Many-to-one comparisons in stratified designs

Vom Fachbereich Gartenbau
der Universität Hannover
zur Erlangung
des akademischen Grades eines

Doktors der Gartenbauwissenschaften
– Dr. rer. hort. –

genehmigte

Dissertation

von

Egbertus Hendrikus Evert Biesheuvel
geboren am 06.01.1967 in Kampen

2002

To my parents

Referent:	Prof. Dr. L.A. Hothorn
Referentin:	Prof. Dr. I. Pigeot
Tag der Promotion:	24 Juni 2002

Zusammenfassung

Gegenstand dieser Dissertation war die Untersuchung multipler Testprozeduren für Many-One-Vergleiche in einem stratifizierten Design unter strikter Einhaltung der globale Irrtumswahrscheinlichkeit auf dem Niveau a. Das Problem des simultanen Vergleichs mehrerer aktiver Behandlungsgruppen mit einer Kontrollgruppe in jeder der verschiedenen Schichten tritt in unterschiedlichen praktischen Situationen auf, wie die Beispiele in der Einleitung dieser Arbeit zeigen. Ein naiver Ansatz würde darin bestehen, das Dunnett-Verfahren innerhalb jeder Schicht anzuwenden, ohne eine weitere Fehlerkorrektur für multiples Testen vorzunehmen. Dies würde zu einer Inflation des multiplen Niveaus a des Gesamtexperiments führen. Andererseits würde eine Bonferroni-Korrektur zur Berücksichtung der Anzahl der Strata auf einen konservativen Ansatz hinauslaufen, unter der in dieser Arbeit getroffenen Annahme einer unbekannten gemeinsamen Varianz.

Cheung und Holland (1992) erweiterten das Dunnett-Verfahren auf die stratifizierte Situation. Allerdings leiteten sie lediglich obere Perzentile für einen gemeinsamen Korrelationskoeffizienten ab und schlugen vor, diese Perzentile für alle anderen Testsituationen zu interpolieren. Die vorliegende Arbeit weist nach, daß diese Approximationen nicht mehr erforderlich sind und daß korrekte Perzentile heute einfach mit verfügbarer Software (SAS) berechnet werden können.

Darüberhinaus beschreibt diese Arbeit, wie Güte-Berechnungen und Fallzahlschätzungen durchgeführt werden können, was bei Cheung und Holland nicht betrachtet worden war.

Obwohl bei den meisten Many-One-Vergleichen in praktischen Testsituationen das Interesse darin liegt, die Überlegenheit bzw. den Unterschied einer aktiven Behandlung zu einer Kontrollbehandlung nachzuweisen, gibt es Testsituationen, wo diese Art von Fragestellungen nicht adäquat ist. In dieser Arbeit wird gezeigt, daß es auch möglich ist, Many-One-Vergleiche in einem stratifizierten Design im Fall eines Nicht-Unterlegenheit-Testproblems oder globalen Äquivalenzproblems unter Einhaltung des multiplen Niveaus a durchzuführen.

Auch für den Fall, daß sich das Testproblem besser mittels Verhältnisraten als durch Differenzen beschreiben läßt, wird in der vorliegenden Arbeit gezeigt, wie Many-One-Vergleiche in einem stratifizierten Versuch durchgeführt werden können.

Alle diese Verfahren treffen die Annahme der Normalverteilung der Daten. Wenn diese Annahme zweifelhaft ist, kann die Verwendung nichtparametrischer Verfahren eher angebracht sein. Munzel und Hothorn (2001) diskutieren einen asymptotischen Ansatz zur Durchführung von Many-One-Vergleichen im Ein-Weg-Design basierend auf einer paarweisen Rangvergabe-Prozedur. Die vorliegende Arbeit illustriert, wie dieses Testverfahren auf den Fall des stratifizierten Zwei-Weg-Designs erweitert werden kann.

Schließlich werden die Methode der stochastischen Approximation, die Bootstrap-Methode und die Permutationsmethode als alternative Methoden diskutiert. Diese drei Methoden werden anhand einer Situation veranschaulicht, für die diese computerintensiven Resampling-Methoden standardmäßig innerhalb der SAS-Software verfügbar sind.

Zusammenfassend zeigt die vorliegende Arbeit, daß es möglich ist Many-One-Vergleiche in einem stratifizierten Zwei-Weg-Design für verschiedene praktische Testsituationen durchzuführen, und stellt den erforderlichen Programm-Code zur Analyse dieser Testprobleme bereit.

Schlagwörter: Many-One-Vergleiche; stratifizierten Design

CONTENTS

1 Introduction

To examine treatment effects in scientific experiments, multiple comparison procedures are useful and popular techniques in various disciplines such as medicine and agrobiology (see for example Hochberg and Tamhane (1987) and references therein). However, as many writers indicate, one has to be very cautious when simultaneous inferences are implemented because he or she may not be aware of the multiplicity effect (Westfall and Young, 1993). As explained by Tukey (1977), when a large data set undergoes extensive data splitting without careful control of the overall error rate, 'false significance' can easily result. For instance, in multiple hypotheses testing, the probability of making at least one false rejection among all the hypotheses being considered can be substantial even though each individual hypothesis is tested with a small α level. Hence to tackle the multiplicity problem, some researchers prefer multiple comparison procedures that are designed to control the familywise error rate (FWE) as defined in Hochberg and Tamhane (1987). A multiple comparison procedure is said to control the FWE in stronge sense if it protects the FWE under all configurations of the null hypothesis and in the weak sense if it controls the FWE under the complete null configuration. However, the control of the FWE may not be necessary in some cases. For a more in-depth discussion of which error rate to control in multiple comparison problems, one can read Chapter 1 of Hochberg and Tamhane (1987).

It is clear that no adjustment for multiple comparisons at all will result in the smallest p-value. On the other hand adjusting for multiple comparisons and incorporating the correlation structure results in a smaller adjusted p-value than Bonferroni style adjustments.

Dunnett (1955) mentioned the common problem in applied research of the comparison of treatments with a control or a standard: '*Such a situation may arise, for example, when an agronomist tests the effects on crop yield of the addition of chemicals to the soil, or when a pharmacologist assays drug sample to determine their potencies. In designing an experiment to measure the effects of such treatments, it is often desirable to include in the experiment a control in the form of either a dummy treatment, to measure the magnitude of the experimental response in the absence of the treatments under investigation, or some recognized standard treatment.*' In his paper, Dunnett described his well known and widely used multiple comparison procedure for simultaneously comparing, by interval estimation or hypothesis testing, all active treatments with a control when sampling from a distribution where the normality assumption is reasonable. Multiple comparisons to a control (MCC) are also referred to as many-to-one comparisons.

The problem of multiple comparisons with a control is a special case of the more general multiple comparisons problem considered by Tukey (1953) and Scheffé (1953). Tukey's procedure based on the Studentized range and Scheffé's procedure based on the F-distribution enables the experimenter to make any number of comparisons among a set of sample means with the assurance that the probability of all confidence statements being correct will be equal to or greater than a specified value. When the experimenter only wishes to make comparisons between one of the means and each of the others, as is the case when one of the means represents a control, use of the Tukey or Scheffé procedure would result in larger p-values and in wider confidence limits than necessary.

Dunnett's procedure is tailored for the one-way layout situation, i.e. a single stratum, and the issue is how one should undertake the comparisons of active treatments with a control in a stratified design. Direct application of Dunnett's procedure might be inappropriate in this situation.

The basic question to consider is how the experimenter should control the familywise error rate:

(a) for active treatments versus control averaged over all strata,

(b) for active treatments versus control separately in each stratum, or

(c) for active treatments versus control globally across all strata.

Situation (a) implies direct application of Dunnett's original procedure.

Situation (b) implies the conduct of a separate Dunnett procedure within each stratum, sometimes called the 'Dunnett-within-group' procedure. Notice that the relevant family of hypotheses under consideration is the set of hypotheses comparing active treatments with control within each of the strata. So there are a number of families of hypotheses in total; for each stratum there is a family of hypotheses.

Situation (c) implies the comparisons of all active treatments with a control within each of the strata simultaneously while controlling the FWE. The relevant family of hypotheses under consideration is the set of all treatment versus control hypotheses in the overall experiment. This can be seen as an extension of Dunnett's multiple comparison procedure to a stratified design.

Cheung and Holland (1991) extended the Dunnett procedure for comparing all active treatments with a control for the one-way layout to instances where it is desired to make such comparisons simultaneously within each of several strata while holding the probability of making any Type I errors at a designated level a in case of a common sample size for any of the stratum-treatment combinations. In 1992 they described this procedure allowing different sample sizes for each of the stratum-treatment categories.

This situation occurs in the setting of a fixed effects two-way factorial layout where one factor has one level that is a control or otherwise specially designated level, among with several active treatment levels, and it is desired to make comparisons between all active treatments with control at each of the levels of a second factor.

To illustrate the situation in a practical setting, some examples are considered.

Oat yield data

Steel and Torrie (1980) presented the results of an experiment to compare the yields of four oat seed lots (strata) following three chemical seed treatments and a control. Two of the seed lots were Vicland; once infected with H.Victoriae (1), once not (2). The experimental design was a split-plot layout with seed lots as whole plots and treatments as subplots, incorporating a randomized complete block design with four-blocks. Yields were measured in bushels per acre. The data are presented in the following table.

Table 1.1 Oat yield data from Steel and Torrie

	Treatment			
Seed lots	Control	Ceresan M	Panogen	Agrox
Vicland (1)	36.1	50.6	45.9	37.3
Vicland (2)	50.9	55.4	53.1	54.3
Clinton	53.9	51.4	55.9	56.1
Branch	61.9	63.4	57.7	61.3

Each cell represents the mean of four observations, in bushels per acre.

The authors analyzed the data with a separate Dunnett procedure for each of the seed lots, i.e. the ‘Dunnett within-group’ procedure. One-sided tests were used because it was expected a priori that treatments would increase yield. The use of ‘Dunnett within-group’ procedure seems to be justified if the purpose of the experiment was to make treatment recommendations for many farmers, each of whom uses only one seed lots. However, if the purpose was to advise a single farmer who uses all four seed lots which seed treatment to use with each seed lot, the extended Dunnett procedure for the stratified design seems to be preferable.

Animal myocardial infarction data

Jugdutt (1988) described the data of an experiment to study the effects of nitroglycerin and ibuprofen on left ventricular topography during healing after myocardial infarction induced in dogs. One group of dogs in which no infarction was induced served as sham and another group of dogs in the dogs with infarction served as the control treatment group. The two strata consisted of the group of dogs measured at one week post-occlusion and at six weeks post-occlusion. The following table shows a part of the results of this experiment.

Table 1.2 Infarct size data from Jugdutt

	Treatment			
Post-occlusion	Control	Nitroglycerin	Ibuprofen	Sham
1-week	37.4	33.8	26.2	24.2
	(8)	(6)	(6)	(12)
6-week	32.5	31.6	29.1	24.2
	(7)	(9)	(8)	(10)

Each cell represents the mean of occluded bed size in percentages. The number of observations is in brackets.

The authors analyzed the data in several ways, including the comparisons of each treatment group at 1 week with 6 weeks and each treatment group versus the control group for both time points separately. Despite these analyses, the data could have been analyzed by the extended Dunnett procedure allowing different sample sizes. Depending on the research question the control group could have been defined as the control treatment but also by the ‘sham treatment’ group.

In vivo bone marrow cell data

Morales-Ramírez and García-Rodríguez (1994) studied the radioprotective capacity of three dosages of chlorophyllin on γ-ray-induced sister chromatid exchange (SCE) in murine bone marrow cells in vivo. The group of mice was divided in two; one group was exposed to ionizing radiation (stratum 1), which is capable of inducing SCE and the other group was not exposed (stratum 2). The following table shows some pooled results of two separate identical experiments, which were considered as one experiment by the authors.

Table 1.3 SCE induction data from Morales-Ramírez and García-Rodríguez

	Treatment			
Radiation	Control	Chlorophyllin 100 µg	Chlorophyllin 50 µg	Chlorophyllin 10 µg
Yes	4.5 (14)	3.5 (8)	4.0 (8)	4.5 (7)
No	3.4 (15)	3.5 (8)	3.6 (8)	3.6 (8)

Each cell represents the mean of SCE per cell. The number of observations is in brackets.

The authors did the statistical evaluation to compare the different dosages of chlorophyllin against the control group with '*Dunnett's test for several groups and different sample sizes (Cheung and Holland, 1992)*'.

Human erythrocytes data

Trevisan et al. (1986) studied the intra-erythrocytic cation metabolism in ureamic patients on different dialysis treatments. The patients in this study underwent two different treatments, regular haemodialysis and continuous ambulatory peritoneal dialysis (CAPD). Also, from a survey on cellular ion transport and hypertension, 67 persons were randomly selected as controls. The subjects were classified according to gender. Blood was drawn from each subject and one of the responses was the haemoglobin content (g/l) as shown in the following table.

Table 1.4 Heamoglobin content data from Trevisan et al.

	Treatment		
Gender	Control	Haemodialysis	CAPD
Males	15.6 (35)	8.8 (18)	10.8 (14)
Females	14.2 (32)	9.4 (16)	10.2 (10)

Each cell represents the mean in g/l. The number of observations is in brackets.

The test of treatment by gender interaction was significant. Therefore it seems to be appropriate to compare both active treatments with control separately for males and females with the extended Dunnett procedure for the two-way layout allowing different sample sizes.

These examples illustrate the topic of this thesis, to describe multiple comparisons procedures for many-to-one comparisons in a stratified design while maintaining the FWE at a designated level a .

The original Dunnett's multiple comparisons procedure for simultaneously comparing all active treatments with a control for a one-way layout assuming normal distributed data is reviewed in Chapter 2.

Chapter 3 and Chapter 4 describe the extended Dunnett's procedure for a stratified two-way layout situation in case of a one-sided alternative testing problem and a two-sided alternative testing problem respectively. The computation of adjusted p-values and simultaneous confidence intervals is discussed as well as the calculation of the different kinds of power and computation of sample sizes.

In practice there are also applications where the control treatment is a well-known standard treatment or concurrent treatment instead of a real placebo. In that particular situation one wants to test for non-inferiority instead of superiority or even to test for equivalence of the active treatments versus the control treatment. How to perform the many-to-one comparisons in those settings is examined in Chapter 5.

So far the testing problems are all formulated in terms of differences between the population means of the active treatment and the control treatment. However, there are also testing situations where it is more common to express the testing problem in proportions rather than differences while the normality assumption for the original variable is still justified. These kind of testing problems are considered in Chapter 6.

Chapter 7 relaxes the assumption of normality and discusses a nonparametric procedure to perform the many-to-one comparisons in a stratified two-way layout. The procedure is based on pairwise rankings and relies on asymptotic results.

A flavor how resampling methods can be applied is discussed in Chapter 8. The resampling methods considered are a stochastic approximation method, the bootstrap method and the permutation method. The methods are illustrated for situations that can be handled by standard available software, i.e. standard procedures available within the SAS software system.

A summary and outlook is given in Chapter 9.

The following table gives a quick reference to the corresponding chapters:

Table 1.5 Overview of chapters

Many-to-one comparisons		Chapter
Original Dunnett's procedure for a one-way layout		2
Stratified two-way layout		
• Testing problem in terms of differences assuming normality	One-sided superiority testing	3
	Two-sided inequality testing	4
	Equivalence testing	5
• Testing problem in terms of ratios assuming normality		6
• Nonparametric procedure based on pairwise rankings		7
• Standard resampling methods (stochastic approximation, bootstrap, permutation)		8

2 Dunnett's procedure

This chapter reviews the multiple comparison procedure proposed by Dunnett (1955) for comparing several treatments with a control in the situation of a one-way design when the observations are assumed to be independently and normally distributed with a common standard deviation. Since each comparison has the same control in common, the procedure incorporates the dependencies between these comparisons. Dunnett's procedure is based on the multivariate Student t-distribution and maintains the familywise error rate at a prespecified level. It is useful to review the Dunnett procedure for the one-way layout before discussing the situation of a stratified two-way layout, because a good understanding of the Dunnett procedure for the one-way situation is helpful to understand the stratified situation. It will be illustrated in the next chapter that the many-to-one comparisons procedure for the stratified two-way layout situation is a rather straightforward extension of the original Dunnett procedure. Hence, the original Dunnett procedure can be considered as a special case of the stratified Dunnett procedure.

The first section introduces some general notation and the test statistic. The second section points out how upper percentage points of the test statistic can be calculated. The last section shows the derivations of multiplicity adjusted p-values and simultaneous confidence intervals based on Dunnett's multiple testing procedure.

2.1 Notation and test statistic

Before introducing the model, the hypothesis and the test statistic we provide the example as used by Dunnett in his 1955 paper is provided to illustrate the approach throughout this chapter.

The following data are blood count measurements on three groups of animals, one of which served as a control while the other two groups were treated with active drugs. Due to accidental losses, the numbers of animals in the three groups are unequal.

Table 2.1 Blood counts (millions of cells per cubic millimeter)

	Treatment		
	Control	Drug A	Drug B
	7.40	9.76	12.80
	8.50	8.80	9.68
	7.20	7.68	12.16
	8.24	9.36	9.20
	9.84		10.55
	8.32		
Sum	49.50	35.60	54.39
N	6	4	5
Mean	8.25	8.90	10.88

The interest of the experimenter was to compare both drug A and drug B with the control.

