = \delta_{upper} = 400$. Here the difference in means is -378.4825, as for this endpoint the true difference in means is set to -400 and the empirical one is given by the true difference minus the simulated effect from the random number generator: -400+21.5175=-378.4825. With the difference

in means not exceeding the corresponding relevance threshold δ_{lower} the unadjusted p-value for this endpoint is 1.

The procedures with a data-driven order of hypotheses are superior to alternative methods when the sample sizes are small. Therefore the exact method is proposed, as it provides an empirically stronger control of the FWER.

Table B2: Nonparametric for relevant differences, asymptotic, $\theta_{lower} \leq \mu_{2j} - \mu_{1j} \leq \theta_{upper}$ for endpoints under H_0

However for the sake of completeness a table with simulation results for the asymptotic procedure is listed. The table gives results on the weak and the strong control of FWER. Analyzed are diverse settings of sample size, relevance thresholds, variances and correlations. The endpoints under H_0 have a random true difference in means.

side	$-\delta_{lower} = \delta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak	strong
one-sided	0	10	10	0.1	**6.10**	**6.08**
	50	10	10	0.1	0.95	0.86
	400	10	10	0.1	0.29	0.21
	0	30	10	0.1	**5.39**	**5.43**
	50	30	10	0.1	0.56	0.58
	400	30	10	0.1	0.11	0.15
	0	30	10	0.9	5.20	5.25
	50	30	10	0.9	1.16	1.22
	400	30	10	0.9	0.26	0.27
	0	30	15	0.1	**5.36**	**5.65**
	50	30	15	0.1	0.61	0.63
	400	30	15	0.1	0.16	0.16
two-sided	0	10	10	0.1	**5.37**	5.28
	50	10	10	0.1	0.44	0.39
	400	10	10	0.1	0.06	0.04
	0	30	10	0.1	5.13	4.89
	50	30	10	0.1	0.29	0.35
	400	30	10	0.1	0.03	0.03
	0	30	10	0.9	5.13	5.13
	50	30	10	0.9	0.75	0.83
	400	30	10	0.9	0.10	0.12
	0	30	15	0.1	5.13	4.89
	50	30	15	0.1	0.28	0.36
	400	30	15	0.1	0.03	0.03

The same pattern as in the former tables can be seen: the np-δ-shift method shows slight exceeds of the FWER when a one-sided testing scenario is chosen and $\delta_{lower} = \delta_{upper} = 0$.

Table B3: Nonparametric for relevant differences, exact, $\theta_{lower} \leq \mu_{2j} - \mu_{1j} \leq \theta_{upper}$ for endpoints under H_0

This table is equivalent to the former one in terms of settings and results, but for the exact version with smaller sample sizes.

side	$-\delta_{lower} = \delta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak	strong
one-sided	0	5	10	0.1	**6.22**	**6.24**
	50	5	10	0.1	0.20	0.17
	400	5	10	0.1	0.02	0.02
	0	10	10	0.1	5.23	5.15
	50	10	10	0.1	0.07	0.07
	400	10	10	0.1	0.01	0.01
	0	5	10	0.9	4.77	4.82
	50	5	10	0.9	0.60	0.58
	400	5	10	0.9	0.13	0.15
	0	5	15	0.1	**6.22**	**6.22**
	50	5	15	0.1	0.38	0.37
	400	5	15	0.1	0.04	0.05
two-sided	0	5	10	0.1	3.47	3.10
	50	5	10	0.1	0.32	0.32
	400	5	10	0.1	0.04	0.03
	0	10	10	0.1	4.48	4.30
	50	10	10	0.1	0.38	0.31
	400	10	10	0.1	0.05	0.03
	0	5	10	0.9	3.13	3.25
	50	5	10	0.9	1.32	1.42
	400	5	10	0.9	0.32	0.31
	0	5	15	0.1	3.47	2.98
	50	5	15	0.1	0.35	0.32
	400	5	15	0.1	0.04	0.03

Table B4: Nonparametric for relevant differences, exact with (non-)normal data and true differences under H_0 equal to $\delta_{lower},\ \delta_{upper}$

The next table shows simulation results for the weak and the strong sense of the FWER by use of the exact method. Different to the setting above, the differences for endpoints under H_0 are set to the margins of the null hypothesis as in table table B1. For the strong control the five endpoints under H_1 have expected values as explained in the introduction of this chapter. Only two-sided testing is analyzed. In the first two columns of the empirical results the individual samples per group are multivariate Gaussian distributed, while in the two following ones the data follows a multivariate skewed distribution with a skewness of 2 and a kurtosis of 7 generated with the algorithm proposed by Fleishman (1978).

$-\delta_{lower} = \delta_{upper}$	n_i	σ	$\rho_{ijj'}$	normal		skewed	
control				weak	strong	weak	strong
0	5	10	0.1	3.47	3.10	3.27	2.87
50	5	10	0.1	2.49	2.27	2.40	1.91
400	5	10	0.1	2.49	2.27	2.40	1.91
0	10	10	0.1	4.48	4.30	4.11	4.32
50	10	10	0.1	2.72	2.81	2.48	2.60
400	10	10	0.1	2.72	2.81	2.48	2.60
0	5	10	0.9	3.13	3.25	3.17	2.99
50	5	10	0.9	2.14	2.20	2.20	2.15
400	5	10	0.9	2.14	2.20	2.20	2.15
0	5	15	0.1	3.47	2.98	3.27	2.26
50	5	15	0.1	2.49	2.16	2.40	1.51
400	5	15	0.1	2.49	2.16	2.40	1.51

A.3 Parametric procedures to test for relevant ratios

This section presents simulation results for parametric procedures with a data-driven order of hypotheses to test for relevant ratios discussed in chapter 5. These are the Sasabuchi selector, the θ-shift and random$^{\theta}$ method.

As for testing against a relevant difference many parameters vary between the scenarios except the expected values. If not stated otherwise, for one-sided testing the expected values for endpoints under the null hypothesis are $\mu_{1j} = 100$ and $\mu_{2j} = \tau_{\theta}^{H_0} \cdot 100$, where for each j $\tau_{\theta}^{H_0}$ is a random value between 1 and θ_{upper} in 0.05 steps. Endpoints under H_1 have the true mean values $\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot (\theta_{upper} + 0.5))$. In the two-sided case $\tau_{\theta}^{H_0}$ is set for each endpoint under H_0 to a random value chosen between θ_{lower} and θ_{upper} in 0.05 steps as described in section 5.1.1. The five endpoints under H_1 receive either the expected values $\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot (\theta_{upper} + 0.5)$ (three endpoints) or $\mu_{1j} = 100 \cdot \theta_{lower}^{-1} + 0.5$ and $\mu_{2j} = 100$ (two endpoints).