Notation

Suppose the following fixed effect one-way layout model

$$X_{jk} = m_j + e_{jk} \qquad j = 0, 1, \ldots, c \text{ and } k = 1, \ldots, n_j \text{ (> 0 for all } j) \qquad (2.1)$$

where *j* = 0 denotes the control treatment and the other *c* active treatments are labeled by *j* = 1 to *c* respectively. There are n_0 observations on the control and n_j observations on active treatment *j*.

Assume that the sample values $\{X_{jk}\}$ are identically and independently normal distributed with unknown means $m_0, m_1, \ldots, m_c$ and an unknown common variance s^2, i.e. $X_{jk} \sim N(m_j, s^2)$.

Let $\bar{X}_j = \sum_{k=1}^{n_j} X_{jk}$ denotes the sample mean (*j* = 0, 1, …, *c*) and let $s^2 = \dfrac{\sum_{j=0}^{c}\sum_{k=1}^{n_j}(X_{jk} - \bar{X}_j)^2}{n}$ be the usual pooled variance estimator of s^2 based on $n = \sum_{j=0}^{c} n_j - (c+1)$ degrees of freedom, which is independent of the sample means $\bar{X}_j$.

The aim is to test the null hypothesis of no effect between any of the *c* active treatments versus control against the one-sided alternative hypothesis that there exists an active treatment, which is superior to control, i.e.

$$H_0 : m_j = m_0 \qquad (j = 1, \ldots, c) \tag{2.2}$$

$$H_1 : \exists j : m_j > m_0 \qquad (j = 1, \ldots, c)$$

Assuming that a higher treatment mean m_j implies an improvement.

In case a lower treatment effect implies improvement and superiority should be demonstrated by showing that $m_j < m_0$, one should use the negative values to end up with the current settings.

Or in case of the two-sided alternative hypothesis that there exists an active treatment, which is different from control, the test situation is as follows

$$H_0 : m_j = m_0 \qquad (j = 1, \ldots, c) \tag{2.3}$$

$$H_1 : \exists j : m_j \neq m_0 \qquad (j = 1, \ldots, c)$$

Similar to the test situation where only one active treatment is compared with control (*c* = 1), Dunnett (1955) proposed to consider the statistics

$$D_j = \frac{\overline{X}_j - \overline{X}_0}{s\sqrt{n_j^{-1} + n_0^{-1}}} \qquad (j = 1, \ldots, c). \tag{2.4}$$

To test the global null hypothesis the test statistic

$$D = \max_{1 \le j \le c} \left\{ D_j \right\} \tag{2.5}$$

is proposed for the one-sided alternative hypothesis, and

$$D = \max_{1 \le j \le c} \left\{ \left| D_j \right| \right\} \tag{2.6}$$

in case of a two-sided alternative hypothesis.

As the distribution of $\left\{\frac{\bar{X}_1-\bar{X}_0}{\sqrt{n_1^{-1}+n_0^{-1}}}, \dots, \frac{\bar{X}_c-\bar{X}_0}{\sqrt{n_c^{-1}+n_0^{-1}}}\right\}$ is multivariate normal under the null hypothesis, the joint distribution of the D_j's is a central c-variate Student t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}=\left\{r_{j_1,j_2}\right\}$, denoted as $\left(D_1,D_2,...,D_c\right)' \sim t_c(n,\mathbf{R})$. (Cornish (1954); see also Appendix 1 for more details about the multivariate normal and multivariate t-distribution.)

The entries of $\mathbf{R}$ consist of the correlation between each pair of D_{j_1} and D_{j_1} $(1\le \mathrm{j}_1 \ne j_1 \le c)$ and is given by

$$r_{j_1,j_2}=\sqrt{\frac{n_{j_1}}{n_0+n_{j_1}}}\sqrt{\frac{n_{j_2}}{n_0+n_{j_2}}}=b_{j_1}b_{j_2} \quad (1\le \mathrm{j}_1 \ne j_2 \le c) \tag{2.7}$$

where

$$b_j=\sqrt{\frac{n_j}{n_0+n_j}}. \tag{2.8}$$

Proof:

$$r_{j_1,j_2}=corr\left(D_{j_1},D_{j_2}\right)=corr\left(\bar{X}_{j_1}-\bar{X}_0,\bar{X}_{j_2}-\bar{X}_0\right)=\frac{\operatorname{cov}\left(\bar{X}_{j_1}-\bar{X}_0,\bar{X}_{j_2}-\bar{X}_0\right)}{\sqrt{\operatorname{var}\left(\bar{X}_{j_1}-\bar{X}_0\right)}\sqrt{\operatorname{var}\left(\bar{X}_{j_2}-\bar{X}_0\right)}}=$$

$$=\frac{\operatorname{var}\left(\bar{X}_0\right)}{\sqrt{\operatorname{var}\left(\bar{X}_{j_1}\right)+\operatorname{var}\left(\bar{X}_0\right)}\sqrt{\operatorname{var}\left(\bar{X}_{j_2}\right)+\operatorname{var}\left(\bar{X}_0\right)}}=$$

$$=\frac{s^2 n_0^{-1}}{\sqrt{s^2\left\{n_{j_1}^{-1}+n_0^{-1}\right\}}\sqrt{s^2\left\{n_{j_1}^{-1}+n_0^{-1}\right\}}}=\sqrt{\frac{n_{j_1}}{n_0+n_{j_1}}}\sqrt{\frac{n_{j_2}}{n_0+n_{j_2}}}$$

A correlation matrix with the special correlation structure $r_{j_1,j_2}=b_{j_1}b_{j_2}$ is said to have the so-called product correlation structure. In the next section it becomes clear that this property simplifies the computations. (See also Appendix 1 for further details.)

The test procedure that rejects the global null hypothesis in favor of the alternative hypothesis if $D > d_a$, where d_a is chosen such that $P_{H_0}(D > d_a) = a$ controls the Type I error rate.

The calculations of p-values and upper percentage points of the distribution will be shown in the next section.

The introduction of the maximum of the D_j's as the test statistic to test the global null hypothesis H_0 might be somewhat artificial in first instance. However, the test statistic arises also in a natural way if one considers the global hypothesis-testing problem as a finite intersection of sub-hypotheses on testing each of the components.

Consider the finite family of *c* individual sub-hypotheses

$$H_{0j} : m_j = m_0$$

against the one-sided alternatives (2.9)

$$H_{1j} : m_j > m_0 .$$

Clearly, the global null hypothesis H_0 consists of the intersection of all sub null hypotheses H_{0j}, i.e. $H_0 = \bigcap_j H_{0j}$ and the alternative hypothesis H_1 is the combination or union of all sub-hypotheses H_{1j}, i.e. $H_1 = \bigcup_j H_{1j}$, in case of a one-sided global testing situation.

So testing the global hypothesis is now represented as what is called a 'Union-Intersection' (UI) multiple testing problem.

The test statistic D_j is used for testing the sub-hypothesis H_{0j} versus the alternative H_{1j} and H_{0j} is rejected if and only if D_j exceeds say x_{a_j}. According to the UI method of Roy (1953), the rejection region for H_0 is given by the union of rejection regions for the H_{0j} $j \in$ I, that is, H_0 is rejected if and only if at least one H_{0j} $j \in$ I, is rejected.

Given this, the critical constants x_{a_j} can then be determined as follows:

$$P\left(\text{reject } H_0\right) = a \Leftrightarrow P\left(\text{reject at least one } H_{0j}\right) = a \Leftrightarrow P\left(D_j > x_{a_j} \text{ for at least one } j\right) = a$$

Notice that there are several configurations of the critical values x_{a_j} that fulfill this requirement. However, usually the x_{a_j}'s are chosen to be identical, i.e. $x_{a_j} = x_a$. In general the sub-testing problems are generally treated symmetrically with regard to the relative importance of Type I errors. This implies that the marginal levels $a_j = P_{H_{0j}}\left(D_j > x_{a_j}\right)$ should be the same for all *j*. Since the D_j's have the same marginal distribution under the H_{0j}'s, it follows that the x_{a_j}'s should be equal.

In addition it also simplifies the task of computing.

So by letting $x_{a_j} = x_a$ for all *j* it follows that H_0 is rejected if $\max_j\left\{D_j\right\} > x_a$, where x_a should be chosen such that $P_{H_0}\left(\max_j\left\{D_j\right\} > x_a\right) = a$.

Similar statements hold true for the two-sided test situation.

Another way to see that the introduction of the maximum of the D_j's is quit natural is by using some theory: notice that the testing procedure, which rejects H_{0j} if $D_j > x_a$ compares each test statistic with the same common critical value. Such a testing procedure is what Gabriel (1969) has called a simultaneous test procedure. Notice further that the testing family $\left\{\left(D_j, H_{0j}\right) \middle| j = 1,\ldots,c\right\}$ is joint, i.e. the distributions of the D_j's $\left(j \in \tilde{J}\right)$ is completely specified under $\bigcap_{j\in\tilde{J}} H_{0j}$, where $\tilde{J} \subseteq \left\{1,\ldots c\right\}$. Then Gabriel showed that this simultaneous test procedure controls the FWE strongly if x_a is chosen such that $P_{H_0}\left(D > d_a\right) = a$ and D is defined as $\max_j\left\{D_j\right\}$. (See also Appendix 1 of Hochberg and Tamhane (1987))

Computations of the example

For the example described above, the test situation is as follows:

$$H_0 : m_0 = m_A = m_B$$
$$H_1 : m_A > m_0 \text{ or } m_B > m_0$$

in case of the one-sided alternative hypothesis or

$$H_0 : m_0 = m_A = m_B$$
$$H_1 : m_A \neq m_0 \text{ or } m_B \neq m_0$$

in case of the two-sided test situation.

Using the introduced notation, the number of active treatments is c = 2, the number of observations are n_0 = 6, n_1 = 4 (j = 1 represent drug A) and n_2 = 5 (j = 2 represents drug B) and s^2 = 1.381 based on n = 12 degrees of freedom.

Easy calculations show that $d_1 = \dfrac{8.90 - 8.25}{\sqrt{1.381}\sqrt{4^{-1} + 6^{-1}}} = 0.86$ and

$$d_2 = \frac{10.88 - 8.25}{\sqrt{1.381}\sqrt{5^{-1} + 6^{-1}}} = 3.69$$

and (D_1, D_2) follows a bivariate Student t-distribution with $n = 12$ degrees of freedom and correlation matrix $\mathbf{R} = \begin{pmatrix} 1 & r_{1,2} \\ r_{1,2} & 1 \end{pmatrix}$ where $r_{1,2} = \sqrt{\dfrac{4}{6+4}}\sqrt{\dfrac{5}{6+5}} = 0.426$.

As will be shown in the next section, the p-value is $p = P\left(\max\limits_{1,2}\{D_j\} > 3.69\right) = 0.003$ in case of the one-sided alternative or $p = P\left(\max\limits_{1,2}\{|D_j|\} > 3.69\right) = 0.006$ in case of the two-sided alternative.

2.2 Calculation of probabilities and upper percentage points

The testing problems under examination require probabilities and upper percentage points from the distribution of the test statistic $D = \max_{1 \le j \le c} \{D_j\}$ in case of the one-sided test situation or the test statistic $D = \max_{1 \le j \le c} \{|D_j|\}$ in case of the two-sided test situation under the null hypothesis. The probability distribution depends on the parameters *c*, n and the correlation matrix $\mathbf{R}$ characterized by the set of the *c* parameters $\{b_j\}$.

This section describes how these probabilities and upper percentage points can be calculated. The first method to compute the probabilities makes use of the fact that the joint distribution of the D_j's follows a multivariate t-distribution. The second method is based on the multivariate normal distribution. Both these methods are general in the sense that they can be applied to a broad class of correlation structures and not necessarily restricted to correlation structures that follows from the many-to-one comparisons. Taking the correlation structure into account simplifies the calculations as shown by the third method.

Percentage points

The aim is to find percentage points $d(a,c,n,\{b_j\})$ such that $P_{H_0}\left(D \le d(a,c,n,\{b_j\})\right) = 1 - a$. If one is able to compute the probabilities $P(D \le t)$ for arbitrary *t* under the null hypothesis, then the problem is mainly solved. There are several methods to compute percentage points given an algorithm that computes probabilities. Popular methods are rejection types of algorithms as proposed by Edwards and Berry (1987) and the class of root finding methods, like the secant method and the bisection method.

Bretz (1999) showed that the bisection method yields good results and is simple to implement. The SAS/IML program code can be found in Appendix 4.

Therefore the problem is reduced to the computation of the probabilities $P(D \le t)$.

Tables from literature / approximate procedures

Instead of computing the probabilities, one can make use of tables available in the literature as has been done in the past for many other problems as well. However, the percentage points $d(a,c,n,\{b_j\})$ depend on a , *c*, n and on the correlations r_{j_1,j_2}'s or equivalently the sample size ratios $n_0 / n_1, ..., n_0 / n_c$. (Note that $b_j = \left(1+\frac{n_0}{n_j}\right)^{-1/2}$ (2.8).) Thus it is not possible to tabulate $d(a,c,n,\{b_j\})$ in general. Replacing the correlations r_{j_1,j_2}'s by a common value r provides natural approximations to the critical points $d(a,c,n,\{b_j\})$, since tables are widely available for the equicorrelated situation. Several values of r are proposed based on Bonferroni type of inequalities or on a suitable average of the r_{j_1,j_2}'s like the arithmetic mean

$$\overline{r} = \frac{2}{c(c-1)} \sum_{1 \le j_1 < j_2 \le c}^{c} r_{j_1,j_2} .$$

Comprehensive tables of the percentage points for the equicorrelated situation are given in Bechhofer and Dunnett (1988).

However, with the current available numerical solutions to handle unequal sample sizes, this approach is not recommended anymore.

Computation of probabilities

Without loss of generalisability only the one-sided alternative testing situation is considered, i.e. the calculation of probabilities $P(D \le t)$ under the null hypothesis where $D = \max_{1 \le j \le c} \{D_j\}$.

The two-sided testing situation will not be described for the one-way layout. However, in case of a two-way layout, the two-sided testing situation will be described in Chapter 4.

1. Multivariate t-distribution

The probability $P(D \le t)$ can be calculated by making use of the joint distribution of the D_j's:

$$P(D \le t) = P\left(\max_{1 \le j \le c} \{D_j\} \le t\right) = P\left(all\, D_j \le t\right) = P\left(D_1 \le t \wedge ... \wedge D_c \le t\right) \qquad (2.10)$$

Under H_0, the joint distribution of the D_j's follows a central c-variate t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$ characterized by $\{b_j\}$. Thus t is the equi-percentage point of this t-distribution, i.e.

$$P(D \le t) = T_c(-\infty, \mathbf{t}; n, \mathbf{R}) \qquad (2.11)$$

So the problem is solved if one is able to compute probabilities of a multivariate t-distribution.

Until recently the direct numerical evaluation of the multivariate t-probabilities for an arbitrary correlation matrix was considered computationally infeasible. And even in case programs were available these numerical computations were too slow to be useful for practical purposes except for very small dimensions.

However, recent developments on the numerical evaluation of the multivariate t-integral have solved this problem for practical settings. See for a detailed and up to date/state of the art discussion Genz and Bretz (1999), Somerville and Bretz (2001) and Bretz, Genz and Hothorn (2001)

Computer programs are available and for example SAS/IML code can be found on the homepage of Bretz (the website with URL http://www.bioinf.uni-hannover.de/~bretz/).

2. Multivariate normal distribution

The multivariate normal distribution is more frequently mentioned in literature than the multivariate t-distribution. For the calculation of the cumulative density function of a multivariate normal distribution are more solutions available than for the cumulative density function of the multivariate t-distribution. Therefore it might be useful to express the probability $P(D \leq t)$ in terms of a multivariate normal distribution rather than a multivariate t-distribution. This can be accomplished by making use of the relationship between the multivariate t-distribution and the multivariate normal distribution as already described by Dunnett (1955). (See also Appendix 1 for the relationship between the multivariate normal and multivariate t-distribution.) Dunnett showed that the distribution function of a *c*-variate t-distribution with ν degrees of freedom and correlation matrix $\mathbf{R}$ could be transformed into a single integral over a *c*-variate normal distribution with the same matrix $\mathbf{R}$ as covariance matrix, i.e.

$$T_c(-\infty, \mathbf{t}; n, \mathbf{R}) = \int_0^\infty \Phi_c(-\infty, \mathbf{t}\sqrt{x}; \mathbf{0}, \mathbf{R}) h_n(x) dx$$

where $T_c(-\infty, \mathbf{t}; n, \mathbf{R})$ and $\Phi_c(-\infty, \mathbf{x}; \boldsymbol{\mu}, \mathbf{S})$ are the cumulative density functions of the multivariate t-distribution and multivariate normal distribution, respectively and $h_n(x)$ is the density function of a c_n^2/n distributed random variable.

And thus

$$P(D \leq t) = \int_0^\infty \Phi_c(-\infty, \mathbf{t}\sqrt{x}; \mathbf{0}, \mathbf{R}) dH_n(x) \qquad (2.12)$$

So the problem of calculating the probabilities from the distribution of the test statistic $D = \max\limits_{1 \le j \le c} \left\{ D_j \right\}$ has been reduced to the calculation of the cumulative density function of the multivariate normal distribution.

3. Univariate normal distribution

Both the above-mentioned approaches don't make use of the special structure of the correlation matrix $\mathbf{R}$. As shown is formula (2.7), the correlation matrix $\mathbf{R} = \left\{ r_{ij} \right\}$ satisfies the product structure condition, i.e. $r_{ij} = b_i b_j \ \forall i \ne j$ with $b_i = \sqrt{\dfrac{n_i}{n_0 + n_i}}$.

It can be shown that given this condition, the calculation of the probability of the cumulative density function does not involve the integration of a *c*-dimensional multivariate normal distribution but can be calculated using the univariate standard normal distribution (for further details see also Appendix 1):

$$P(D \le t) = \int_0^\infty \Phi_c(\mathbf{-\infty}, \mathbf{t}\sqrt{x}; \mathbf{0}, \mathbf{R}) h_n(x) dx = \int_0^\infty \left[\int_{-\infty}^{\infty} \prod_{j=1}^{c} \Phi\left(\frac{b_j y + t\sqrt{x}}{\sqrt{1 - b_j^2}} \right) d\Phi(y) \right] h_n(x) dx \qquad (2.13)$$

where $\Phi(y)$ is the cumulative density function of the univariate standard normal distribution and $h_n(x)$ is the density function of a c_n^2 / n distributed random variable.

The advantage to express the probability $P(D \le t)$ in terms of the univariate standard normal distribution due to the product correlation structure is that the computation times are reduced considerably. In particular for increasing dimensions of *c*.

Dunnett (1984) described an algorithm (in FORTRAN) that computes multivariate normal probability integrals with product correlation structure. The outer-integral could be evaluated using an appropriate numerical integration routine.

In SAS, the function PROBMC can compute this probability $P(D \le t)$, say *prob*, directly with the following statement

prob = PROBMC('DUNNETT1', t, ., n, c, b_1, b_2, . . . , b_c)

In addition, the PROBMC function allows computing the upper percentage points $d(a, c, n, \{b_j\})$ with only one statement:

d = `PROBMC('DUNNETT1',.,`$1 - a$`,`n`,`$c, b_1, b_2,$`...,`b_c`)`.

(See Appendix 2 for further details of this SAS/STAT function)

All three methods are exact in the sense that they only have a numerical error, which can be kept under control.

Computations of the example

The computation of the one-sided p-value in case of the example can be easily computed with the following statement

pval = `1 – PROBMC('DUNNETT1',3.69,.,12,2,SQRT(4/10),SQRT(5/11))`

which returns a value of *pval* = 0.003.

The correct upper percentage point $d(a,c,n,\{b_j\})$ at a = 0.05, i.e. $t_{2,12,\mathbf{R};0.95}$, can be computed with the statement

d = `PROBMC('DUNNETT1',.,0.95,12,2,SQRT(4/10),SQRT(5/11))`

which returns a value of *d* = 2.121.

2.3 Implementation for testing and estimation

In practice the test of an individual hypothesis whether a particular active treatment is superior to the control is often more relevant than testing the global hypothesis. This section describes how this can be achieved.

Consider the finite family of *c* sub-hypotheses

$$H_{0j} : m_j = m_0$$

against the one-sided alternatives

$$H_{1j} : m_j > m_0 .$$

The test statistic D_j of formula (2.4) is used for testing H_{0j} versus H_{1j} and H_{0j} is rejected if and only if D_j exceeds say x_a .

As already discussed in section 2.1, $H_0 = \bigcap_j H_{0j}$ and $H_1 = \bigcup_j H_{1j}$ in case of the one-sided global testing situation, where H_0 is the global null hypothesis of no effect between all of the c+1 treatments and H_1 is the one-sided alternative hypothesis that there exists an active treatment which is superior to control as formulated in (2.3).

The Union Intersection (UI) test rejects H_0 if $D = \max_j \{D_j\} > x_a$, where x_a should be chosen such that $P_{H_0}(D > x_a) = a$. It follows easily from the previous section that x_a is the upper percentage point of the c-variate t-distribution, i.e. $t_{c,n,\mathbf{R};1-a}$. (See also Appendix 1 for notation.)
Roy and Bose (1953) showed that, if the single inference given by the UI test of H_0 is of level a , then all multiple inferences, tests and confidence estimates, for the parameters on which the hypotheses H_{0j} are postulated have the family wise error rate controlled at level a .
Thus all individual hypotheses H_{0j} with corresponding $D_j > t_{c,n,\mathbf{R};1-a}$ can be rejected if the global hypothesis H_0 can be rejected because of $D > t_{c,n,\mathbf{R};1-a}$ while strongly controlling the FWE at level a .

Adjusted p-values

However, for most testing applications, it is more informative to determine a p-value for each individual hypothesis than merely noting whether a specific level a has been reached. Therefore, in line with the definition of an unadjusted p-value for a single hypothesis test, a multiplicity-adjusted p-value for an individual hypothesis is defined as the smallest overall significance level at which that hypothesis can be rejected using a particular multiple testing procedure and the observed test statistic (Wright (1992)). Sometimes these adjusted p-values are called joint p-values; see e.g. Dunnett and Tamhane (1991).
In our testing situation, the adjusted p-value $\tilde{p}_j$ belonging to the testing of the null hypothesis H_{0j} is

$$\tilde{p}_j = \min\left\{a \mid H_{0j} \text{ is rejected at FWE} = a\right\} = P_{H_0}\left(D > d_j\right) = 1 - T_c(\text{-}\infty, \mathbf{d}_j; n, \mathbf{R}) \qquad (2.14)$$

where d_j is the observed value of the test statistic D_j (j = 1, …, c).

Simultaneous confidence intervals

Notice that the testing procedure is a simultaneous test procedure in the sense of Gabriel (1969); see also Section 2.1. As a result, it has all of the desirable properties of a simultaneous test procedure and simultaneous confidence intervals can be obtained as indicated by Gabriel (1969). See Appendix 1 of Hochberg and Tamhane (1987) for further details of simultaneous test procedures and simultaneous confidence regions.

Therefore, corresponding upper one-sided 100(1-a)% simultaneous confidence intervals for $m_j - m_0$ are given by

$$(\bar{X}_j - \bar{X}_0 - t_{c,n,\mathbf{R};1-a}\, s\sqrt{n_j^{-1} + n_0^{-1}}, \infty) \quad (j = 1, ..., c). \tag{2.15}$$

It is clear that the simultaneous coverage probability of these c intervals is 1 - α.

The computation of the simultaneous confidence intervals requires the correct upper percentage point from the probability distribution of $D = \max_{1 \le j \le c} \{D_j\}$ which, as we have seen, is the 1 - a upper percentage point of the central c-variate t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$ characterized by the set of the c parameters $\{b_j\}$, denoted as $t_{c,n,\mathbf{R};1-a}$.

Computations of the example

The computations for the example are summarized in the following table.

Table 2.2 Analysis of blood count data

Contrast	Difference	Adjusted p-value $\tilde{p}_j$	95% Confidence interval
Low dose - Plac.	0.650	0.325	(-0.959, ∞)
High dose - Plac.	2.628	0.003	(1.119, ∞)

3 Dunnett’s procedure extended to the stratified two-way layout

This chapter describes the many-to-one comparisons in the situation of a stratified two-way layout. More specifically, it describes the comparisons of the mean of all active treatments with the control mean within each of the strata simultaneously while controlling the familywise error rate. As will be shown, it can be seen as an extension of Dunnett’s multiple comparison procedure, which is discussed in Chapter 2, to the case of several strata. The examples provided in Chapter 1 illustrated the situation of many-to-one comparisons in each of several strata in different practical settings. This chapter is restricted to the one-sided alternative testing situation only. The two-sided testing situation can be handled very similar to the one-sided testing situation although the formulas are slightly more complicated. The required adaptations to handle the two-sided situation are described in Chapter 4.

Section 3.1 and Section 3.2 outline the derivation of the probability distribution of the appropriate test statistic and show how percentage points and simultaneous confidence intervals can be derived as described by Cheung and Holland (1991, 1992). Power considerations are discussed in Section 3.3 and the related issue of sample size calculations is described in Section 3.4. The step-down procedure as proposed by Cheung and Holland (1994) is discussed in Section 3.5.

Example

The following example of an experiment/trial will be used throughout this chapter as an illustration. Two active dosages of a new drug, a low and a high dose, are compared against placebo. The subjects were classified to the two gender groups. Twenty subjects in this trial were randomly assigned to the control treatment. The outcome parameter of interest is a continuous variable that can be assumed to be normally distributed. A high outcome indicates improvement. The aim is to compare both active treatments with control separately for males and females.

Table 3.1 Summary statistics of example dataset

	Treatment								
	Placebo			Low dose			High dose		
Stratum	N	Mean	SD	N	Mean	SD	N	Mean	SD
Males	10	10.29	0.81	7	11.16	0.55	5	12.46	0.90
Females	10	14.64	0.73	6	15.22	1.26	5	15.90	0.55

3.1 Notation and test statistic

This section introduces the test statistic using identical notation used to describe the one-way layout situation in Section 2.1 as far as possible.

Suppose the following fixed effect two-way layout model:

$$X_{ijk} = m_{ij} + e_{ijk} \quad i = 1, \ldots, r\ ,\ j = 0, 1, \ldots, c \text{ and } k = 1, \ldots, n_{ij} \ (> 0 \text{ for all } (i,j)) \qquad (3.1)$$

Let X_{ijk} denotes the *k*-th observation on treatment *j* in stratum *i*. Again, let *j* = 0 denotes the control treatment or other designated treatment level.

Without loss of generalization the number of treatments contained in each stratum are assumed to be equal, although the formulas will also apply in case this situation does not hold, i.e. *c* varies across *i*. (See also Cheung and Holland (1994))

Assume that the sample values $\left\{X_{ijk}\right\}$ are independently normal distributed with mean m_{ij} and an unknown common variance s^2, i.e. $X_{ijk} \sim N(m_{ij}, s^2)$.

Let $\overline{X}_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk}$ denotes the sample mean (*i* = 1, …, *r* and *j* = 0, 1, …, *c*) and let $s^2 = \dfrac{\sum_{i=1}^{r}\sum_{j=0}^{c}\sum_{k=1}^{n_{ij}}(X_{ijk} - \overline{X}_{ij})^2}{n}$ be the usual pooled variance estimator of s^2 based on $n = \sum_{i=1}^{r}\sum_{j=0}^{c}(n_{ij} - 1))$ degrees of freedom, which is independent of the sample means $\overline{X}_{ij}$.

The global null hypothesis to be tested is the hypothesis of no effect between any of the *c* active treatments versus control within each of the *r* strata

$$H_0 : m_{ij} = m_{i0} \qquad (i = 1, \ldots, r\ \ j = 1, \ldots, c) \qquad (3.2)$$

against the one-sided alternative hypothesis that there exists an active treatment which is superior to control within at least one of the *r* strata

$$H_1 : \exists ij : m_{ij} > m_{i0} \qquad (i = 1, \ldots, r\ \ j = 1, \ldots, c).$$

(Assuming that a higher treatment mean m_{ij} implies an improvement. In case a lower treatment effect implies improvement and superiority should be demonstrated by showing that $m_{ij} < m_{i0}$, one should use the negative values to end up with the current settings.)

Similar to the test statistic as proposed by Dunnett (1955) in his original procedure, Cheung and Holland (1991, 1992) proposed the pivotal statistics:

$$D_{ij} = \frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c) \tag{3.3}$$

and to test the global null hypothesis H_0 the test statistic

$$D = \max_{1 \le i \le r; 1 \le j \le c} \{D_{ij}\} \tag{3.4}$$

is defined.

The procedure that rejects the global null hypothesis H_0 in favor of the one-sided alternative hypothesis H_1 if $D > d_a$, where d_a is chosen such that $P_{H_0}(D > d_a) = a$, controls the FWE.

Notice that this test statistic is a direct extension of the test statistic proposed by Dunnett to perform many-to-one comparisons for a one-way layout situation, because the test statistic in formula (3.4) reduces to the test statistic in formula (2.5) in case there is only one single stratum, i.e. $r = 1$.

Similar to the one-way layout situation, it can easily be shown that under the global null hypothesis the joint distribution of the D_{ij}'s follows a central rc-variate Student t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$, denoted as $(D_{11},\ldots,D_{rc})' \sim t_{rc}(n, \mathbf{R})$.

The correlation $r_{(i_1 j_1),(i_2 j_2)}$ between each pair of $D_{i_1 j_1}$ and $D_{i_2 j_2}$ is given as follows:

if $i_1 \ne i_2$ $\quad r_{(i_1 j_1),(i_2 j_2)} = 0$ because the X_{ijk}'s of different strata are assumed to be independent,

if $i_1 = i_2$ $\quad r_{(i_1 j_1),(i_2 j_2)} = r_{i(j_1, j_2)} = b_{ij_1} b_{ij_2}$ $(1 \le j_1 \ne j_2 \le c)$ where $b_{ij} = \sqrt{\dfrac{n_{ij}}{n_{i0} + n_{ij}}}$. (3.5)

So the correlation matrix $\mathbf{R}$ is given by:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R_1} & .. & 0 \\ .. & .. & .. \\ 0 & .. & \mathbf{R_r} \end{pmatrix} \text{ where } \mathbf{R_i} = \begin{pmatrix} 1 & r_{i(1,2)} & .. & r_{i(1,c)} \\ r_{i(2,1)} & 1 & .. & .. \\ .. & .. & 1 & r_{i(c-1,c)} \\ r_{i(c,1)} & .. & r_{i(c,c-1)} & 1 \end{pmatrix} \quad (i = 1,\ldots, r) \qquad (3.6)$$

Thus the correlation matrix $\mathbf{R}$ has a block diagonal structure that partially satisfies the product correlation structure, which means that this condition holds within each of the strata, i.e. each $\mathbf{R_i}$ satisfies the product correlation structure.

<u>Computations of the example</u>

For the example described, the test situation is as follows:

$$H_0 : m_{10} = m_{11} = m_{12} \text{ and } m_{20} = m_{21} = m_{22}$$

versus

$$H_1 : m_{11} > m_{10} \text{ or } m_{12} > m_{10} \text{ or } m_{21} > m_{20} \text{ or } m_{22} > m_{20}$$

where i = 1 and i = 2 represent the males and females respectively, and where j = 1 and j = 2 represent the low and high dose respectively.

Thus one is interested to test that any of the dosages is better than placebo.

Using the introduced notation, the number of strata is r = 2, active treatments is c = 2, s^2 = 0.671 based on n = 37 degrees of freedom.

Applying ANOVA on these data shows the following results:

Table 3.2 ANOVA output

Stratum	Contrast	Estimate	Std error	Df	T	Pr(t)*
M	Plac-Low	0.864	0.404	37	2.139	0.020
	Plac-High	2.163	0.449	37	4.820	<0.001
F	Plac-Low	0.582	0.423	37	1.375	0.089
	Plac-High	1.265	0.449	37	2.819	0.004

* one-sided unadjusted p-value

The D_{ij}'s follows a 4-variate central t-distribution with n = 37 degrees of freedom and correlation matrix $\mathbf{R}$, i.e. $\left(D_{11}, D_{12}, D_{21}, D_{22}\right)' \sim t_4(37, \mathbf{R})$.

The correlation matrix $\mathbf{R}$ is given by $\mathbf{R} = \begin{pmatrix} 1 & r_{1(1,2)} & & \\ r_{1(2,1)} & 1 & \multicolumn{2}{c}{\mathbf{0}} \\ & & 1 & r_{2(1,2)} \\ \multicolumn{2}{c}{\mathbf{0}} & r_{2(2,1)} & 1 \end{pmatrix}$ with

$$r_{1(1,2)} = r_{1(2,1)} = \sqrt{\frac{7}{17}}\sqrt{\frac{5}{15}} = 0.3705 \text{ and } r_{2(1,2)} = r_{2(2,1)} = \sqrt{\frac{6}{16}}\sqrt{\frac{5}{15}} = 0.3536.$$

3.2 Probabilities, upper percentage points and simultaneous confidence intervals

Like the situation of a one-way layout, the one-sided testing problems under examination require probabilities and upper percentage points from the probability distribution of $D = \max_{1 \le i \le r; 1 \le j \le c} \{D_{ij}\}$ which depends on the parameters r, c, n and the correlation matrix $\mathbf{R}$ characterized by the set of the rc parameters $\left\{b_{ij}\right\}$.

The upper percentage point $d(a, r, c, n, \{b_{ij}\})$ such that $P_{H_0}\left(D \le d(a, r, c, n, \{b_{ij}\})\right) = 1 - a$ is the 1 - a percentage point of the central rc variate t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$, which will be denoted as $t_{rc,n,\mathbf{R};1-a}$.