Table C1: Parametric for relevant ratios - weak control / one-sided

The first table gives the simulation results for the weak control of the FWER by testing one-sided for a relevant increase. In all scenarios 50 endpoints under the null hypothesis are tested. The expected values are set as denoted in the introduction of this section. All other relevant information is given in the table.

θ_{upper}	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector			θ-shift			random$^{\theta}$		
				$\alpha = 1\%$	5%	10%	1%	5%	10%	1%	5%	10%
1	5	10	0.1	1.05	5.09	10.03	**1.21**	**5.84**	**11.61**	1.01	4.90	9.61
1.5	5	10	0.1	0	0	0	0.17	0.89	1.75	0.11	0.55	1.22
5	5	10	0.1	0	0	0	0.04	0.17	0.33	0.03	0.17	0.33
1	10	10	0.1	1.01	4.87	9.85	1.10	5.35	**10.97**	0.97	4.58	9.12
1.5	10	10	0.1	0	0	0	0.14	0.63	1.37	0.10	0.45	0.89
5	10	10	0.1	0	0	0	0.03	0.13	0.26	0.02	0.08	0.16
1	30	10	0.1	1.05	5.17	10.20	1.12	**5.49**	**10.89**	0.86	4.54	8.86
1.5	30	10	0.1	0	0	0	0.14	0.58	1.08	0.05	0.31	0.66
5	30	10	0.1	0	0	0	0.02	0.11	0.25	0.01	0.05	0.11
1	5	10	0.9	0.95	4.89	9.95	0.95	4.89	9.95	0.89	5.08	10.09
1.5	5	10	0.9	0.03	0.09	0.09	0.80	3.51	5.96	0.74	2.84	5.03
5	5	10	0.9	0	0	0	0.41	1.59	2.55	0.38	1.51	2.39
1	10	10	0.9	0.92	4.94	9.72	0.92	4.94	9.72	1.11	5.09	10.11
1.5	10	10	0.9	0	0	0	0.78	2.82	4.47	0.66	2.19	3.70
5	10	10	0.9	0	0	0	0.30	0.96	1.48	0.26	0.93	1.48
1	30	10	0.9	1.05	5.02	10.08	1.05	5.02	10.08	1.08	5.11	10.13
1.5	30	10	0.9	0	0	0	0.05	1.59	2.59	0.38	1.49	2.68
5	30	10	0.9	0	0	0	0.09	0.37	0.063	0.12	0.43	0.83
1	5	50	0.1	1.05	5.09	10.03	1.21	**5.84**	**11.61**	1.01	4.89	9.62
1.5	5	50	0.1	0.33	1.72	3.77	0.37	1.98	4.33	0.30	1.50	3.17
5	5	50	0.1	0.01	0.04	0.09	0.11	0.69	1.64	0.08	0.46	1.08

In total analogy to the tests for relevant differences, the only procedure which shows slight exceeds of the FWER is the θ-shift procedure in the special case of one-sided testing and $\theta_{lower} = \theta_{upper} = 1$.

Table C2: Parametric for relevant ratios - weak control / two-sided

This table gives results for the same scenarios as the last table but for the two-sided problem. Hence no exceeds of the FWER are to be expected.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector			θ-shift			random$^{\theta}$		
				$\alpha = 1\%$	5%	10%	1%	5%	10%	1%	5%	10%
1	5	10	0.1	1.08	5.11	10.21	1.08	5.11	10.21	1.08	5.11	10.21
1.5	5	10	0.1	0	0	0	0.09	0.44	0.88	0.08	0.30	0.59
5	5	10	0.1	0	0	0	0.03	0.10	0.22	0.01	0.06	0.12
1	10	10	0.1	0.99	4.82	9.90	0.99	4.82	9.90	0.99	4.82	9.90
1.5	10	10	0.1	0	0	0	0.07	0.38	0.72	0.05	0.22	0.45
5	10	10	0.1	0	0	0	0.02	0.06	0.13	0	0.06	0.09
1	30	10	0.1	0.98	5.09	10.05	0.98	5.09	10.05	0.98	5.09	10.05
1.5	30	10	0.1	0	0	0	0.05	0.31	0.61	0.04	0.20	0.38
5	30	10	0.1	0	0	0	0	0.06	0.13	0	0.02	0.04
1	5	10	0.9	1.02	5.02	9.99	1.02	5.02	9.99	1.02	5.02	9.99
1.5	5	10	0.9	0	0	0	0.71	3.05	5.47	0.69	2.63	3.98
5	5	10	0.9	0	0	0	0.33	1.25	2.09	0.34	1.00	1.50
1	10	10	0.9	0.98	4.90	10.09	0.98	4.90	10.09	0.98	4.90	10.09
1.5	10	10	0.9	0	0	0	0.67	2.50	3.95	0.047	1.46	2.19
5	10	10	0.9	0	0	0	0.23	0.69	1.08	0.14	0.43	0.59
1	30	10	0.9	1.11	4.94	9.89	1.11	4.94	9.89	1.11	4.94	9.89
1.5	30	10	0.9	0	0	0	0.43	1.38	2.11	0.21	0.57	0.81
5	30	10	0.9	0	0	0	0.07	0.19	0.35	0.03	0.07	0.11
1	5	50	0.1	1.07	5.11	10.21	1.07	5.11	10.21	1.07	5.11	10.21
1.5	5	50	0.1	0.17	0.94	1.85	0.20	1.12	2.24	0.10	0.71	1.64
5	5	50	0.1	0.01	0.01	0.03	0.06	0.37	0.82	0.04	0.25	0.52

Table C3: Parametric for relevant ratios - strong control / one-sided

This table gives simulation results for the strong control of the FWER. Apart from the inclusion of endpoints under the alternative hypothesis it is equal to table C1. Out of the 50 endpoints 45 are under the null hypothesis and 5 show differences between the treatment groups. The expected values for these 5 endpoints are given as explained in the introduction of this section.