In principle the same algorithms to calculate the probabilities $P(D \le t)$ for arbitrary t, as described for the one-way layout model (see Section 2.2) can be applied.
The algorithms making use of the multivariate t-distribution and multivariate normal distribution can be applied without any changes.
Only the third algorithm that makes use of the fact that the correlation matrix of the D_{ij}'s satisfies the product correlation matrix to break the computations down to the integration of univariate standard normal distributions should be slightly adapted because the block diagonal correlation matrix $\mathbf{R}$ doesn't satisfy the product correlation structure completely, although the $\mathbf{R}_i$'s satisfy this condition. (See formula (3.5))

Bechhofer and Tamhane (1974) already showed that a multivariate normal probability integral over a rectangular region could be expressed as an iterated integral that is much easier to evaluate numerically in case the covariance matrix has a certain block covariance structure.

The third algorithm can be worked out as follows:

3. Univariate normal distribution

Make use of the block diagonal structure of the correlation matrix $\mathbf{R}$, where each of the $\mathbf{R}_i$'s satisfies the product correlation structure.

Then:

$$P\left(D \le t\right) = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j=1}^c \Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) j\,(y) dy \right] h(u) du \qquad (3.7)$$

where $u = s^2 / s^2$, $h(u)$ is the density function of a c_n^2 / n distributed variable, i.e. $h(u) = \frac{n^{n/2} e^{-un/2} u^{n/2-1}}{\Gamma(n/2) 2^{n/2}}$ and $\Phi(y)$ and $j\,(y)$ are the standard cumulative distribution function and probability density function respectively.

Proof:

$$P\left(D \le t\right) = P\left(\max_{1\le i\le r;1\le j\le c} \left\{ D_{ij} \right\} \le t \right) = \int_0^\infty P\left(\max_{1\le i\le r;1\le j\le c} \left\{ \frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right\} \le t\sqrt{u}; \sqrt{\frac{s^2}{s^2}} = \sqrt{u} \right) h(u) du =$$

$$= \int_0^\infty \prod_{i=1}^r P\left(\max_{1\le j\le c} \left\{ \frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right\} \le t\sqrt{u} \right) h(u) du = \int_0^\infty \prod_{i=1}^r P\left(\frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \le t\sqrt{u}\ ;\ \forall j \right) h(u) du =$$

$$= \int_0^\infty \prod_{i=1}^r \Phi_c\left(-\infty, \mathbf{t}\sqrt{\mathbf{u}}; \mathbf{0}, \mathbf{R}_i \right) h(u) du = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j=1}^c \Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) j\,(y) dy \right] h(u) du$$

where $\Phi_c\left(\mathbf{-}\infty, \mathbf{t}\sqrt{\mathbf{u}}; \mathbf{0}, \mathbf{R}_i\right)$ is the c-variate normal integral with expectation $\mathbf{0}$, correlation matrix $\mathbf{R}_i$ over the rectangular region with upper integration bounds $t\sqrt{u}$.

Notice that the inner integrand

$$prob_i = \int_{-\infty}^{\infty} \prod_{j=1}^{c} \Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) j\,(y)dy \tag{3.8}$$

is the probability provided by the original Dunnett procedure applying infinite degrees of freedom.

The probability $prob_i$ can directly be computed using the PROBMC function, which is available within SAS with the following statement:

$prob_i$ = PROBMC('DUNNETT1',$t\sqrt{u}$, ., ., c, b_{i1}, b_{i2}, . . ., b_{ic})
(See Appendix 2 for further details of this SAS/STAT function)

The probability $P(D \le t)$ can be calculated within SAS by using the subroutine QUAD available within PROC IML to perform the required numerical integration in one dimension.
The complete SAS program code to compute these probabilities can be found in Appendix 3.

Adjusted p-values

Analogue to the testing situation with a single stratum, the adjusted p-value $\tilde{p}_{ij}$ corresponding to the sub-hypothesis $H_{0ij} : m_{ij} = m_{i0}$ versus the alternative $H_{1ij} : m_{ij} > m_{i0}$ is defined as

$$\tilde{p}_{ij} = \min\left\{ a \;\middle|\; H_{0ij} \text{ is rejected at FWE} = a \right\} = P_{H_0}\left(D > d_{ij}\right) = 1 - T_{rc}\left(-\infty, \mathbf{d}_{ij}; n, \mathbf{R}\right) \tag{3.9}$$

where d_{ij} is the observed value of the test statistic D_{ij} (i = 1, …, r and j = 1, …, c).

(See Sections 2.1 and 3.3 for the relationship between these sub-hypotheses and the global hypothesis.)

Upper one-sided 100(1-α)% simultaneous confidence intervals

The derivation of simultaneous confidence intervals in case of the two-way situation is identical to the one-way situation as described in Section 2.3.

Therefore, upper one-sided 100(1-α)% simultaneous confidence intervals for $m_{ij} - m_{i0}$ are given by

$$\left(\bar{X}_{ij} - \bar{X}_{i0} - d_a s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}, \infty\right) \quad \text{for all } i = 1, \dots, r \text{ and } j = 1, \dots, c \tag{3.10}$$

where

$$d_a = d(a, r, c, n, \{b_{ij}\}) = t_{rc,n,\mathbf{R};1-a} \ .$$

(For further details see Section 2.3.)

Computations of the example

The computation of the p-value for the example by making use of the univariate normal distribution results in a p-value of <0.001.

The program code to compute the adjusted p-values, the upper-percentage point and the one-sided 95% simultaneous confidence intervals in the setting of the example is shown in program Ch3_12.sas of Appendix 3. Running this program provides an upper percentage point of $d_{0.05} = d(0.05, 2, 2, 37, \{b_{ij}\}) = t_{4,37,\mathbf{R};0.95} = 2.306$ and the results presented in the following table.

Table 3.3 Adjusted p-values and simultaneous confidence intervals

Stratum	Contrast	Estimate	Adjusted p-value $\tilde{p}_{ij}$	95% Confidence interval
M	Plac-Low	0.864	0.072	(-0.067, ∞)
	Plac-High	2.163	<0.001	(1.128, ∞)
F	Plac-Low	0.582	0.286	(-0.394, ∞)
	Plac-High	1.265	0.015	(0.230, ∞)

Looking at the points estimates, the low and the high dosage seems to show improvement over placebo in both genders. But only the high dosage shows a statistically significant (p<0.05) improvement in both genders. For the males, the improvement is highly significant (p<0.001) and estimated as 2.163, with at least an improvement of 1.128 with 95% confidence. For the females, the effect is less pronounced; with 95% confidence the effects is at least 0.230 and estimated as 1.265, which is statistically significant (p = 0.015). Noticing that the lower bound of the 95% confidence interval is larger than 0 can also conclude this.

3.3 Power

The power can be defined analogue to the definition of the power for a univariate test. In univariate testing applications, the power of a test is defined as

$$Power = P\left(\text{reject } H_0 \mid H_0 \text{ is false}\right).$$

To perform this calculation, the condition 'H_0 is false' should be specified precisely. For example when testing $H_0 : m_1 = m_2$, the condition 'H_0 is false' must be specified by giving a particular non-null value for $m_1 - m_2$.

To test the global null hypothesis

$$H_0 : m_{ij} = m_{i0} \;\; (i = 1,\ldots,r \;\; j = 1,\ldots,c)$$

against the one-sided alternative hypothesis that there exists an active treatment which is superior to control within at least one of the *r* strata

$$H_1 : \exists i, j : m_{ij} > m_{i0} \;\; (i = 1,\ldots,r \;\; j = 1,\ldots,c)$$

using the statistic

$$D = \max_{1 \le i \le r; 1 \le j \le c} \{D_{ij}\}$$

the power can be defined similarly to the univariate testing situation as

$$Power = P\left(D > d_a \mid H_1\right) \tag{3.11}$$

where the configuration of the m_{ij}'s should be specified under the alternative.

This power is often referred to as the global power, i.e. the power of the global hypothesis.

A closed formed expression of the power can be derived as described by Genz and Bretz (1999):

$$Power = P\left(D > d_a \mid H_1\right) P\left(\max_{1 \le i \le r; 1 \le j \le c} \{D_{ij}\} > d_a \mid H_1\right) = 1 - P\left(\max_{1 \le i \le r; 1 \le j \le c} \{D_{ij}\} \le d_a \mid H_1\right) =$$

$$= 1 - P\left(\frac{\bar{X}_{11} - \bar{X}_{10}}{s\sqrt{n_{11}^{-1} + n_{10}^{-1}}} \le d_a \wedge \quad \ldots \quad \wedge \quad \frac{\bar{X}_{rc} - \bar{X}_{r0}}{s\sqrt{n_{rc}^{-1} + n_{r0}^{-1}}} \le d_a \mid H_1 \right) = \tag{3.12}$$

$$= 1 - P\left(\frac{\dfrac{(\bar{X}_{11} - m_{11}) - (\bar{X}_{10} - m_{10})}{s\sqrt{n_{11}^{-1} + n_{10}^{-1}}} + \dfrac{m_{11} - m_{10}}{s\sqrt{n_{11}^{-1} + n_{10}^{-1}}}}{s/s} \le d_a \wedge \quad \ldots \quad \wedge \quad \frac{\dfrac{(\bar{X}_{rc} - m_{rc}) - (\bar{X}_{r0} - m_{r0})}{s\sqrt{n_{rc}^{-1} + n_{r0}^{-1}}} + \dfrac{m_{rc} - m_{r0}}{s\sqrt{n_{rc}^{-1} + n_{r0}^{-1}}}}{s/s} \le d_a \right.$$

It can be shown that the joint distribution of the D_{ij}'s under the alternative hypothesis H_1 follows a noncentral *rc*-variate t-distribution with correlation matrix $\mathbf{R}$, n degrees of freedom and noncentrality vector $\mathbf{d} = \left(d_{ij}\right)_{1 \le i \le r; 1 \le j \le c} = \left(\dfrac{m_{ij} - m_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right)_{1 \le i \le r; 1 \le j \le c}$.

So the power can be calculated if the values of $m_{ij} - m_{i0}$ are specified and if one is able to compute probabilities of a noncentral multivariate t-distribution. Like the evaluation of the central multivariate t-distribution, no algorithm was available, until recently, to numerically compute this probability directly.

Nowadays, computer programs are available to compute the probabilities of a noncentral multivariate t-distribution; see Genz and Bretz (1999) and Bretz, Genz and Hothorn (2001). For example SAS/IML code can be found on the homepage of Bretz (the website with URL http://www.bioinf.uni-hannover.de/~bretz/).

Making use of the relationship between the multivariate t-distribution and multivariate normal distribution a different expression of the power can be derived. (See for further details also Section 2.2 and Appendix 1)

Assume that $\mathbf{Z}$ is a standardized *k*-variate normal distribution random variable with correlation matrix $\mathbf{R}$ and independently U is a c_n^2 / n distributed random variable with density $h(u)$, i.e. $h(u) = \dfrac{n^{n/2} e^{-un/2} u^{n/2-1}}{\Gamma(n/2) 2^{n/2}}$.

Then the power can be expressed as:

$$Power = 1 - P\left(D \le d_a \mid H_1\right) = 1 - P\left(\frac{\mathbf{Z}+\mathbf{d}}{\sqrt{U}} \le \mathbf{d}_a\right) = 1 - \int_0^{\infty} P\left(\mathbf{Z}+\mathbf{d} \le \mathbf{d}_a\sqrt{u}\right)h(u)du =$$

$$= 1 - \int_0^{\infty} \Phi_{rc}\left(-\infty, \mathbf{d}_a\sqrt{u} - \mathbf{d}; \mathbf{0}; \mathbf{R}\right)h(u)du = 1 - \int_0^{\infty}\prod_{i=1}^{r} \Phi_c\left(-\infty, \mathbf{d}_a\sqrt{u} - \mathbf{d}_i; \mathbf{0}; \mathbf{R}_i\right)h(u)du =$$

$$= 1 - \int_0^{\infty}\prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\prod_{j=1}^{c}\Phi\left(\frac{b_{ij}y + d_a\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}}\right)j\,(y)dy\right]h(u)du \qquad (3.13)$$

where $\Phi_{rc}\left(-\infty, \mathbf{d}_a\sqrt{u} - \mathbf{d}; \mathbf{0}; \mathbf{R}\right)$ is the *rc*-variate normal integral with expectation $\mathbf{0}$, correlation matrix $\mathbf{R}$ over the rectangular region with upper integration bounds $d_a\sqrt{u} - d$, $\Phi(.)$ and $j\,(.)$ are the univariate standard cumulative distribution function and probability density function respectively and $b_{ij} = \sqrt{\dfrac{n_{ij}}{n_{i0} + n_{ij}}}$ (3.5).

So the power can be computed by using similar algorithms as those applied to compute the upper percentage points d_a.

This definition of global power, i.e. $Power = P\left(\text{reject } H_0 \mid H_0 \text{ is false}\right)$, is in line with the definition given by Hayter and Liu (1992). They define the power as the probability of rejecting the overall hypothesis $H_0 : m_0 = m_1 = \ldots = m_k$, if $\max\limits_{1 \le j \le k} | m_j - m_0 | / s \ge d$ for preassigned d. So H_0 is rejected if Dunnett's test rejects at least one of the sub-hypotheses $H_{0j} : m_0 = m_j$ (*j* = 1, …, *k*), no matter which one. That is, a rejected H_{0j} doesn't need to belong to a treatment with $| m_j - m_0 | / s \ge d$. However, the hypothesis belonging to the largest difference from the control will have the greatest chance of being rejected.

Therefore another way to define the power is to look at the single global null hypothesis H_0 that all *rc* treatment means are equal to the control means and the one-sided alternative hypothesis H_1 as a finite family of *rc* individual sub-hypotheses:

$$H_{0ij} : m_{ij} = m_{i0} ,$$

against (3.14)

$$H_{1ij} : m_{ij} > m_{i0} .$$

Notice that $H_0 = \bigcap_{ij} H_{0ij}$ and $H_1 = \bigcup_{ij} H_{1ij}$.

In the situation of multiple hypotheses testing power can be defined in many different ways. The most common used definitions include the so-called *all-pairs* power and *any-pair* power definitions introduced by Ramsey (1978) and the so-called *per-pair* power. The all-pairs power is the probability of detecting all true differences, the any-pair power is the probability of detecting at least one true difference and the per-pair power is the probability of detecting a particular difference:

All-pairs power = P(reject all H_{0ij} that are false),

Any-pair power = P(reject at least one H_{0ij} that is false), (3.15)

Per-pair power = P(reject a particular H_{0ij} that is false).

In general the all-pairs power appears to be attractive because obviously one would like to reject all false hypotheses. However, this is a stringent definition, since reasonable practical designs often have low power to obtain rejections for all false hypotheses.

In contrast the any-pair power is the probability that at least one significant result will be found in the experiment. The any-pair power is most compatible with multiple testing methods that aim to control the FWE at a , since the power function approaches the nominal FWE level a as the parameters approach the complete null configuration.

The per-pair power is most closely related to the power definition in the univariate testing situation. The difference is that the test uses the multiplicity-adjusted critical value instead of the unadjusted critical value. Notice however, that it seems unnatural to be interested in only one particular selected hypothesis while in a multiple comparison setting.

Thus which power definition one wants to apply in the experiment should be considered carefully.

(See Westfall, Tobias, et al. (1999) for further details and alternative names of these power definitions.)

Suppose that S is the subset of $\{ij\}$ such that the null hypotheses H_{0ij} are false when $ij \in S$ and all remaining null hypotheses are true. Assume that k is the dimension of S, i.e. there are k false null hypotheses H_{0ij}.

Then the all-pairs power can be written as

$$Power_{all-pairs} = P\left(D_{ij} > d_a \forall ij \in S\right) = P\left(\frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} > d_a \forall ij \in S \right) =$$

$$= P\left(\frac{\dfrac{(\bar{X}_{ij} - m_{ij}) - (\bar{X}_{i0} - m_{i0})}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} + \dfrac{m_{ij} - m_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}}}{s/s} > d_a \forall ij \in S \right) \tag{3.16}$$

which is the probability of a k-variate noncentral t-distribution with correlation matrix $\mathbf{R}_k$, n degrees of freedom and noncentrality vector $\mathbf{d} = \left(\dfrac{m_{ij} - m_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right)_{ij \in S}$, where $\mathbf{R}_k = \left\{ r_{(i_1 j_1),(i_2 j_2)} \right\}_{(i_1 j_1),(i_2 j_2) \in S}$ is the correlation matrix $\mathbf{R}$ restricted to the subset of $ij \in S$.

Note that the computation of the all-pairs power requires that the alternatives be specified precisely.

The all-pairs power can also be expressed in terms of the univariate standard normal distribution:

$$Power_{all-pairs} = \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j:ij \in S} \Phi\left(\frac{-b_{ij} y - d_a\sqrt{u} + d_{ij}}{\sqrt{1 - b_{ij}^2}} \right) j(y) dy \right] h(u) du \tag{3.17}$$

where $\Phi(.)$ and $j(.)$ are the univariate standard cumulative distribution function and probability density function and *h(u)* is the density function of a c_n^2/n distributed variable.

Proof:

Define $Y_{ij} = \frac{X_{ij} - m_{ij}}{s}\sqrt{n_{ij}}$, $Y_{i0} = \frac{X_{i0} - m_{i0}}{s}\sqrt{n_{ij}}$ and $U = \frac{s^2}{s^2}$. The Y_{ij} 's and Y_{i0} 's are i.i.d. standard normal random variables being independent of U which is c_n^2 /n distributed.

Notice that $\frac{1}{\sqrt{n_{ij}}}\frac{1}{\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} = \sqrt{1 - b_{ij}^2}$ and $\frac{1}{\sqrt{n_{i0}}}\frac{1}{\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} = b_{ij}$, where $b_{ij} = \sqrt{\frac{n_{ij}}{n_{i0} + n_{ij}}}$ (3.5).

Then:

$$Power_{all-pairs} = P\left(D_{ij} > d_a \forall ij \in S\right) = P\left(\frac{\sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij}}{\sqrt{U}} > d_a \forall ij \in S\right) =$$

$$= \int_0^\infty P\left(\sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij} > d_a\sqrt{u} \ \forall ij \in S\right) h(u)du =$$

$$= \int_0^\infty \left[\int_{-\infty}^\infty P\left(\sqrt{1-b_{ij}^2}Y_{ij} + d_{ij} > b_{ij}y + d_a\sqrt{u} \ \forall ij \in S\right) j\,(y)dy\right] h(u)du =$$

$$= \int_0^\infty \left[\int_{-\infty}^\infty P\left(Y_{ij} > \frac{b_{ij}y + d_a\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \ \forall ij \in S\right) j\,(y)dy\right] h(u)du =$$

$$= \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:ij \in S} \Phi\left(\frac{-b_{ij}y - d_a\sqrt{u} + d_{ij}}{\sqrt{1-b_{ij}^2}}\right) j\,(y)dy\right] h(u)du$$

A similar proof can be given by making use of the relationship between the multivariate t-distribution and normal distribution and the partial product correlation structure of $\mathbf{R}_k$. But this proof is not valid in the situation of only one false null hypothesis, i.e. *k* = 1.

$$Power_{all-pairs} = P\left(D_{ij} > d_a \forall ij \in S\right) = \int_0^\infty \Phi_k\left(\mathbf{d}_a\sqrt{u} - \mathbf{d}, \infty; \mathbf{0}; \mathbf{R}_k\right) h(u)du =$$

$$= \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:ij \in S} \Phi\left(\frac{-b_{ij}y - d_a\sqrt{u} + d_{ij}}{\sqrt{1-b_{ij}^2}}\right) j\,(y)dy\right] h(u)du$$

where $\Phi_k\left(\mathbf{d}_a\sqrt{u} - \mathbf{d}, \infty; \mathbf{0}; \mathbf{R}_k\right)$ is the *k*-variate normal integral with expectation $\mathbf{0}$, correlation matrix $\mathbf{R}_k$ over the rectangular region with lower integration bounds $d_a\sqrt{u} - d$.

Similarly the any-pair power can be expressed as

$$Power_{any-pair} = P\left(D_{ij} > d_a \exists ij \in S\right) = 1 - P\left(D_{ij} \le d_a \forall ij \in S\right) =$$

$$= 1 - P\left(\frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \le d_a \forall ij \in S \right) =$$

$$= 1 - P\left(\frac{\dfrac{(\bar{X}_{ij} - m_{ij}) - (\bar{X}_{i0} - m_{i0})}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} + \dfrac{m_{ij} - m_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}}}{s/s} \le d_a \forall ij \in S \right) \qquad (3.18)$$

which can also be computed as the probability of a *k*-variate noncentral t-distribution with correlation matrix $\mathbf{R}_k$, n degrees of freedom and noncentrality vector $\mathbf{d} = \left(\dfrac{m_{ij} - m_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right)_{ij \in S}$.

Alternatively, the any-pair power can also be expressed in terms of univariate distributions:

$$Power_{any-pair} = 1 - \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j|ij \in S} \Phi\left(\frac{b_{ij} y + d_a \sqrt{u} - d_{ij}}{\sqrt{1 - b_{ij}^2}} \right) j(y) dy \right] h(u) du . \qquad (3.19)$$

(The proof is similar to the proof of the all-pairs power.)

Notice that in the global power can be considered as a special case of the any-pair power by assuming that all null hypotheses H_{0ij} are false, such that S has a dimension of *rc.*

Suppose that one is interested in detecting only one particular difference corresponding with the sub-hypothesis $H_{0\tilde{i}\tilde{j}}$. Then the per-pair power can be expressed as follows

$$Power_{per-pair} = P\left(D_{\tilde{i}\tilde{j}} > d_a\right) = P\left(\frac{\dfrac{(\bar{X}_{\tilde{i}\tilde{j}} - m_{\tilde{i}\tilde{j}}) - (\bar{X}_{\tilde{i}0} - m_{\tilde{i}0})}{s\sqrt{n_{\tilde{i}\tilde{j}}^{-1} + n_{\tilde{i}0}^{-1}}} + \dfrac{m_{\tilde{i}\tilde{j}} - m_{\tilde{i}0}}{s\sqrt{n_{\tilde{i}\tilde{j}}^{-1} + n_{\tilde{i}0}^{-1}}}}{s/s} > d_a \right) \qquad (3.20)$$

which is the probability of a univariate noncentral Student t-distribution with n degrees of freedom and noncentrality parameter $d_{\tilde{ij}} = \frac{m_{\tilde{ij}} - m_{\tilde{i}0}}{s\sqrt{n_{\tilde{ij}}^{-1} + n_{\tilde{i}0}^{-1}}}$.

Notice that the per-pair power can be considered as a special case of the all-pairs power or of the any-pair power in the situation that there is only one false null hypothesis, i.e. $k = 1$.

Computations of the example

Assume that the trial described in the example was planned with 10 subjects on placebo, 7 on the low dosage and 5 on the high dosages for both males and females. (Thus one female in the low dose group resulted in a missing value at the end of the trial.) And that the variance is assumed to be $s^2 = 0.70$.

These settings determine the correlation matrix $\mathbf{R} = \mathbf{R}\{b_{ij}\}$ ($b_{11} = b_{12} = \sqrt{5/15}$ and $b_{21} = b_{22} = \sqrt{7/17}$) such that the critical value d_a can be calculated as $d_a = t_{4,38,\mathbf{R};0.95} = 2.304$ for $a = 0.05$.

Suppose that one assumes that both dosages are superior to placebo in males and females, i.e. all sub-hypotheses H_{0ij}'s are assumed to be false and S consists of all indices: $S = \{(11),(12),(21),(22)\}$.

Then the all-pairs power and any-pair power can be computed by using formulas (3.16) and (3.18) respectively, if one specifies the differences $m_{ij} - m_{i0}$ for S, because these differences determine the noncentrality vector $\mathbf{d}$. Program Ch3_3.sas of Appendix 3 shows how these probabilities of a noncentral multivariate t-distribution can be computed applying the SAS/IML code of Bretz available on his homepage (website with URL http://www.bioinf.uni-hannover.de/~bretz/).

The following table shows the results for several configurations of the vector of differences $\Delta\boldsymbol{\mu} = (m_{11} - m_{10}, \ldots, m_{22} - m_{20})$:

Table 3.4 All-pairs and any-pair power for the complete set of indices

$\Delta\boldsymbol{\mu}$	All-pairs power	Any-pair power
(0.5, 1 , 0.5, 1)	0.014	0.739
(1 , 1 , 1 , 1)	0.113	0.895
(1 , 1.5, 1 , 1.5)	0.263	0.982
(1.5, 1.5, 1.5, 1.5)	0.604	0.998
(1 , 2 , 1 , 2)	0.312	>0.999
(2 , 2 , 2 , 2)	0.945	>0.999

(The error in the computations is less than 0.0001)

Thus with an improvement of 1 point of all four dosages compared to placebo, the probability to reject all four sub-hypotheses is somewhat more than 11% and the probability to reject at least one of the sub-hypotheses is almost 90%.

Another scenario would be that one assumes that only the high dosage in both males and females is superior to placebo. So S is now a real subset consisting of $S = \{(12),(22)\}$.

The critical value remains the same, i.e. $d_a = t_{4,38,\mathbf{R};0.95} = 2.304$ but the correlation matrix $\mathbf{R}$ should be restricted to S. It turns out that $\mathbf{R}_2 = \mathbf{I}_2$ the identity matrix of dimension 2, because there is only active treatment to compare with the control treatment for each of the two strata.

The noncentrality vector $\mathbf{d}$ is specified by the vector $\Delta\boldsymbol{\mu} = \left(m_{12} - m_{10}, m_{22} - m_{20}\right)$.

The all-pairs power and any-pair power for several configurations of the vector $\Delta\boldsymbol{\mu}$ are shown in the following table.

Table 3.5 All-pairs and any-pair power for a subset of indices

$\Delta\boldsymbol{\mu}$	All-pairs power	Any-pair power
(0.5, 0.5)	0.018	0.229
(1 , 1)	0.223	0.698
(1.5, 1.5)	0.693	0.967
(2 , 2)	0.956	>0.999

(The error in the computations is less than 0.0001)

Comparing the results of both sets of indices S gives an idea of the impact of the choice of S. For example comparing the power of $\Delta\boldsymbol{\mu} = (1 , 1 , 1 , 1)$ for the complete set of indices with

$\Delta\boldsymbol{\mu} = (1\ ,\ 1)$ for the subset of indices shows that the all-pairs power is increased from 0.113 to 0.223 and that the any-pair power is decreased from 0.895 to 0.698.

Intuitively, this is obvious. Assume that the true effects are equal, then it is harder to reject all false hypotheses if the number of false hypotheses is increasing, i.e. a smaller all-pairs power, but it is easier to reject at least one false hypothesis if the number of false hypotheses is increasing, i.e. a higher any-pair power.

3.4 Sample size

An important aspect in the design of studies and the planning of experiments is to know how large the sample size must be in order to detect certain relevant differences with a preassigned probability.

Horn and Vollandt (1998, 2000) showed how sample size formulas could be derived in the single-stratum situation for any of the three types of power, the all-pairs, the any-pair and the per-pair power. This section expands these formulas for the stratified situation under consideration.

The determination of sample sizes demands that a minimum difference Δ (Δ > 0) between the active treatment and the placebo group in their population means should be preassigned which is worth detecting. For example, in a clinical trial, Δ may represent the minimum clinical relevant difference.

Similar to the discussion of how to calculate the power in the situation of multiple hypotheses testing as stated in the previous section, one should decide whether all hypotheses H_{0ij} with differences between m_{ij} and m_{i0} of at least Δ should be rejected with a given probability $1-b$, or whether at least one hypothesis H_{0ij} with a difference between m_{ij} and m_{i0} of at least Δ should be rejected with a given probability $1-b$ or whether a particular single hypothesis H_{0ij} should be rejected with a given probability $1-b$ if the difference between m_{ij} and m_{i0} is at least Δ.

Denote the difference between m_{ij} and m_{i0} as $m_{i,j-0}$, i.e. $m_{i,j-0} = m_{ij} - m_{i0}$ for shortening the notation, then the following power definitions can be added:

The *all-pairs* Δ power is defined as the probability of rejecting all hypotheses with $m_{i,j-0} \geq \Delta$, the *any-pair* Δ power is defined as the probability of rejecting at least one hypothesis with $m_{i,j-0} \geq \Delta$ and the *per-pair* Δ power is the probability of rejecting a particular single hypothesis with $m_{i,j-0} \geq \Delta$ i.e.

All-pairs Δ power = P(reject all H_{0ij} with $m_{i,j-0} \geq \Delta$),

Any-pair Δ power = P(reject at least one H_{0ij} with $m_{i,j-0} \geq \Delta$), (3.21)

Per-pair Δ power = P(reject a particular H_{0ij} with $m_{i,j-0} \geq \Delta$).

Using the expressions of different kind of powers as derived in the previous section and filling in $b_{ij} = \sqrt{\dfrac{n_{ij}}{n_{i0} + n_{ij}}}$ (3.5) and $d_{ij} = \dfrac{m_{i,j-0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}}$ (Section 3.3) result in the following expressions for the all-pairs Δ power, any-pair Δ power and the per-pair Δ power:

$$Power_{all-pairs\ \Delta} = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0} \geq \Delta} \Phi\left(-\sqrt{\frac{n_{ij}}{n_{i0}}} y - \sqrt{1 + \frac{n_{ij}}{n_{i0}}} d_a \sqrt{u} + \frac{\sqrt{n_{ij}}\, m_{i,j-0}}{s} \right) j(y)dy \right] h(u)du \quad (3.22)$$

$$Power_{any-pair\ \Delta} = 1 - \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0} \geq \Delta} \Phi\left(\sqrt{\frac{n_{ij}}{n_{i0}}} y + \sqrt{1 + \frac{n_{ij}}{n_{i0}}} d_a \sqrt{u} - \frac{\sqrt{n_{ij}}\, m_{i,j-0}}{s} \right) j(y)dy \right] h(u)du \quad (3.23)$$

$$Power_{per-pair\ \Delta} = \int_0^\infty \int_{-\infty}^\infty \Phi\left(-\sqrt{\frac{n_{\tilde{i}\tilde{j}}}{n_{\tilde{i}0}}} y - \sqrt{1 + \frac{n_{\tilde{i}\tilde{j}}}{n_{\tilde{i}0}}} d_a \sqrt{u} + \frac{\sqrt{n_{\tilde{i}\tilde{j}}}\, m_{\tilde{i},\tilde{j}-0}}{s} \right) j(y)dy\, h(u)du \quad (3.24)$$

where again $h(u)$ is the density function of a c_n^2/n distributed variable and $\Phi(y)$ and $j(y)$ are the standard cumulative distribution function and probability density function respectively.

The per-pair Δ power can be considered as a special case of the all-pairs Δ power or of the any-pair Δ power, which is also mentioned in the previous section. Therefore, it is sufficient to deal with the all-pairs Δ power and the any-pair Δ power. The per-pair Δ power won’t be discussed in the sequel of this section.

Notice that the effect size worthwhile to be detected is supposed to be identical across all strata, i.e. Δ does not depend on *i*. However, it might be that one wants to detect different effect sizes within each of the strata. In that situation Δ should be replaced by Δ_i. The formula's used throughout this section can be easily extended but are somewhat more complex and won't be considered here further.

In addition, sample sizes n_{ij} are only determined in the case of $n_{i1} = n_{i2} = ... = n_{ic} = n_i$, for every *i*, i.e. all active treatment arms have the same sample size within each stratum. However n_{i0} may be different from n_i. Notice that the power expressions are monotone increasing in n_i.

The power expressions depend on the ratios $l_i = \frac{n_i}{n_{i0}}$. In the sequel l_i is supposed to be constant, i.e. $l_i = l$. Two common values for l are $l = 1$ and $l = \frac{1}{\sqrt{c}}$. Where the latter value represents the well-known square root allocation that was shown to be nearly optimal by Dunnett (1955). See also Spurrier and Nizam (1990) for optimal sample size allocation in comparing several treatments with a control in a one-way layout.

Furthermore, these expressions depend on the real and unknown number of differences $m_{i,j-0} \geq \Delta$. Denote the unknown number of differences $m_{i,j-0} \geq \Delta$ by k_i. Notice that this number k_i is not restricted to be the same for all strata. In most cases k_i is completely unknown, i.e. it is only known that $0 \leq k_i \leq c$ with at least one $k_i \geq 1$. However, it is easy and useful to consider the more general case where a priori knowledge states that $g_i \leq k_i \leq h_i$ for some lower bound integers g_i and upper bound integers h_i with $0 \leq g_i \leq h_i \leq c$ and at least one $g_i \geq 1$. Thus, the most common situation where no a priori knowledge is available is regarded as the special case $g = 1$ (at least one treatment has a true difference of at least Δ) and $h_i = c$ (all treatments have a true difference of at least Δ), where $g = \sum_{i=1}^{r} g_i$.

The task is to determine the minimal integers n_i, which guarantee that the Δ power is not smaller than a preassigned probability $1 - b$ for any values $m_{i,j-0}$ provided that $g_i \leq k_i \leq h_i$ (*i* = 1, …, *r* *j* = 1, …, *c*). For that reason n_i is determined for the least favorable configurations (LFC) of $m_{i,j-0}$ with $g_i \leq k_i \leq h_i$ that provide minima of the Δ power.

For the all-pairs Δ power it can be shown that

$$Power_{all-pairs\ \Delta} = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0} \geq \Delta} \Phi\left(-\sqrt{l}y - \sqrt{1+l}d_a\sqrt{u} + \frac{\sqrt{n_i}\,m_{i,j-0}}{s} \right) j(y)dy \right] h(u)du \geq$$

$$\geq \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0} \geq \Delta} \Phi\left(-\sqrt{l}y - \sqrt{1+l}d_a\sqrt{u} + \frac{\sqrt{n_i}\,\Delta}{s} \right) j(y)dy \right] h(u)du \geq$$

$$\geq \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \Phi^{h_i}\left(-\sqrt{l}y - \sqrt{1+l}d_a\sqrt{u} + \frac{\sqrt{n_i}\,\Delta}{s} \right) j(y)dy \right] h(u)du =$$

$$= P\left(D_{i1} > d_a, ..., D_{ih_i} > d_a\ \forall i \,\middle|\, m_{i,1-0} = ... = m_{i,h_i-0} = \Delta \forall i \right) \qquad (3.25)$$

which means that $m_{i,1-0} = ... = m_{i,h_i-0} = \Delta$, $m_{i,(h_i+1)-0} = ... = m_{i,c-0} = 0$ (i = 1, …, r) is a LFC for the all-pairs Δ power if $g_i \leq k_i \leq h_i$ (i = 1, …, r). Notice that there are many LFC 's.