θ_{upper}	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector			θ-shift			random$^{\theta}$		
				$\alpha = 1\%$	5%	10%	1%	5%	10%	1%	5%	10%
1	5	10	0.1	1.09	5.00	10.09	1.13	**5.69**	**11.67**	0.93	4.96	9.87
1.5	5	10	0.1	0	0	0	0.18	0.90	1.82	0.11	0.53	1.06
5	5	10	0.1	0	0	0	0.01	0.16	0.38	0.02	0.11	0.25
1	10	10	0.1	0.92	4.88	9.86	1.01	**5.43**	**10.85**	0.87	4.40	9.15
1.5	10	10	0.1	0	0	0	0.11	0.59	1.25	0.07	0.49	0.97
5	10	10	0.1	0	0	0	0.03	0.17	0.29	0.02	0.12	0.25
1	30	10	0.1	0.94	5.10	10.09	1.15	**5.47**	**10.79**	0.88	4.32	8.78
1.5	30	10	0.1	0	0	0	0.11	0.53	1.11	0.05	0.31	0.65
5	30	10	0.1	0	0	0	0.02	0.12	0.23	0	0.05	0.13
1	5	10	0.9	1.05	5.07	10.13	1.01	5.07	10.20	0.93	5.19	10.02
1.5	5	10	0.9	0.05	0.07	0.07	0.83	3.45	5.80	0.75	2.96	5.08
5	5	10	0.9	0	0	0	0.40	1.60	2.59	0.31	1.32	2.36
1	10	10	0.9	0.95	5.06	10.09	0.91	5.06	9.96	0.96	4.98	10.03
1.5	10	10	0.9	0	0	0	0.70	2.64	4.36	0.56	2.15	3.82
5	10	10	0.9	0	0	0	0.31	1.00	1.49	0.21	0.79	1.32
1	30	10	0.9	1.07	5.10	10.13	1.05	5.10	9.95	1.02	5.02	10.14
1.5	30	10	0.9	0	0	0	0.52	1.69	2.57	0.38	1.50	2.63
5	30	10	0.9	0	0	0	0.14	0.38	0.61	0.10	0.47	0.80
1	5	50	0.1	0.76	4.33	9.35	0.81	4.85	**10.69**	0.69	4.23	9.21
1.5	5	50	0.1	0.22	1.48	3.34	0.25	1.67	3.86	0.18	1.31	2.96
5	5	50	0.1	0.01	0.06	0.17	0.11	0.69	1.43	0.08	0.49	1.06

Table C4: Parametric for relevant ratios - strong control / two-sided

In the following table the same scenarios as in the last one are observed, but for two-sided testing.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector			θ-shift			random$^{\theta}$		
				$\alpha = 1\%$	5%	10%	1%	5%	10%	1%	5%	10%
1	5	10	0.1	0.99	4.97	9.97	0.99	4.97	9.97	0.99	4.97	9.97
1.5	5	10	0.1	0	0	0	0.10	0.46	0.87	0.08	0.27	0.53
5	5	10	0.1	0	0	0	0.02	0.07	0.16	0.01	0.05	0.12
1	10	10	0.1	1.00	5.00	10.09	1.00	5.00	10.09	1.00	5.00	10.09
1.5	10	10	0.1	0	0	0	0.08	0.37	0.74	0.07	0.25	0.49
5	10	10	0.1	0	0	0	0.01	0.07	0.16	0	0.05	0.12
1	30	10	0.1	0.98	5.06	10.07	0.98	5.06	10.07	0.98	5.06	10.07
1.5	30	10	0.1	0	0	0	0.09	0.30	0.63	0.01	0.12	0.29
5	30	10	0.1	0	0	0	0.02	0.06	0.12	0.01	0.03	0.07
1	5	10	0.9	1.02	5.12	10.03	1.02	5.12	10.03	1.02	5.12	10.03
1.5	5	10	0.9	0	0	0	0.66	2.95	5.18	0.57	2.34	3.76
5	5	10	0.9	0	0	0	0.27	1.13	1.93	0.26	0.92	1.47
1	10	10	0.9	1.03	4.91	10.12	1.03	4.91	10.12	1.03	4.91	10.12
1.5	10	10	0.9	0	0	0	0.67	2.36	3.81	0.47	1.45	2.09
5	10	10	0.9	0	0	0	0.14	0.73	1.19	0.11	0.32	0.47
1	30	10	0.9	1.03	5.18	10.15	1.03	5.18	10.15	1.03	5.18	10.15
1.5	30	10	0.9	0	0	0	0.40	1.28	1.95	0.24	1.45	0.90
5	30	10	0.9	0	0	0	0.07	0.27	0.38	0.02	0.32	0.15
1	5	50	0.1	0.72	4.15	8.74	0.72	4.15	8.74	0.72	4.15	8.74
1.5	5	50	0.1	0.13	0.71	1.61	0.16	0.82	1.85	0.14	0.70	1.43
5	5	50	0.1	0	0.01	0.02	0.08	0.33	0.70	0.02	0.20	0.42

As well as for the weak control an exceed of the FWER for the θ-shift method appears in the one-sided case only.

Table C5: Parametric for relevant ratios - varying number of endpoints under H_0

In the following table the empirical FWER is shown for a varying number of endpoints under the null hypothesis. The upper half of the table gives results for one-sided hypothesis testing, and the lower part presents the results for two-sided testing. The true means for the endpoints are set as explained in the beginning of this section. The further parameters are: $n_i = 5$, $\rho_{ijj'} = 0.1$ and $\sigma = 10$. Note that the rows 3, 6, 9 and 12 show the results of the weak control and the eight others give results for the strong one.

side	relevance thresholds	number of endpoints under H_0 (out of 50)	empirical FWE		
			Sasabuchi selector	θ-shift	random$^\theta$
one-sided	1	1	4.92	4.92	5.07
	1	25	5.12	**5.81**	5.14
	1	50	5.09	**5.84**	4.90
	5	1	0	0	0
	5	25	0	0.10	0.08
	5	50	0	0.17	0.17
two-sided	1 / 1	1	5.13	5.13	5.13
	1 / 1	25	5.25	5.25	5.25
	1 / 1	50	5.11	5.11	5.11
	0.2 / 5	1	0	0	0
	0.2 / 5	25	0	0.03	0.01
	0.2 / 5	50	0	0.10	0.06

As well as for tables C1 to C4 the empirical FWERs are close to the nominal levels only for relevance threshold(s) of 1. In the two-sided case for $\theta_{lower} = \theta_{upper} = 1$ all error rates are again the same because no endpoints have to be transformed by the θ-shift and the random$^\theta$ method. For one-sided testing, $\theta_{lower} = \theta_{upper} = 1$ and one endpoint under H_0 the Sasabuchi-selector and the θ-shift procedure have the same empirical FWER: Either the ratio of the endpoint is less than 1; in this case it will not be a false positive (the data transformation of the θ-shift procedure changes the values of the second treatment group, which has no effect on the FWER in this setting). Or the ratio of the endpoint is greater than 1; then the data transformation is not used by the θ-shift method and the algorithms

of both procedures are equal. The random$^{\theta}$ procedure would show exactly the same error rate of 4.92 %, but the data transformation of this procedure includes the use of the random number generator. Therefore other random numbers are used for the generation of the data. Although the true FWER is equal for all three methods, the empirical one is slightly different for the random$^{\theta}$ method.

Tables C6: Parametric for relevant ratios - ratio under H_0 equal to $\theta_{lower},\ \theta_{upper}$
The following four tables give simulation results of the FWER-control concerning the weak sense for $\alpha = 5\%$. These results present the maximal exceeds of the FWER of all parametric procedures, which occurred in the one-sided case only. Basically the scenarios are the same as in tables C1 to C4 but the true ratios under the null hypothesis are set to the relevance thresholds. In the first table all mean data levels are set to 100 and in the second table the mean level of the endpoints are exponential distributed including a constant coefficient of variation as presented in section 7.6. Table 3 and 4 are the same but for the control of the FWER in the strong sense. The means of the 5 endpoints under H_1 are set as introduced in the beginning of this section.

Table C6.1: Weak control with mean level of 100

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
side				one	two	one	two	one	two
1	5	10	0.1	5.09	5.11	**5.84**	5.11	4.90	5.11
1.5	5	10	0.1	5.19	2.57	**5.87**	3.27	5.02	2.73
5	5	10	0.1	5.13	2.69	**5.92**	3.25	4.97	2.57
1	10	10	0.1	4.87	4.82	5.35	4.82	4.58	4.82
1.5	10	10	0.1	4.77	2.54	5.29	2.99	4.56	2.40
5	10	10	0.1	4.85	2.58	5.33	2.98	4.66	2.38
1	30	10	0.1	5.17	5.09	**5.49**	5.09	4.54	5.09
1.5	30	10	0.1	5.16	2.43	**5.47**	5.01	4.49	2.18
5	30	10	0.1	5.10	2.68	**5.41**	4.93	4.43	2.23
1	5	50	0.1	5.09	5.11	**5.84**	5.11	4.89	5.11
1.5	5	50	0.1	**5.46**	2.92	**5.87**	3.27	5.00	2.65
5	5	50	0.1	5.13	2.68	**5.91**	3.25	4.97	2.57

Table C6.2: Weak control with exponential distributed mean level

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	cv	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
side				one	two	one	two	one	two
1	5	0.1	0.1	5.11	5.00	**5.39**	5.00	**5.36**	5.00
1.5	5	0.1	0.1	5.03	2.54	**5.38**	2.72	**5.41**	2.85
5	5	0.1	0.1	5.05	2.47	**5.37**	2.63	**5.45**	2.71
1	10	0.1	0.1	4.86	4.93	5.02	4.93	5.11	4.93
1.5	10	0.1	0.1	4.81	2.44	4.97	3.34	5.22	2.57
5	10	0.1	0.1	4.96	2.44	5.12	2.55	5.30	2.67
1	30	0.1	0.1	5.19	5.02	5.09	5.02	4.90	5.02
1.5	30	0.1	0.1	5.28	2.62	5.10	2.54	4.96	3.79
5	30	0.1	0.1	5.21	2.67	5.05	2.54	5.02	2.47
1	5	0.5	0.1	5.11	5.00	**5.39**	5.00	5.34	5.00
1.5	5	0.5	0.1	5.13	2.58	**5.38**	2.72	**5.38**	2.64
5	5	0.5	0.1	5.05	2.46	**5.37**	2.63	**5.49**	2.71

Table C6.3: Strong control with mean level of 100

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
side				one	two	one	two	one	two
1	5	10	0.1	5.00	4.97	**5.69**	4.97	4.93	4.97
1.5	5	10	0.1	5.08	2.46	**5.76**	3.13	5.01	2.53
5	5	10	0.1	4.38	2.06	4.98	2.45	4.39	2.02
1	10	10	0.1	4.88	5.00	**5.43**	5.00	4.66	5.00
1.5	10	10	0.1	4.90	2.69	**5.36**	3.13	4.32	2.41
5	10	10	0.1	4.84	2.50	5.26	2.84	4.41	2.19
1	30	10	0.1	5.10	5.06	5.07	5.06	4.23	5.06
1.5	30	10	0.1	5.10	2.48	5.13	2.78	4.49	2.22
5	30	10	0.1	4.99	2.47	5.07	2.75	4.35	2.27
1	5	50	0.1	4.33	4.15	4.85	4.15	4.20	4.15
1.5	5	50	0.1	4.71	2.35	5.03	2.59	4.33	2.31
5	5	50	0.1	4.57	2.35	5.23	2.83	4.15	2.27

Table C6.4: Strong control with exponential distributed mean level

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	cv	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
side				one	two	one	two	one	two
1	5	0.1	0.1	4.87	5.01	5.17	5.01	5.18	5.01
1.5	5	0.1	0.1	4.85	2.70	5.18	2.89	5.30	2.78
5	5	0.1	0.1	4.67	2.39	4.95	2.53	5.00	2.45
1	10	0.1	0.1	5.05	5.00	5.17	5.00	4.89	5.00
1.5	10	0.1	0.1	5.02	2.54	5.15	2.63	**5.36**	2.55
5	10	0.1	0.1	5.05	2.60	5.18	2.69	5.20	2.65
1	30	0.1	0.1	5.33	5.09	**5.36**	5.09	5.20	5.09
1.5	30	0.1	0.1	5.25	2.69	5.29	2.73	5.32	2.60
5	30	0.1	0.1	5.30	2.64	5.35	2.77	5.12	2.46
1	5	0.5	0.1	4.53	4.50	4.79	4.50	4.72	4.50
1.5	5	0.5	0.1	4.54	2.48	4.73	2.61	5.11	2.46
5	5	0.5	0.1	4.47	2.36	4.72	2.52	5.01	2.42

Here random$^{\theta}$ procedure exceeds slightly the FWER for one-sided testing as well. However in the two-sided case this method still controls the false positive rate.