Thus assuming that the unknown number of differences between m_{ij} and m_{i0} of at least Δ is equal to the upper bound h_i guarantees a minimum of the all-pairs Δ power. Without any a priori knowledge, i.e. $h_i = c$ (i = 1, …, r), the LFC is given by $m_{i,1-0} = ... = m_{i,c-0} = \Delta$ (i = 1, …, r).

This LFC for the all-pairs Δ power is also intuitively clear. A true difference of $m_{i,j-0} = \Delta$ is harder to detect than a true difference of $m_{i,j-0} > \Delta$. And an increasing number of true differences $m_{i,j-0} \geq \Delta$ decreases the probability to detect them all; the maximum number of true differences within each stratum equals h_i.

So, the smallest integers n_i have to determined such that the probability $P\left(D_{11} > d_a, ..., D_{rh_r} > d_a \,\middle|\, m_{1,1-0} = ... = m_{r,h_r-0} = \Delta \right)$ is at least $1-b$, for given a, Δ, l and h_i (i = 1, …, r).

This power probability can be calculated by noticing that $\left(D_{11}, ..., D_{1h_1}, ..., D_{r1}, ..., D_{rh_r} \right)$ follows a noncentral h-variate t-distribution with correlation matrix $\mathbf{R}_h$, $n = \sum_{i=1}^r \left(n_i l_i^{-1} + cn_i - c - 1 \right)$ degrees of freedom and noncentrality vector $\mathbf{d} = \left(\frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} \right)$, where $h = \sum_{i=1}^r h_i$. The correlation matrix $\mathbf{R}_h$ has a block-diagonal structure, with correlation coefficients $r_{i(j_1,j_2)} = \frac{n_i}{n_i + n_{i0}} = \frac{l}{1+l}$ ($1 \le j_1 \ne j_2 \le h_i$) and zero's elsewhere.

The degrees of freedom n depends on the sample sizes n_i. Therefore no explicit expression of n_i can be obtained if the variance s^2 is unknown and the determination of n_i should be performed iteratively. Notice that the critical value $d_a = t_{hn,\mathbf{R}_h;1-a}$ also depends on the sample sizes through the degrees of freedom. However, in the situation that the variance s^2 is known explicit formulas of n_i can be obtained in some situations as will be shown later on.

Similarly for the any-pair Δ power it can be shown that

$$Power_{any-pair\ \Delta} = 1 - \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0}\ge\Delta} \Phi\left(\sqrt{l}y + \sqrt{1+l}d_a\sqrt{u} - \frac{\sqrt{n_i}m_{i,j-0}}{s}\right) j(y)dy \right] h(u)du \ge$$

$$\ge 1 - \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{i,j-0}\ge\Delta} \Phi\left(\sqrt{l}y + \sqrt{1+l}d_a\sqrt{u} - \frac{\sqrt{n_i}\Delta}{s}\right) j(y)dy \right] h(u)du \ge$$

$$\ge 1 - \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \Phi^{g_i}\left(\sqrt{l}y + \sqrt{1+l}d_a\sqrt{u} - \frac{\sqrt{n_i}\Delta}{s}\right) j(y)dy \right] h(u)du =$$

$$= 1 - P\left(D_{i1} < d_a, ..., D_{ig_i} < d_a\ \forall i : g_i \ge 1 \middle| m_{i,1-0} = ... = m_{i,g_i-0} = \Delta\ \forall i : g_i \ge 1\right) \quad (3.26)$$

This implies that $m_{i,1-0} = ... = m_{i,g_i-0} = \Delta,\ m_{i,(g_i+1)-0} = ... = m_{i,c-0} = 0$ (*i* = 1, …, *r*) is a LFC for the any-pair Δ power if $g_i \le k_i \le h_i$ (*i* = 1, …, *r*). So assuming that the unknown number of differences between m_{ij} and m_{i0} of at least Δ is equal to the lower bound g_i guarantees a minimum of the any-pairs Δ power. In case no a priori knowledge is available, i.e. $g = \sum_{i=1}^r g_i = 1$, a LFC is given by any configuration were exactly one of the differences $m_{i,j-0}$ equals Δ and all others are smaller than Δ, for example $m_{1,1-0} = \Delta,\ m_{1,1-0} = ... = m_{1,c-0} = 0$ and $m_{i,1-0} = \Delta,\ m_{i,1-0} = ... = m_{i,c-0} = 0$ (*i* = 2, …, *r*).

The LFC for the any-pair Δ power is also intuitively clear by noticing that the probability to detect at least one true difference is smaller with a lower number of true differences of $m_{i,j-0} \ge \Delta$; the minimum number of true differences within each stratum equals g_i.

In general, it is not uncommon to assume that the variance s^2 is known, i.e. assuming infinite degrees of freedom, in order to perform sample size calculations.
In that case the test statistics have the simple form:

$$D_{ij} = \frac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \quad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

Under that assumption the minimal all-pairs Δ power can be written as

$$\begin{aligned}\text{Minimal}Power_{all-pairs\ \Delta} &= P\left(D_{i1} > d_a, \ldots, D_{ih_i} > d_a \ \forall i \,\middle|\, m_{i,1-0} = \ldots = m_{i,h_i-0} = \Delta \forall i\right) = \\ &= P\left(Z_{i1} > d_a - \frac{\Delta}{s\sqrt{n_i^{-1} + n_{i0}^{-1}}}, \ldots, Z_{ih_i} > d_a - \frac{\Delta}{s\sqrt{n_i^{-1} + n_{i0}^{-1}}} \ \forall i\right) = \\ &= P\left(Z_{i1} \le \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_a, \ldots, Z_{ih_i} \le \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_a \ \forall i\right) \qquad (3.27)\end{aligned}$$

where the Z_{ij}'s are jointly distributed as a standardized h-variate normal random variable with correlation matrix $\mathbf{R}_h$.
Therefore, the smallest integers n_i have to be determined for which

$$P\left(Z_{i1} \le \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_a, \ldots, Z_{ih_i} \le \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_a \ \forall i\right) \ge 1 - b\ , \qquad (3.28)$$

where $1 - b$ is the preassigned minimal required all-pairs Δ power.

However, there are many h-vectors $\mathbf{b} = \left(b_1, \ldots, b_1, \ \ldots\ , b_r, \ldots, b_r\right)'$, with $b_i = \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_a$ (i = 1, …, r) such that $\Phi_h(\text{-}\infty, \mathbf{b}, \mathbf{0}, \mathbf{R}_h) \ge 1 - b$. Thus no unique solution for the sample sizes n_i can be derived, unless the ratio of the sample sizes between the strata is defined upfront.
For example in the situation that the sample sizes are equal across all strata, i.e. $n_i = n$, and treating the problem symmetrically with regard to all hypotheses with $m_{i,j-0} \ge \Delta$, the solution is the smallest integer n for which

$$n \geq (1+l)\left(d_a + x_{h,\mathbf{0},\mathbf{R}_h;1-b}\right)^2 s^2 / \Delta^2 \tag{3.29}$$

where $x_{h,\mathbf{0},\mathbf{R}_h;1-b}$ is the $1-b$ percentage point of an *h*-variate standardized normal distribution with correlation matrix $\mathbf{R}_h$. Notice that $d_a = x_{rc,\mathbf{0},\mathbf{R}_{rc};1-a}$ (see also Appendix 1 for notation). Without any a priori knowledge about the unknown number of differences $m_{i,j-0} \geq \Delta$, *h* has to be replaced by *rc*.

Similarly for the any-pair Δ power, the smallest n_i have to be determined for which

$$1-P\left(Z_{i1} < d_a - \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}},\ldots,Z_{ig_i} < d_a - \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} \ \forall i\right) \geq 1-b\ , \tag{3.30}$$

where $1-b$ is the preassigned minimal required any-pair Δ power.

Under the same assumptions it follows that the solutions are the smallest n for which

$$n \geq (1+l)\left(d_a - x_{g,\mathbf{0},\mathbf{R}_g;b}\right)^2 s^2 / \Delta^2 \tag{3.31}$$

Without any a priori knowledge about the unknown number of differences $m_{i,j-0} \geq \Delta$, g has to be replaced by 1, which in that case leads to $x_{1,0,\mathbf{R}_1;b} = u_b = -u_{1-b}$.

The sample size requirements in this section are formulated in terms of power considerations, which is consistent with hypothesis testing. Another way to determine sample sizes is based on confidence intervals. Though this will not be discussed here any further. For more details see for example Pan and Kupper (1999) who showed sample size calculations for several multiple comparison procedure, including the one-way Dunnett's procedure, treating the confidence width as random. They illustrated that ignoring the underlying stochastic nature of the confidence width could lead to serious underestimation of the required sample sizes.

Computations of the example

Some sample size calculations are illustrated in the context of the example.

Suppose for a moment that a clinical relevant difference is represented by an improvement of at least 1.5, i.e. Δ = 1.5 and that the square-root rule is used to allocate the sample sizes between the active and control treatment, i.e. $l = \frac{1}{\sqrt{c}} = 0.707$. Furthermore, assume that there is no a-priori knowledge concerning the real but unknown number of differences $m_{i,j-0} \geq \Delta$, which implies that g = 1 and $h = rc$ = 4.

Then $m_{1,1-0} = m_{1,2-0} = m_{2,1-0} = m_{2,2-0} = 1.5$ is a LCF for the all-pairs Δ power (see (3.25)) and $m_{1,1-0} = 1.5$, $m_{1,2-0} = m_{2,1-0} = m_{2,2-0} = 0$ is a LFC for the any-pair Δ power (see (3.26)).

In addition, assume that the variance is known, say s^2 = 0.70.

Then the minimal all-pairs Δ power can be expressed as the probability of a standardized multivariate normal random vector with a correlation matrix $\mathbf{R}$, which has a block-diagonal structure with correlation coefficients $r = \frac{l}{1+l} = \frac{1}{\sqrt{c}+1} =$ 0.414. (see also (3.27))

Given the variance the critical value is given by $d_a = x_{4,\mathbf{0},\mathbf{R};0.95} =$ 2.215 (a = 0.05).

Program Ch3_4.sas of Appendix 3 shows how the probability of a standardized multivariate normal distribution can be computed applying the SAS/IML code of Bretz available on his homepage (website with URL http://www.bioinf.uni-hannover.de/~bretz/).

Assume that the required all-pairs Δ power is at least 80%, i.e. b = 0.20. Then according to formula (3.28) the sample sizes n_1 and n_2 have to be determined as the minimal integers for which $\Phi_4(-\infty,(b_1,b_1,b_2,b_2)',\mathbf{0},\mathbf{R}) \geq 0.80$ where $b_i = \frac{1.5\sqrt{n_i}}{1.093} - 2.215$ (i = 1, 2).

There is no unique solution, for example n_1 = 9 and n_2 = 7 results in a probability of 0.817, and n_1 = n_2 = 8 results in a probability of 0.834; so both solutions fulfill the requirement.

But restricting the sample sizes to be equal for males and females, the sample sizes is given by formula (3.29). Filling in $x_{4,\mathbf{0},\mathbf{R};0.80} = 1.563$ results in $n_1 = n_2$ $\geq 1.7071(2.215+1.563)^2\, 0.7/1.5^2 = 7.581$.

Thus 8 subject in each of the two dosages groups and 11 subject in the control group for both males and females are required to have a probability of at least 80% that all dosages that are superior to placebo by at least 1.5, will be detected at an a level of 5%. Assuming that no a-priori knowledge is available about how many treatments are superior to placebo by at least 1.5.

Similarly the any-pair Δ power can be determined. For example, under the same assumptions and equal sample sizes for both males and males, the sample sizes can be calculated by formula (3.29) and filling in $x_{1,\mathbf{0},\mathbf{R};0.80} = -u_{0.80} =$-0.842:

$$n_1 = n_2 \geq 1.7071(2.215 + 0.842)^2\, 0.7/1.5^2 = 9.118\,.$$

3.5 Step-down procedure

The extended Dunnett procedure discussed in the previous sections is a so-called single-step procedure. In general performing them in a stepwise manner can increase the power of single-step procedures. However, these stepwise procedures do have their drawbacks. For example, the application of stepwise procedures is mainly restricted to hypothesis testing problems, because it is only known for a few situations how to invert them to obtain simultaneous confidence.

Stepwise procedures can be divided into step-down and step-up types of procedures. A step-down procedure starts by testing the overall intersection hypothesis and then steps down through the hierarchy of implied hypotheses. If any hypothesis is not rejected, then all implied hypotheses are retained without any further testing. So a hypothesis is tested if and only if all of its implying hypotheses are rejected. The step-up procedure starts the other way around: it begins by testing all minimal hypotheses and then steps up through the hierarchy of hypotheses. If any hypothesis is rejected, then all implied hypotheses are rejected without any further testing. So a hypothesis is tested if and only if all of its implying hypotheses are retained.

The closure method proposed by Marcus, Peritz and Gabriel (1976) provides a general theorem for constructing step-down procedures. Any multiple comparison procedure based on this closure principle is called a closed testing procedure. An analogous theory to construct step-up procedures does not exist.

This section discusses the step-down procedure described by Cheung and Holland (1994). But first the closure principle is explained. At the end of the section some words are said about a step-up procedure.

Closed testing procedure

The closed testing procedure based on the closure method by Marcus et.al. (1976) works as follows:

Let $\left\{H_j, 1 \leq j \leq k\right\}$ be a finite family of hypotheses.

Form the 'closure' of this family by taking all nonempty intersection hypotheses $H_P = \bigcap_{j \in P} H_j$ for $P \subseteq \left\{1, 2, ..., k\right\}$.

Suppose that an appropriate a -level test of each hypothesis H_P is available. This test might be any test that is valid for the given intersection. Each method results in a different closed testing procedure.

Then, any hypothesis H_P is rejected if and only if H_P and every intersection hypothesis that includes H_P is rejected by its associated a -level test (i.e. H_P is rejected $\Leftrightarrow H_Q$ is rejected $\forall Q \supseteq P$).

This closed testing procedure strongly controls the FWE at a .

(Proof: see for example pages 54-55 of Hochberg and Tamhane (1987))

In general the number of tests in a closed testing procedure increases rapidly with increasing *c*. Therefore it makes sense to consider a shortcut version of the closed testing situation that can be applied in a particular setting.

Suppose there are *k* hypotheses H_1 to H_k which have the free combination property as defined by Holm (1979); i.e. the partition of the *k* hypotheses into any subset of *m* hypotheses $\left\{H_{j_1}, ..., H_{j_m}\right\}$ which are true and a subset of all remaining *k-m* hypotheses which are simultaneously false is a plausible event. (Or in other words, each of the 2^k outcomes of the *k* hypothesis problem is possible.) Note that this condition is satisfied for the many-to-one comparisons in a stratified two-way layout.

Consider a closed testing procedure that uses a UI statistic (see also section 2.1) for testing all intersection hypotheses $H_P = \bigcap_{j \in P} H_j$. Then this closed testing procedure can be applied in a shortcut manner because the UI tests have the property that whenever any intersection hypothesis H_P is rejected at least one of the H_j 's implied by H_P is rejected. Thus it is sufficient to make a rejection decision on H_j only, instead of testing all the intersections H_P containing that H_j . However, the H_j 's must be ordered to ensure that a hypothesis is automatically retained if any intersection hypothesis implying that hypothesis is retained.

In the special situation that the UI test T is of the form $T = \max_j \{T_j\}$, in particular this is true if the rejection regions of the individual hypotheses H_j's are of the form $T_j > x$, this requirement can be ensured by testing the H_j's in the order of the corresponding test statistics T_j's, starting with the hypothesis with the largest T_j.

Thus the hypothesis with the largest T_j is tested first. Notice that rejecting this hypothesis implies rejecting any intersection hypothesis containing this hypothesis, including the overall hypothesis as well. Next the hypothesis with the second largest T_j is tested. This procedure is continued until some T_j is found to be not significant. At that point all the hypotheses whose test statistic values are less than or equal to the current T_j are automatically retained. (See also Grechanovsky and Hochberg (1999))

Holm's (1979) well-known sequentially rejective Bonferroni procedure is a shortcut version of the closed testing procedure based on the Bonferroni inequality that can be applied for multiple testing problems with arbitrary correlation structures. Even if the free combination property doesn't hold, this method strongly controls the FWE, but then it can be modified to give more powerful tests. (Shaffer, 1986)

The shortcut version of the closed testing procedure in the situation of many-to-one comparisons in a stratified one-way layout was already proposed by Naik (1975) and also by Marcus et al (1976).

In most testing applications, it is more informative to determine p-values for each hypothesis than simply recording whether a specific level a has been reached. Therefore Dunnett and Tamhane (1991) showed a p-value version of the step-down procedure for comparing treatments with a control in unbalanced one-way layouts. Their method computed adjusted or so called 'joint' p-values associated with the observed treatment versus control mean differences. (See also Section 2.3) They showed that this procedure is more powerful than the sequentially rejective Bonferroni procedure of Holm (1979) and the single-step procedure of Dunnett (1955).