Table C7: Parametric for relevant ratios - varying ratio of endpoints under H_0
In the following table the control of the FWER depending on the ratio of means of the endpoints under H_0 is observed. Both the weak and the strong sense are simulated. The expected values for the endpoints under the alternative hypothesis are computed as denoted in the introduction of this section. And for the endpoints under H_0 the true ratio of means is set to θ_{upper} (one-sided) or to one of the two relevance thresholds (two-sided). The remaining parameters are set to $n_i = 5$, $\sigma = 10$, $\rho_{ijj'} = 0.1$, $\theta_{lower}^{-1} = \theta_{upper} = 5$.

side	$\tau_\theta^{H_0}$	empirical FWER (weak sense)			empirical FWER (strong sense)		
		Sasabuchi selector	θ-shift	random$^\theta$	Sasabuchi selector	θ-shift	random$^\theta$
one-sided	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	0	0	0	0	0
	4	0	0	0	0	0	0
	5	5.13	**5.92**	4.97	4.38	4.98	4.39
two-sided	1	0	0	0	0	0	0
	(0.5 or) 2	0	0	0	0	0	0
	(1/3 or) 3	0	0	0	0	0	0
	(0.25 or) 4	0	0	0	0	0	0
	(0.2 or) 5	2.69	3.25	2.57	2.06	2.45	2.02

If the endpoints under H_0 receive a true ratio less than θ_{upper} (one-sided) or less than θ_{upper} and greater than θ_{lower} (two-sided) the error rates are close or exactly equal to 0, because the probability that the empirical ratio exceeds the threshold(s) is small. Only with true ratios equal to the relevance threshold(s) is the probability relatively high that false positives appear; hence only in these settings is the empirical error rate larger than 0.

Table C8: Parametric for relevant ratios - varying correlation of endpoints under H_0

The next two tables show the weak and the strong control for varying correlations between endpoints. In the simulation settings for the first table three of the endpoints under H_0 have a correlation close to zero ($\rho_{ijj'} = 0.01$) and in the second table the correlation of three endpoints is set to -0.3. For all other endpoints the correlation is 0.3. Only two-sided testing is observed.

The expected values of the endpoints under H_1 (strong control) are chosen as given in the introduction of this section. And the endpoints under H_0 have true ratios equal to one of the relevance thresholds.

C8.1: Three endpoints with $\rho_{ijj'} = 0.01$

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
control				weak	strong	weak	strong	weak	strong
1	5	10	0.3	5.18	4.96	5.18	4.96	5.18	4.96
1.5	5	10	0.3	2.68	2.63	3.60	3.50	3.05	2.96
5	5	10	0.3	3.01	2.35	3.56	2.76	2.97	2.45
1	10	10	0.3	4.94	5.11	4.94	5.11	4.94	5.11
1.5	10	10	0.3	2.55	2.60	3.26	3.27	2.56	2.66
5	10	10	0.3	2.69	2.56	3.04	2.98	2.78	2.51
1	30	10	0.3	4.98	5.04	4.98	5.04	4.98	5.04
1.5	30	10	0.3	2.66	2.43	3.01	2.81	2.14	2.19
5	30	10	0.3	2.85	2.51	3.09	2.72	2.33	2.27
1	5	50	0.3	5.18	4.12	5.18	4.12	5.18	4.12
1.5	5	50	0.3	3.21	2.68	3.60	2.95	2.95	2.58
5	5	50	0.3	3.00	2.54	3.55	2.99	2.95	2.62

C8.2: Three endpoints with $\rho_{ijj'} = -0.3$

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
control				weak	strong	weak	strong	weak	strong
1	5	10	0.3	5.10	4.91	5.10	4.91	5.10	4.91
1.5	5	10	0.3	2.70	2.51	3.72	3.48	3.04	2.86
5	5	10	0.3	3.02	2.35	3.67	2.74	3.06	2.24
1	10	10	0.3	4.74	4.96	4.74	4.96	4.74	4.96
1.5	10	10	0.3	2.50	2.59	3.15	3.24	2.52	2.61
5	10	10	0.3	2.74	2.64	3.10	3.00	2.41	2.29
1	30	10	0.3	5.05	4.91	5.05	4.91	5.05	4.91
1.5	30	10	0.3	2.60	2.40	2.92	2.80	2.29	2.26
5	30	10	0.3	2.72	2.55	2.94	2.82	2.30	2.24
1	5	50	0.3	5.09	4.12	5.09	4.12	5.09	4.12
1.5	5	50	0.3	3.30	2.67	3.71	2.95	3.04	2.50
5	5	50	0.3	3.02	2.57	3.66	3.02	3.06	2.55

Table C9: Parametric for relevant ratios - unbalanced design

The last table for the parametric tests on relevant ratios gives results of the empirical FWER for unbalanced designs. All scenarios are tested two-sided and the endpoints under H_0 receive a true ratio of means equal to one of the relevance thresholds.

$\theta_{lower}^{-1} = \theta_{upper}$	n_1	n_2	σ	$\rho_{ijj'}$	Sasabuchi selector		θ-shift		random$^{\theta}$	
control					weak	strong	weak	strong	weak	strong
1	15	5	10	0.1	5.11	5.05	5.11	5.05	5.11	5.05
1.5	15	5	10	0.1	2.50	2.53	3.00	2.93	2.59	2.64
5	15	5	10	0.1	2.60	2.39	3.03	2.73	2.52	2.34
1	5	15	10	0.1	5.07	5.03	5.07	5.03	5.07	5.03
1.5	5	15	10	0.1	2.54	2.67	3.02	3.22	2.52	2.43
5	5	15	10	0.1	2.49	2.49	2.85	2.85	2.59	2.34
1	15	5	10	0.9	5.18	4.99	5.18	4.99	5.18	4.99
1.5	15	5	10	0.9	2.87	2.90	4.83	4.86	4.45	4.67
5	15	5	10	0.9	4.01	4.06	4.64	4.71	4.56	4.67
1	5	15	10	0.9	4.67	4.93	4.67	4.93	4.67	4.93
1.5	5	15	10	0.9	2.72	2.82	4.64	4.76	4.71	4.75
5	5	15	10	0.9	4.11	4.11	4.65	4.72	4.49	4.55

As expected, the procedures control empirically the FWER in the analyzed experiments.