Cheung and Holland (1994) presented an extension of this step-down procedure to the stratified situation. This procedure is now discussed in more detail using the notation introduced in Section 3.1.

Consider the finite family of *rc* individual sub-hypotheses:

$$H_{0ij} : m_{ij} = m_{i0}\,,$$

against the upper one-sided alternatives

$$H_{1ij} : m_{ij} > m_{i0}\,.$$

Then the shortcut version of the closed testing procedure can be applied as follows:

- Order all observed test statistics d_{ij}'s from smallest to largest, say $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(rc)}$. Let $H_{0(1)}$, $H_{0(2)}$, …, $H_{0(rc)}$ be the corresponding null-hypotheses and let $E_{(m)}$ be the subset of indices *ij*'s corresponding to the *m* smallest d_{ij}'s (*m* = 1, …, *rc*). Thus $E_{(rc)}$ is the set of all indices and $E_{(1)}$ refers to the indices corresponding to $d_{(1)}$.

  Denote with $\mathbf{R}_{(m)}$ the sub-matrix of the correlation matrix $\mathbf{R}$ restricted to $E_{(m)}$ (*m* = 1, …, *rc*).
- Start with testing $H_{0(rc)}$ and reject $H_{0(rc)}$ if $d_{(rc)} > t_{rc\,n,\mathbf{R}_{(rc)};1-a}$; otherwise retain all sub-hypotheses without further tests.
- The general step *m* is, reject $H_{0(m)}$ if $H_{0(rc)}$, …, $H_{0(m+1)}$ are rejected and $d_{(m)} > t_{m,n,\mathbf{R}_{(m)};1-a}$. If $H_{0(m)}$ is not rejected, then also retain $H_{0(m-1)}$, …,$H_{0(1)}$ without any further testing (*m* = 1, …, *rc*).

(Notice that the notation is slightly different than Cheung and Holland (1994), because they ordered the d_{ij}'s within each stratum).

It can easily be shown that the critical constants $t_{mn,\mathbf{R}_{(m)};1-a}$ are monotonically increasing in *m*. The single-step procedure of Cheung and Holland (1992) uses the largest critical constant $t_{rc,n,\mathbf{R}_{(rc)};1-a}$ for testing all the hypotheses (see Section 3.2), regardless of the order, and hence the single-step procedure is less powerful than its step-down counterpart.

The p-value version, which provides adjusted p-values, can be described as follows.
Compute

$$\tilde{p}_{(m)} = P\left(\text{at least one } D_{ij} > d_{(m)},\ \{ij\} \in E_{(m)}\right) = 1 - P\left(D_{ij} \le d_{(m)},\ \{ij\} \in E_{(m)}\right) \qquad (3.32)$$

$$(m = 1, \ldots, rc)$$

Then define the adjusted p-value for $H_{0(m)}$ as

$$p_{(m)} = \max\left\{\tilde{p}_{(m)}, \tilde{p}_{(m+1)}, \ldots, \tilde{p}_{(rc)}\right\} \ (m = 1, \ldots, rc). \qquad (3.33)$$

Once these p-values are determined, hypothesis testing can be conducted at any fixed specified level a, if desired, by comparing any $p_{(m)}$ with a and rejecting $H_{0(m)}$ if $p_{(m)} \le a$ (m = 1, …, rc). In other situations it may be more useful to simply report the adjusted p-values and perhaps use them as inverse measures of the strength of evidence in favor of $H_{1(m)}$.
Notice that these adjusted p-values are monotonically ordered. Thus if $p_{(m)} > a$ and hence $H_{0(m)}$ is accepted, then monotonicity ensures acceptance also of $H_{0(m-1)}, \ldots, H_{0(1)}$.
Therefore the classical version based on critical constants for a specified a level and the p-value version are in accordance with each other.

The implementation of the step-down procedure requires the computation of either the critical constants $t_{m,n,\mathbf{R}_{(m)};1-a}$ or the adjusted p-values $p_{(m)}$. These can be computed using the same algorithms to compute the critical constants or p-values for the single-step procedure as described in Section 3.2. However, the step-down procedure is more computer intensive then the single step procedure, because the critical constants $t_{m,n,\mathbf{R}_{(m)};1-a}$ and the p-values $p_{(m)}$ have to be computed at each step. Basically the critical constants have to be computed only for the first steps, until one retain a hypothesis in which case one retain all remaining hypotheses without any further testing.

Appendix 3 contains the program code to illustrate the step-down algorithm for the example shown at the end of this Chapter.

It was though that stepwise procedures didn't have corresponding confidence sets in contrast to the single-step procedures. Stefansson, Kim and Hsu (1988) and Hayter and Hsu (1994) showed that common stepwise procedures, like the single-step procedure, do have corresponding confidence sets. In particular the step-down procedure described above does have corresponding confidence sets. Here ‘correspond’ is taken to mean that the decision that a treatment is superior to the control based on the stepwise procedure occurs only when the generated confidence interval for that treatment difference is contained within $\left(0,\infty\right)$. This confidence bounds version is not presented here. (See Hsu (1996) Chapter 3 for details in the situation of many-to-one comparisons for a one-way layout, which can be easily extended to the stratified two-way situation.)

The other type of stepwise procedures is the step-up procedure. The step-up procedure starts by testing the hypothesis corresponding to the treatment that appears to be least significant from the control group. If the hypothesis is retained the procedure proceeds towards the hypothesis with the most significant difference until the first time a hypothesis is rejected. The procedure stops and also all other remaining hypotheses are rejected without any further testing. Dunnett and Tamhane (1992, 1995) described a step-up multiple test procedure, which cover the situation of many-to-one comparisons in the single stratum setting for the equal correlated and unequal correlated situation. They showed that the proposed step-up procedure is more powerful than the single-step procedure except when only one hypothesis is false, in which case it is slightly less powerful. Similarly, it can be shown that the step-up procedure is slightly less powerful than the step-down procedure when a few hypotheses are false, but it is more powerful when most or all of the hypotheses are false.

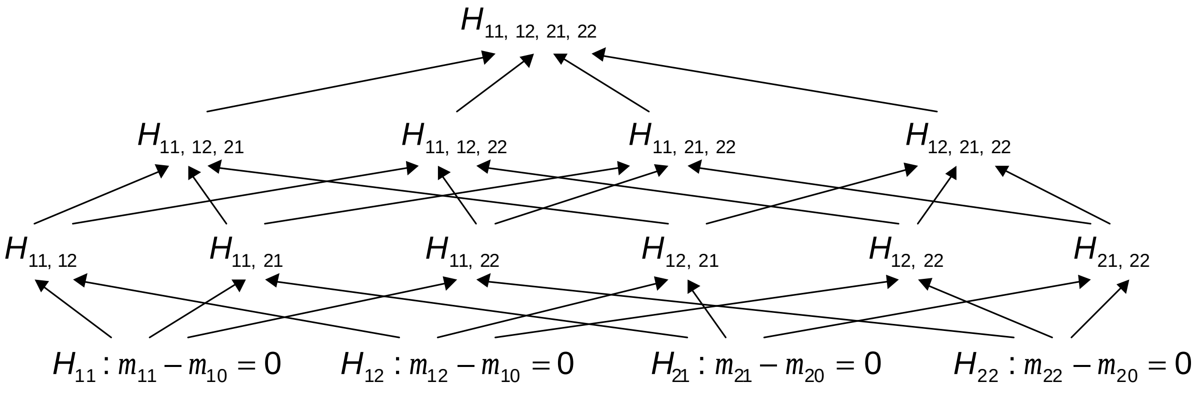
However, this procedure does not control the FWE at the pre-specified a level; see for example Liu (1997a). In addition, Liu (1997a) stated that the computation of the critical constants is very time consuming even under the equal correlation assumption and with moderate number of active treatment groups.

Finner and Roters (1998) showed the closeness of the critical constants for the step-up and step-down procedures based on asymptotic results.

The determination of sample sizes for the single-step procedure was discussed in the previous section, section 3.4. It was an adaptation of the sample size formula derived by Horn and Vollandt (1998). Similar to the determination of sample sizes for the single-step procedure it is possible to determine sample sizes for the step-down and step-up procedures as shown by Dunnett, Horn and Vollandt (2001). Although it will not be shown here, these formulas can also be extended to derive sample sizes for the may-to-one comparisons in a stratified two-way layout.

Computations of the example

The closed testing procedure in the settings of the example is as follows:



where $H_{ij,\ ...,\ lm}$ denotes the null hypothesis $H_{ij,\ ...,\ lm} : m_{ij} - m_{i0} = ... = m_{lm} - m_{l0} = 0$.

To reject, for example, the null hypothesis H_{11}, all hypotheses that include the indices *(11)* have to be rejected. Those hypotheses are all indicated in the figure above.

Notice that:

The global null hypothesis $H_{11,\ 12,\ 21,\ 22}$ is rejected at the significance level of 5% if and only if the test statistic $D = \max\{D_{11}, D_{12}, D_{21}, D_{22}\} > t_{4,37,\mathbf{R};0.95}$.

The null hypothesis $H_{11,\ 12,\ 21}$ is rejected if and only if the test statistic $\tilde{D} = \max\{D_{11}, D_{12}, D_{21}\} > t_{3,37,\tilde{\mathbf{R}};0.95}$, where $\tilde{\mathbf{R}} = \begin{pmatrix} 1 & r_{1(1,2)} & 0 \\ r_{1(1,2)} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Etc…

The calculation of the shortcut version of the closed testing procedure goes as follows:

Ordering the observed test statistics $(d_{11}, d_{12}, d_{21}, d_{22}) = (2.139,\ 4.820,\ 1.375,\ 2.819)$ from smallest to largest results in $d_{(1)} \leq d_{(2)} \leq d_{(3)} \leq d_{(4)} = d_{21} \leq d_{11} \leq d_{22} \leq d_{12}$. Thus the corresponding null-hypotheses $H_{0(1)}$, $H_{0(2)}$, $H_{0(3)}$ and $H_{0(4)}$ are H_{21}, H_{11}, H_{22} and H_{12} respectively.

- Start with testing the null hypothesis H_{12} and reject H_{12} if $d_{12} > t_{4,37,\mathbf{R};0.95}$, because $d_{12} = 4.820 > 2.306 = t_{4,37,\mathbf{R};0.95}$ the null hypothesis $H_{12} : m_{12} = m_{10}$ is rejected.
- Next, tests the null hypothesis H_{22}. Also $H_{22} : m_{22} = m_{20}$ is rejected because $d_{22} = 2.819 > 2.187 = t_{3,37,\mathbf{R}_{(3)};0.95}$.

- The null hypothesis $H_{11}: m_{11} = m_{10}$ is now tested and rejected because $d_{22} = 2.139 > 2.019 = t_{2,37,\mathbf{R}_{(2)};0.95}$.
- The last hypothesis to be tested is H_{21}. The null hypothesis $H_{21}: m_{21} = m_{20}$ is rejected if and only if $d_{21} > t_{1,37,\mathbf{R}_{(1)};0.95} = t_{37;0.95}$. Notice that $d_{21} = 1.375 < 1.688 = t_{37;0.95}$ and therefore H_{21} is retained.

The p-value version, provides the following results:
First compute the $\tilde{p}_{(m)}$'s: $(\tilde{p}_{(1)}, \tilde{p}_{(2)}, \tilde{p}_{(3)}, \tilde{p}_{(4)}) = (0.089,\ 0.039,\ 0.011,\ {<}0.001)$;
for example $\tilde{p}_{(3)} = P\left(\text{at least one } D_{ij} > d_{(3)} = d_{22} = 2.819,\ \{\text{ij}\} \in \{(11),(21),(22)\}\right) = 0.011.$
Then the adjusted p-values are

$p_{11} = p_{(2)} = \max\{\tilde{p}_{(2)}, \tilde{p}_{(3)}, \tilde{p}_{(4)}\} = \tilde{p}_{(2)} = 0.039$ for the null hypothesis H_{11},

$p_{12} = p_{(4)} = \tilde{p}_{(4)} < 0.001$ for the null hypothesis H_{12},

$p_{21} = p_{(1)} = \max\{\tilde{p}_{(1)}, \tilde{p}_{(2)}, \tilde{p}_{(3)}, \tilde{p}_{(4)}\} = \tilde{p}_{(1)} = 0.089$ for the null hypothesis H_{21},

$p_{22} = p_{(3)} = \max\{\tilde{p}_{(3)}, \tilde{p}_{(4)}\} = \tilde{p}_{(3)} = 0.011$ for the null hypothesis H_{22}.

(Notice that in this example the $\tilde{p}_{(m)}$'s are already ordered by accident.)

The single-step adjusted p-values (see also Section 3.2) and the step-down adjusted p-values are presented in the following table.

Table 3.6 Single-step and step-down adjusted p-values

Stratum	Contrast	Estimate	Single-step adjusted p-value	Step-down adjusted p-value
M	Plac-Low	0.864	0.072	0.039
	Plac-High	2.163	<0.001	<0.001
F	Plac-Low	0.582	0.286	0.089
	Plac-High	1.265	0.015	0.011

This table shows that the step-down adjusted p-values are smaller than the single-step adjusted p-values. Using the step-down adjusted p-values, the low dose for the males is statistically significant superior to placebo at the 5% level.

The SAS program code can be found in Appendix 3.

4 Two-sided testing situation

There is long and ongoing debate in the literature concerning the use of one-sided or two-sided tests in experiments. Some claim that when the research question expects a change in one direction only, the hypothesis test should reflect this by using a one-sided test. Others insist on the use of a two-sided test in case the treatment effect might be in the opposite direction than the expected direction. In particular in the conduct of clinical trials this topic has been heavily discussed. When to use one-sided tests or two-sided tests will not be discussed in this chapter any further. For a discussion of the pros and cons of both approaches see for example Peace (1991), Dubey (1991), Fisher (1991), Overall (1991), Dunnett and Gent (1996) and Senn (1997).

This chapter illustrates how the many-to-one comparisons in the situation of a stratified two-way layout for the one-sided alternative hypothesis, as described in Chapter 3, works out in the situation of a two-sided alternative hypothesis.

Section 4.1 introduces the problem of making an incorrect directional decision by rejecting a null hypothesis in case of a two-sided testing situation.

Section 4.2 illustrates step-by-step all the adaptations needed in the procedures applicable to the testing situation with a one-sided alternative hypothesis in order to make them suitable for the testing situation with a two-sided alternative hypothesis.

Section 4.3 shows how the two-sided testing problems can be approached as one-sided testing problems.

4.1 Directional decisions and Type III errors

The two-sided testing situation is slightly more complex than the one-sided testing situation as will be illustrated in this section.

Consider the family of *rc* individual sub-hypotheses

$$H_{0ij} : m_{ij} = m_{i0}$$

against the two-sided alternatives (4.1)

$$H_{1ij} : m_{ij} \neq m_{i0} .$$

Rejecting the null hypothesis $H_{0ij} : m_{ij} = m_{i0}$ in favor of the two-sided alternative hypothesis $H_{1ij} : m_{ij} \neq m_{i0}$ allows the conclusion that $\left| m_{ij} - m_{i0} \right| \neq 0$. Additionally, in practical applications it seems to be meaningful to conclude that $m_{ij} - m_{i0} > 0$ if $D_{2ij} > d_{2a}$ and to conclude that $m_{ij} - m_{i0} < 0$ if $D_{2ij} < -d_{2a}$. Of course, such a directional decision may be wrong, e.g. it is possible that $D_{2ij} > d_{2a}$ despite $m_{ij} - m_{i0} < 0$. Such an incorrect directional decision is called Type III error. The Type III FWE is the probability that the sign of any tested effect is misclassified by a multiple comparisons procedure.

Suppose that the multiple comparison procedure allows the following three decisions:

$$\begin{aligned} & m_{ij} - m_{i0} > 0 \text{ if } D_{2ij} > d_{2a}, \\ & m_{ij} - m_{i0} < 0 \text{ if } D_{2ij} < -d_{2a} \text{ and} \\ & \text{no directional decision if } \left| D_{2ij} \right| < d_{2a}. \end{aligned} \tag{4.2}$$

Then it can be shown that this single-step multiple testing procedure controls both the Type I and Type III FWE, i.e. $P\left(\textit{any one Type}\ \mathrm{I}\ \textit{or Type}\ \mathrm{III}\ \textit{error}\right) \leq a$, if the test procedure controls the Type I FWE. (For a proof and further details see Chapter 2.2 of Hochberg and Tamhane (1987))

However, with stepwise procedures it is possible that $P\left(\textit{any one Type}\ \mathrm{I}\ \textit{or Type}\ \mathrm{III}\ \textit{error}\right) > a$, as shown in Shaffer (1980) and Liu (1997b). Dunnett, Horn et. al. (2001) stated '*It is an unsolved problem in the two-sided SD and SU testing of treatments vs. a control whether the combined type I and III FWE is controlled to be $\leq a$ (as it is in the single-step test). …; Bauer (1991) speculated that for most closed test procedures applied to practical problems the combined types I and III errors do not go out of control to a noticeable extent.*'.