A.4 Nonparametric procedures to test for relevant ratios

The final section presents first the simulation results for the nonparametric procedure with a data-driven order of hypotheses (np-θ-shift), which use tests for relevant ratio and afterwards some results of the modified permutation algorithm (minP) are given. As discussed in chapter 6 these procedures have a different assumption on the data compared to the other nonparametric tests. In the one-sided case for endpoints under H_0 the expected values are set to $\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot \kappa^{H_0}$ and the true standard deviation is $\sigma_{1j} = 10$ and $\sigma_{2j} = 10 \cdot \kappa^{H_0}$, with κ^{H_0} denoting a random value between 1 and θ_{upper} in steps of 0.05 units. For two-sided testing the parameters are set to $\mu_{1j} = 100$, $\mu_{2j} = 100 \cdot \kappa^{H_0}$, $\sigma_{1j} = 10$ and $\sigma_{2j} = 10 \cdot \kappa^{H_0}$ (increase) and $\mu_{1j} = 100 \cdot (1/\kappa^{H_0})$, $\mu_{2j} = 100$, $\sigma_{1j} = 10 \cdot (1/\kappa^{H_0})$ and $\sigma_{2j} = 10$ (decrease), where κ^{H_0} denotes in this case a random value with $\theta_{lower} \leq \kappa^{H_0} \leq \theta_{upper}$. For the strong control and one-sided testing the endpoints under H_1 have parameters set as: $\mu_{1j} = 100$, $\mu_{2j} = 100 \cdot \theta_{upper} + 50$, $\sigma_{1j} = 10$ and $\sigma_{2j} = 10 \cdot \theta_{upper} + 5$. In the two-sided case the parameters as selected as: $\mu_{1j} = 100$, $\mu_{2j} = 100 \cdot \theta_{upper} + 50$, $\sigma_{1j} = 10$ and $\sigma_{2j} = 10 \cdot \theta_{upper} + 5$ (increase) and $\mu_{1j} = 100 \cdot (1/\theta_{lower}) + 50$, $\mu_{2j} = 100$, $\sigma_{1j} = 10 \cdot (1/\theta_{upper}) + 5$ and $\sigma_{2j} = 10$.

A.4.1 Procedure with a data-driven order of hypotheses

Table D1: Nonparametric for relevant ratios, FWER dependent on n_i

The first table shows the empirical FWER for exact and asymptotic testing for varying sample size per group. Endpoints under H_0 receive a random true ratio of means equal to one of the relevance thresholds. Only two-sided testing scenarios are observed.

assumption	$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak	strong
exact	1	4	10	0.1	2.99	2.45
	5	4	10	0.1	2.18	1.60
	1	7	10	0.1	3.73	3.82
	5	7	10	0.1	2.43	2.18
	1	15	10	0.1	4.68	4.61
	5	15	10	0.1	2.75	2.37
asymptotic	1	4	10	0.1	**5.85**	**5.37**
	5	4	10	0.1	4.27	3.37
	1	7	10	0.1	5.25	5.33
	5	7	10	0.1	3.37	3.00
	1	15	10	0.1	5.17	5.05
	5	15	10	0.1	3.04	2.66

Only for a sample size of 4 and $\theta_{lower} = \theta_{upper} = 1$ an empirical exceed of the FWER by the asymptotic versions can be found. However as the procedures with a data-driven order of hypotheses are superior for small sample sizes, the exact versions have to be used.

Table D2: Nonparametric for relevant ratios - asymptotic

However to show that the np-θ-shift method controls the FWER when the asymptotic versions are used on experiments with higher sample sizes, one table for the asymptotic methods is listed. The endpoints under H_0 have a random treatment effect as described in the beginning of this section.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak		strong	
side				one	two	one	two
1	10	10	0.1	**6.10**	5.23	**6.08**	5.23
1.5	10	10	0.1	0.95	0.51	0.86	0.47
5	10	10	0.1	0.29	0.12	0.21	0.09
1	30	10	0.1	**5.39**	5.14	**5.51**	4.89
1.5	30	10	0.1	0.56	0.28	1.96	0.38
5	30	10	0.1	0.11	0.05	0.76	0.06
1	10	10	0.9	5.03	5.13	**5.36**	5.09
1.5	10	10	0.9	2.01	1.66	1.96	1.65
5	10	10	0.9	0.87	0.58	0.76	0.49
1	30	10	0.9	5.20	5.13	5.25	5.04
1.5	30	10	0.9	1.16	0.76	1.22	0.78
5	30	10	0.9	0.26	0.17	0.27	0.18
1	30	5	0.1	**5.44**	5.11	**5.50**	4.92
1.5	30	5	0.1	0.57	0.29	0.58	0.36
5	30	5	0.1	0.08	0.03	0.11	0.06
1	30	15	0.1	**5.36**	5.09	**5.65**	4.90
1.5	30	15	0.1	0.61	0.32	0.63	0.38
5	30	15	0.1	0.16	0.05	0.16	0.08

If the sample sizes are not too small the behavior of the empirical error rate is similar to the former sections: only if tested one-sided and $\theta_{lower} = \theta_{upper} = 1$ slight exceeds appear.

Table D3: Nonparametric for relevant ratios - exact

The same behavior of the empirical FWER is observed for smaller sample sizes and the exact version:

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak		strong	
side				one	two	one	two
1	5	10	0.1	**6.25**	3.42	**6.25**	3.05
1.5	5	10	0.1	1.14	0.39	1.10	0.34
5	5	10	0.1	0.36	0.11	0.22	0.08
1	10	10	0.1	5.24	4.39	5.07	4.22
1.5	10	10	0.1	0.82	0.42	0.74	0.35
5	10	10	0.1	0.25	0.10	0.17	0.07
1	5	10	0.9	4.75	3.12	4.83	3.13
1.5	5	10	0.9	2.35	1.40	2.33	1.17
5	5	10	0.9	1.16	0.58	1.17	0.42
1	10	10	0.9	4.55	4.27	4.66	4.27
1.5	10	10	0.9	1.77	1.41	1.68	1.45
5	10	10	0.9	0.76	0.50	0.66	0.41
1	5	5	0.1	**6.30**	3.43	**6.29**	3.14
1.5	5	5	0.1	0.95	0.34	0.98	0.31
5	5	5	0.1	0.22	0.06	0.14	0.01
1	5	15	0.1	**6.27**	3.40	**5.84**	2.52
1.5	5	15	0.1	1.34	0.46	1.16	0.31
5	5	15	0.1	0.50	0.16	0.39	0.12

Table D4: Nonparametric for relevant ratios - weak and strong control for $\alpha =$ 1% and 10%

The next table gives results for the nominal FWER of 1% and 10%. All tested scenarios are two-sided and - as for all following simulations - only exact versions are used. For endpoints under H_0 the true ratio of means are a random value.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	$\alpha = 1\%$		$\alpha = 10\%$	
control				weak	strong	weak	strong
1	5	10	0.1	0.86	0.62	9.76	9.50
1.5	5	10	0.1	0.11	0.05	1.30	1.14
5	5	10	0.1	0.01	0.01	0.35	0.30
1	10	10	0.1	0.91	0.89	8.92	8.91
1.5	10	10	0.1	0.07	0.09	0.82	0.81
5	10	10	0.1	0.03	0	0.21	0.16
1	5	10	0.9	0.86	0.64	9.65	9.68
1.5	5	10	0.9	0.47	0.28	3.57	3.39
5	5	10	0.9	0.18	0.16	1.55	1.52
1	5	15	0.1	0.88	0.33	9.78	9.00
1.5	5	15	0.1	0.11	0.07	1.52	1.32
5	5	15	0.1	0.02	0.02	0.52	0.43

Table D5: Nonparametric for relevant ratios - varying number of endpoints under H_0

This table list the empirical FWER for varying a number of endpoints under H_0. It is basically the same as C5 but for the np-θ-shift procedure. Again all settings are two-sided.

relevance thresholds	number of endpoints under H_0 (out of 50)	empirical FWE
1 / 1	1	2.49
1 / 1	25	3.04
1 / 1	50	3.42
0.2 / 5	1	0
0.2 / 5	25	0.02
0.2 / 5	50	0.11

Table D6: Nonparametric for relevant ratios - ratio under H_0 equal to θ_{lower}, θ_{upper}
In the following table the data conditions vary. Three conditions are tested: in I the mean level of the data is as usual 100 and the variance is constant. In condition II the design proposed by Attoor *et al.* (2004) is used: the mean level is exponentially distributed and the coefficient of variation is set to a constant value among the endpoints; the data of the individual endpoints follow a multivariate normal distribution. Condition III shows the results for scenarios where the data is taken from a skewed distribution, with skewness 2 and kurtosis 7. Here the mean level of the data is 100 and the variance is constant.
In all settings the endpoints under H_0 have a true ratio of means equal to either θ_{lower} or θ_{upper} and it is tested two-sided.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ /	$\rho_{ijj'}$	condition I		condition II		condition III	
control		cv		weak	strong	weak	strong	weak	strong
1	5	10 / 0.1	0.1	3.42	3.05	2.93	3.22	3.31	2.87
1.5	5	10 / 0.1	0.1	2.53	1.98	1.98	2.18	2.40	1.57
5	5	10 / 0.1	0.1	2.53	1.87	1.98	2.03	2.40	1.78
1	10	10 / 0.1	0.1	4.39	4.22	4.36	4.43	4.15	4.32
1.5	10	10 / 0.1	0.1	2.70	2.81	2.73	2.66	2.48	2.64
5	10	10 / 0.1	0.1	2.70	2.48	2.73	2.33	2.48	2.37
1	5	10 / 0.1	0.9	3.12	3.13	3.28	3.09	3.19	2.96
1.5	5	10 / 0.1	0.9	2.83	2.61	2.90	2.46	2.78	2.45
5	5	10 / 0.1	0.9	2.83	2.80	2.90	2.65	2.78	2.64
1	5	15 / 0.15	0.1	3.40	2.52	2.97	2.66	3.27	2.41
1.5	5	15 / 0.15	0.1	2.48	1.73	1.98	1.91	2.40	1.50
5	5	15 / 0.15	0.1	2.48	1.98	1.98	2.12	2.40	1.83

Table D7: Nonparametric for relevant ratios - varying correlation of endpoints under H_0

The table is based on the tables of C8, where 47 endpoints have a true correlation of 0.3 and for the remaining three variables the correlation is set to a different value. The expected values of the endpoints under H_1 (strong control) are chosen as given in the introduction of this section and the endpoints under H_0 have a true ratio of means equal to one of the relevance thresholds. All settings are two-sided.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	$\rho_{ijj'} = 0.01$		$\rho_{ijj'} = -0.3$	
control				weak	strong	weak	strong
1	5	10	0.3	3.18	3.19	3.36	3.17
1.5	5	10	0.3	2.56	2.14	2.50	2.09
5	5	10	0.3	2.56	2.00	2.50	1.98
1	10	10	0.3	4.18	4.50	4.12	4.29
1.5	10	10	0.3	2.86	2.78	2.66	2.98
5	10	10	0.3	2.86	2.53	2.66	2.66
1	5	15	0.3	3.14	2.71	3.34	2.61
1.5	5	15	0.3	2.47	1.94	2.50	1.90
5	5	15	0.3	2.47	2.07	2.50	2.08

Table D8: Nonparametric for relevant ratios - unbalanced design

In the following table results of the empirical FWER for unbalanced designs are given. All scenarios are tested two-sided and the endpoints under H_0 receive a true ratio of means equal to one of the relevance thresholds.

$\theta_{lower}^{-1} = \theta_{upper}$	n_1	n_2	σ	$\rho_{ijj'}$	weak	strong
1	15	5	10	0.1	4.18	4.12
1.5	15	5	10	0.1	2.81	2.54
5	15	5	10	0.1	2.81	2.54
1	5	15	10	0.1	4.30	4.31
1.5	5	15	10	0.1	2.75	2.67
5	5	15	10	0.1	2.75	2.28
1	15	5	10	0.9	4.17	4.35
1.5	15	5	10	0.9	3.60	3.47
5	15	5	10	0.9	3.60	3.44
1	5	15	10	0.9	4.05	4.00
1.5	5	15	10	0.9	3.47	3.39
5	5	15	10	0.9	3.47	3.37

A.4.2 Relevance-shifted permutation algorithm

Table D9: minP - FWER for various condition using Gaussian distributed data

The first table of the FWER results corresponding to the relevance-shifted permutation algorithm shows scenarios for various conditions by use of Gaussian distributed data. All settings of the parameters are listed in table. Results are given for the weak and the strong control of the FWER. For half of the scenarios the expected values are set as described in the introduction. Simulation results for this random ratio of the means are printed in the first two columns including empirical FWEs. In the last two columns the ratio of means is set to one of the margins of the null hypothesis. That is, the ratio is equal to one of the thresholds. All scenarios in this table and the further ones are analyzed with two-sided tests. If not stated otherwise, an error rate of 5% is chosen.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	random		margin	
control				weak	strong	weak	strong
1	5	10	0.1	0	0	0	0
1.5	5	10	0.1	0	0	0	0
5	5	10	0.1	0.05	0.08	0	0
1	10	10	0.1	4.65	4.12	4.65	4.12
1.5	10	10	0.1	0.53	0.56	3.30	3.14
5	10	10	0.1	0.15	0.11	3.30	3.01
1	5	10	0.9	1.42	0.91	1.42	0.91
1.5	5	10	0.9	0.02	0	1.37	1.45
5	5	10	0.9	0.22	0.16	1.37	1.44
1	10	10	0.9	4.77	3.97	4.77	3.97
1.5	10	10	0.9	0.92	0.87	3.37	3.13
5	10	10	0.9	0.35	0.35	3.37	3.13
1	5	5	0.9	1.42	1.01	1.42	1.01
1.5	5	5	0.9	0.16	0.27	1.37	1.45
5	5	5	0.9	0.29	0.28	1.37	1.45
1	5	15	0.9	1.42	0.66	1.42	0.66
1.5	5	15	0.9	0	0	1.27	1.35
5	5	15	0.9	0.06	0.01	1.37	1.43

Table D10: minP - FWER for various condition using skewed distributed data
This table gives empirical FWERs for scenarios with random numbers taken from a skewed distribution with a skewness of 2 and a kurtosis of 7. Beside the non-normal distributed data, the tested experiments are exactly the same as the ones analyzed in the last two columns of the former table.

$\theta_{lower}^{-1} = \theta_{upper}$	n_i	σ	$\rho_{ijj'}$	weak control	strong control
1	5	10	0.1	0	0
1.5	5	10	0.1	0	0
5	5	10	0.1	0	0
1	10	10	0.1	4.88	4.18
1.5	10	10	0.1	3.46	3.42
5	10	10	0.1	3.46	3.15
1	5	10	0.9	0.80	0.39
1.5	5	10	0.9	0.83	1.12
5	5	10	0.9	0.84	1.07
1	10	10	0.9	4.65	4.25
1.5	10	10	0.9	3.26	0.33
5	10	10	0.9	3.26	0.33
1	5	5	0.9	0.80	0.46
1.5	5	5	0.9	0.84	1.12
5	5	5	0.9	0.84	1.12
1	5	15	0.9	0.80	0.31
1.5	5	15	0.9	0.80	1.09
5	5	15	0.9	0.84	1.03

Table D11: minP - unbalanced design, large sample size, varying α

In the last table various scenarios, such as unbalanced data, larger sample sizes and the nominal FWER levels of 1% and 10% are tested. All ratio of medians corresponding to endpoints under H_0 are set to one margin of the null hypothesis.

nominal FWER [%]	$\theta_{lower}^{-1} = \theta_{upper}$	n_1	n_2	σ	$\rho_{ijj'}$	weak control	strong control
5	1	15	5	10	0.9	4.94	3.82
5	1.5	15	5	10	0.9	2.92	2.91
5	5	15	5	10	0.9	2.92	2.84
5	1	5	15	10	0.9	4.64	3.79
5	1.5	5	15	10	0.9	2.70	2.69
5	5	5	15	10	0.9	2.70	2.58
5	1	10	10	10	0.5	4.74	4.61
5	1.5	10	10	10	0.5	3.36	3.71
5	5	10	10	10	0.5	3.36	3.59
1	1	10	10	10	0.9	0.89	0.79
1	1.5	10	10	10	0.9	0.66	0.59
1	5	10	10	10	0.9	0.66	0.59
10	1	10	10	10	0.9	9.96	8.13
10	1.5	10	10	10	0.9	6.37	6.48
10	5	10	10	10	0.9	6.37	6.48
5	1	20	20	10	0.9	4.90	4.58
5	1.5	20	20	10	0.9	3.29	3.64
5	5	20	20	10	0.9	3.29	3.64

In all tables corresponding to the minP algorithm the empirical FWER is less than the nominal error rate.

Eine Promotion ist ohne Hilfe und Unterstützung nicht zu leisten. Viele haben mir auf die eine oder andere Weise geholfen und zum Gelingen beigetragen. Ihnen gilt mein Dank:

Mein Doktorvater Prof. Dr. Ludwig Hothorn hat mir in allen Phasen der Arbeit beigestanden, mir viele wertvolle Ratschläge gegeben und mich immer wieder ermutigt. Allzu oft hat er meine Arbeit zurecht (zu Recht) rücken müssen. Fortwährende Unterstützung bekam ich von Dr. Siegfried Kropf, an den ich mich mit allen Fragen und Problemen wenden konnte und der mir stets weiterhalf. Dr. Frank Bretz war mir ein wichtiger Gesprächspartner, der mir in vielen Diskussionen hilfreiche Anregungen gab.
Einen großen Rückhalt erfuhr ich bei meinen Kollegen im Lehrgebiet Biostatistik. Hannelore Visser und Clemens Buczilowski halfen mir nicht nur bei den vielen (verwaltungs-) technischen Fragen, sondern waren oft auch eine moralische Unterstützung bei der Arbeit. Mit meinen Kollegen Dr. Gemechis Dilba, Frank Schaarschmidt, Mario Hasler und Donghui Ma stand ich in regem Austausch, was die alltäglichen Kleinigkeiten betrifft, nicht nur in Lehre und Forschung.
In den stressigsten Zeiten dieser Arbeit fand ich guten Rat und schnelle Hilfe: bei Dr. Peter Westfall durch seine ebenso aufmunternden wie hilfreichen Kommentare zum Permutationsalgorithmus und bei Dr. Markus Eszlinger und Dr. Anthony Passerini durch die großzügige Bereitstellung von experimentellen Daten und den dazugehörigen Informationen.
Viele sind während der Doktorarbeit zu kurz gekommen, und trotzdem haben sie mir immer wieder den Rücken freigehalten und gestärkt: meine Freunde, die viel Verständnis für meine wenige freie Zeit hatten und mich trotz Laptop und Artikel im Gepäck mit in den Urlaub nahmen. Meine Eltern Christiane und Josef Schratz und meine Schwester Sabine Schratz, die mir unermüdlich und geduldig zur Seite standen. Und schließlich und vor allem mein Ehemann Frank und meine Tochter Lea, die mich in den Tiefpunkten meiner Arbeit getragen haben.

Ihnen allen sei an dieser Stelle noch einmal sehr herzlich gedankt!