For more details about this topic see Finner (1999), who provides an overview of the current state of knowledge.

4.2 Adaptations for the two-sided testing situation

The topics of the many-to-one comparisons in a stratified two-way layout for a one-sided alternative testing situation as described in Chapter 3 are re-discussed in this section to make the procedures suitable for the two-sided alternative testing situation.

It is assumed that the same standard conditions remain true: the sample values $\{X_{ijk}\}$ are independently normal distributed with mean m_{ij} and unknown but common variance s^2, i.e. $X_{ijk} \sim N(m_{ij}, s^2)$, and S^2 is the usual pooled variance estimator of s^2 based on n degrees of freedom and there are *c* active treatments within each of the *r* strata.

Notation and test statistic

In the two-sided testing situation, the global null hypothesis of no effect between any of the *c* active treatments versus control within each of the *r* strata

$$H_0 : m_{ij} = m_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c) \tag{4.3}$$

is tested against the two-sided alternative hypothesis that at least one of the active treatments is different from control within any of the *r* strata

$$H_1 : \exists ij : m_{ij} \neq m_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c).$$

In line with Dunnett's statistic for the situation of a single stratum, Cheung and Holland (1991, 1992) proposed the test statistic

$$D_2 = \max_{1\le i\le r;1\le j\le c}\left\{D_{2ij}\right\} \quad \text{with} \quad D_{2ij} = \frac{\left|\bar{X}_{ij} - \bar{X}_{i0}\right|}{S\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \;\; (i = 1,\ldots,r \;\;\; j = 1,\ldots,c) \tag{4.4}$$

to test the global null hypothesis H_0 against the two-sided alternative hypothesis H_1. (To distinguish the two-sided test situation from the one-sided test situation, a subscript 2 is added.)

The global null hypothesis H_0 is rejected in favor of the two-sided alternative hypothesis H_1 if $D_2 > d_{2a}$ where d_{2a} is chosen such that $P\left(D_2 > d_{2a}\right) = a$.

Percentage points

Percentage points $d_2(a,r,c,n,\{b_{ij}\})$ such that $P_{H_0}\left(D_2 \le d_2(a,r,c,n,\{b_{ij}\})\right) = 1-a$ can be calculated by applying the same algorithms to compute the upper percentage points for the one-sided testing situation (see Sections 2.2 and 3.2). Notice that $P_{H_0}\left(D_2 \le d_2\right) = P_{H_0}\left(D_{211} \le d_2,...,D_{2rc} \le d_2\right)$.

This percentage point is called the two-sided 1 - a equi-percentage point of the central *rc* variate t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$ (see formula (3.6)), which is denoted as $|t|_{rc,n,\mathbf{R};1-a}$, i.e. $T_{rc}\left(-|t|_{rc,n,\mathbf{R};1-a}, |t|_{rc,n,\mathbf{R};1-a}; n, \mathbf{R}\right) = 1-a$. (More about this notation in Appendix 1.)

Formula (4.5) below expresses the probability $P(D_2 \le t)$, with t > 0, in terms of univariate normal distributions by making use of the block diagonal structure of the correlation matrix $\mathbf{R}$, of which each $\mathbf{R}_i$ satisfies the product correlation structure (see (3.6)):

$$P\left(D_2 \le t\right) = \int_0^\infty \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j=1}^{c} \left[\Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) - \Phi\left(\frac{b_{ij} y - t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) \right] j\,(y) dy \right] h(u) du \qquad (4.5)$$

where $h(u)$ is the density function of a c_n^2/n distributed variable and $\Phi(y)$ and $j\,(y)$ are the standard cumulative distribution function and probability density function respectively.

Proof:

$$P\left(D_2 \le t\right) = P\left(\max_{1\le i\le r; 1\le j\le c} \{D_{2ij}\} \le t \right) = \int_0^\infty P\left(\max_{1\le i\le r; 1\le j\le c} \left\{ \frac{\left|\bar{X}_{ij} - \bar{X}_{i0}\right|}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right\} \le t\sqrt{u}; \sqrt{\frac{s^2}{s^2}} = \sqrt{u} \right) h(u) du =$$

$$= \int_0^\infty \prod_{i=1}^{r} P\left(\max_{1\le j\le c} \left\{ \frac{\left|\bar{X}_{ij} - \bar{X}_{i0}\right|}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right\} \le t\sqrt{u} \right) h(u) du = \int_0^\infty \prod_{i=1}^{r} P\left(\frac{\left|\bar{X}_{ij} - \bar{X}_{i0}\right|}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \le t\sqrt{u}\ \forall j \right) h(u) du =$$

$$= \int_0^\infty \prod_{i=1}^{r} \Phi_c\left(\mathbf{-t\sqrt{u}}, \mathbf{t\sqrt{u}}; \mathbf{0}, \mathbf{R}_i \right) h(u) du =$$

$$= \int_0^\infty \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j=1}^{c} \left[\Phi\left(\frac{b_{ij} y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) - \Phi\left(\frac{b_{ij} y - t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) \right] j\,(y) dy \right] h(u) du$$

where $\Phi_c\left(\mathbf{-t\sqrt{u},t\sqrt{u};0,R}_i\right)$ is the c-variate normal integral with expectation $\mathbf{0}$, correlation matrix $\mathbf{R}_i$ over the rectangular region with lower and upper integration bounds $-t\sqrt{u}$ and $t\sqrt{u}$ respectively.

The computation of this expression within the SAS system can be simplified, like in the one-sided test situation. Thereto, notice that the inner integrand

$$prob_{2i} = \int_{-\infty}^{\infty} \prod_{j=1}^{c} \left[\Phi\left(\frac{b_{ij}y + t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) - \Phi\left(\frac{b_{ij}y - t\sqrt{u}}{\sqrt{1-b_{ij}^2}} \right) \right] j\,(y)dy \tag{4.6}$$

is the probability provided by the original Dunnett procedure for two-sided inference applying infinite degrees of freedom.

The probability $prob_{2i}$ can again directly be computed with the function PORBMC using the statement:

$prob_{2i}$ = PROBMC('DUNNETT2', $t\sqrt{u}$, ., ., c, b_{i1}, b_{i2}, . . ., b_{ic})

(See Appendix 2 for further details of this SAS/STAT function)

Notice that this probability can also be computed by means of the multivariate t-distribution, by making use of the relationship between the multivariate t-distribution and multivariate normal distribution. (See also Appendix 1) Although the D_{ij}'s themselves are not simultaneously multivariate t-distributed like in the one-sided testing situation.

$$\int_{0}^{\infty} \prod_{i=1}^{r} \Phi_c\left(\mathbf{-t\sqrt{u},t\sqrt{u};0,R}_i\right) h(u)du = \prod_{i=1}^{r} T_c\left(\mathbf{-t\sqrt{u},t\sqrt{u};}n,\mathbf{R}_i\right) = T_{rc}\left(\mathbf{-t\sqrt{u},t\sqrt{u};}n,\mathbf{R}\right) \tag{4.7}$$

where $T_c\left(\mathbf{-t\sqrt{u},t\sqrt{u};}n,\mathbf{R}_i\right)$ is the central c-variate t-integral with n degrees of freedom and correlation matrix $\mathbf{R}_i$ over the rectangular region with lower and upper integration bounds of $-t\sqrt{u}$ and $t\sqrt{u}$ respectively.

Adjusted p-values

Analogue to the one-sided testing situation (formula (3.9)), the adjusted p-value $\tilde{p}_{2ij}$ corresponding to the sub-hypothesis H_{0ij} is defined as

$$\tilde{p}_{2ij} = \min\left\{a \;\middle|\; H_{0ij} \text{ is rejected at FWE} = a\right\} = P_{H_0}\left(D_2 > d_{2ij}\right) = \tag{4.8}$$

$$= 1 - T_{rc}\left(\mathbf{-d}_{2ij}, \mathbf{d}_{2ij}; n, \mathbf{R}\right)$$

where d_{2ij} (> 0) is the observed value of the two-sided test statistic D_{2ij} (i = 1, …, r and j = 1, …, c).

Two-sided 100(1-α)% simultaneous confidence intervals

Two-sided 100(1-α)% simultaneous confidence intervals can be calculated for the two-sided test situation analogue to the one-sided test situation (formula (3.10)). However, now the percentage point $d_{2a} = d_2(a, r, c, n, \{b_{ij}\})$ should be used instead. Then two-sided 100(1-α)% simultaneous confidence intervals for $m_{ij} - m_{i0}$ are given by

$$\bar{X}_{ij} - \bar{X}_{i0} \pm d_{2a} s \sqrt{n_{ij}^{-1} + n_{i0}^{-1}} \qquad (i = 1, \ldots, r \quad j = 1, \ldots, c) \tag{4.9}$$

Power

Whether Type III errors should generally be controlled only together with Type I errors, or better together with Type II errors as both Type II and Type III errors occur with false hypotheses is still under discussion. (See e.g. Hayter and Tamhane (1991) and Horn and Vollandt (2000).) Here it is decided that the power with a two-sided test should include the requirement of a correct directional decisions, i.e., power includes the probability that $D_{2ij} > d_{2a}$ if $m_{ij} - m_{i0} > 0$ and $D_{2ij} < -d_{2a}$ if $m_{ij} - m_{i0} < 0$.

Let S^+ be the subset of $\{ij\}$ such that the null hypotheses H_{0ij} are not true and such that $m_{ij} - m_{i0} > 0$ and let S^- be the subset of $\{ij\}$ of false null hypotheses H_{0ij} with $m_{ij} - m_{i0} < 0$.

Then the all-pairs power for two-sided comparisons with correct directional decisions can be expressed in terms of the univariate normal distribution:

$$Power_{all-pairs} = \tag{4.10}$$

$$= \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \left(\prod_{j:ij\in S^-} \Phi\left(\frac{b_{ij}y - d_{2a}\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \right) x \prod_{j:ij\in S^+} \Phi\left(\frac{-b_{ij}y - d_{2a}\sqrt{u} + d_{ij}}{\sqrt{1-b_{ij}^2}} \right) \right) j(y)dy \right] h(u)du \, .$$

Proof:

The proof is very similar to the proof for the one-sided situation (see Section 3.3).

Define $Y_{ij} = \frac{X_{ij} - m_{ij}}{s}\sqrt{n_{ij}}$, $Y_{i0} = \frac{X_{i0} - m_{i0}}{s}\sqrt{n_{ij}}$ and $U = \frac{S^2}{s^2}$. Then, the Y_{ij}'s and Y_{i0}'s are i.i.d. standard normal random variables being independent of U which is c_n^2 / n distributed.

Then:

$$Power_{all-pairs} = P\left(D_{2ij} < -d_{2a} \forall ij \in S^- \cap D_{2ij} > d_{2a} \forall ij \in S^+\right) =$$

$$= P\left(\frac{\sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij}}{\sqrt{U}} < -d_{2a} \forall ij \in S^- \cap \frac{\sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij}}{\sqrt{U}} > d_{2a} \forall ij \in S^+ \right) =$$

$$= \int_0^{\infty} P\left(\sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij} < -d_{2a}\sqrt{u} \;\; ij \in S^- \cap \sqrt{1-b_{ij}^2}Y_{ij} - b_{ij}Y_{i0} + d_{ij} > d_{2a}\sqrt{u} \;\; ij \in S^+\right) h(u)du =$$

$$= \int_0^{\infty} \left[\int_{-\infty}^{\infty} P\left(\sqrt{1-b_{ij}^2}Y_{ij} + d_{ij} < b_{ij}y - d_{2a}\sqrt{u} \;\; ij \in S^-\right) P\left(\sqrt{1-b_{ij}^2}Y_{ij} + d_{ij} > b_{ij}y + d_{2a}\sqrt{u} \;\; ij \in S^+\right) j(y)dy \right]$$

$$h(u)du =$$

$$= \int_0^{\infty} \left[\int_{-\infty}^{\infty} P\left(Y_{ij} < \frac{b_{ij}y - d_{2a}\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \;\; \forall ij \in S^- \right) P\left(Y_{ij} > \frac{b_{ij}y + d_{2a}\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \;\; \forall ij \in S^+ \right) j(y)dy \right] h(u)du =$$

$$= \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \left(\prod_{j:ij\in S^-} \Phi\left(\frac{b_{ij}y - d_{2a}\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \right) x \prod_{j:ij\in S^+} \Phi\left(\frac{-b_{ij}y - d_{2a}\sqrt{u} + d_{ij}}{\sqrt{1-b_{ij}^2}} \right) \right) j(y)dy \right] h(u)du \, .$$

Analogue the any-pair power can be expressed as

$$Power_{any-pair} = P\left(\bigcup_{ij \in S^-} D_{2ij} \leq -d_{2a} \cup \bigcup_{ij \in S^+} D_{2ij} \geq d_{2a} \right) = \tag{4.11}$$

$$= 1 - P\left(D_{2ij} \geq -d_{2a} \ \ \forall ij \in S^- \cap D_{2ij} \leq d_{2a} \ \ \forall j \in S^+ \right) =$$

$$= 1 - \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \left(\prod_{j:ij \in S^-} \Phi\left(\frac{-b_{ij} y + d_{2a}\sqrt{u} + d_{ij}}{\sqrt{1-b_{ij}^2}} \right) x \prod_{j:ij \in S^+} \Phi\left(\frac{b_{ij} y + d_{2a}\sqrt{u} - d_{ij}}{\sqrt{1-b_{ij}^2}} \right) \right) j(y)dy \right] h(u)du \, .$$

Suppose that one is interested in detecting only one particular difference corresponding with the sub-hypothesis $H_{0\tilde{i}\tilde{j}}$.

The per-pair power corresponding to the sub-hypothesis $H_{0\tilde{i}\tilde{j}}$ with $m_{\tilde{i}\tilde{j}} - m_{\tilde{i}0} > 0$ is described in Section 3.3, formula (3.20). In case the sub-hypothesis $H_{0\tilde{i}\tilde{j}}$ is assumed to have $m_{\tilde{i}\tilde{j}} - m_{\tilde{i}0} < 0$ the per-pair power can be expressed as:

$$Power_{per-pair} = P\left(D_{2\tilde{i}\tilde{j}} < -d_{2a} \right) = P\left(\frac{\dfrac{(\bar{X}_{\tilde{i}\tilde{j}} - m_{\tilde{i}\tilde{j}}) - (\bar{X}_{\tilde{i}0} - m_{\tilde{i}0})}{s\sqrt{n_{\tilde{i}\tilde{j}}^{-1} + n_{\tilde{i}0}^{-1}}} + d_{\tilde{i}\tilde{j}}}{s/s} < -d_{2a} \right) \tag{4.12}$$

which is the probability of a univariate noncentral Student t-distribution with n degrees of freedom and noncentrality parameter $d_{\tilde{i}\tilde{j}} = \dfrac{m_{\tilde{i}\tilde{j}} - m_{\tilde{i}0}}{s\sqrt{n_{\tilde{i}\tilde{j}}^{-1} + n_{\tilde{i}0}^{-1}}}$.

<u>Sample size</u>

To perform the sample size determinations, the all-pairs Δ power and any-pair Δ power have to be defined. As with the definition of the all-pairs power and any-pair power discussed above, it includes the requirement of a correct directional decision:

All-pairs Δ power = P(reject all false H_{0ij} with $\left| m_{i,j-0} \right| \geq \Delta$ with
correct directional decision), (4.13)

Any-pair Δ power = P(reject at least one false H_{0ij} with $\left| m_{i,j-0} \right| \geq \Delta$ with
correct directional decision),

where $m_{i,j-0}$ denotes again $m_{i,j-0} = m_{ij} - m_{i0}$.

Then the following expressions can be derived:

$$\begin{aligned} Power_{all-pairs\ \Delta} = \int_0^\infty \prod_{i=1}^r \Bigg[\int_{-\infty}^\infty \prod_{j:m_{i,j-0}\le -\Delta} \Phi\left(\sqrt{\frac{n_{ij}}{n_{i0}}}y - \sqrt{1+\frac{n_{ij}}{n_{i0}}}d_{2a}\sqrt{u} - \frac{\sqrt{n_{ij}}\,m_{i,j-0}}{s} \right) x \\ x \prod_{j:m_{i,j-0}\ge \Delta} \Phi\left(-\sqrt{\frac{n_{ij}}{n_{i0}}}y - \sqrt{1+\frac{n_{ij}}{n_{i0}}}d_{2a}\sqrt{u} + \frac{\sqrt{n_{ij}}\,m_{i,j-0}}{s} \right) j\,(y)dy \Bigg] h(u)du \end{aligned} \tag{4.14}$$

$$\begin{aligned} Power_{any-pair\ \Delta} = 1 - \int_0^\infty \prod_{i=1}^r \Bigg[\int_{-\infty}^\infty \prod_{j:m_{i,j-0}\le -\Delta} \Phi\left(-\sqrt{\frac{n_{ij}}{n_{i0}}}y + \sqrt{1+\frac{n_{ij}}{n_{i0}}}d_{2a}\sqrt{u} + \frac{\sqrt{n_{ij}}\,m_{i,j-0}}{s} \right) x \\ x \prod_{j:m_{i,j-0}\ge \Delta} \Phi\left(\sqrt{\frac{n_{ij}}{n_{i0}}}y + \sqrt{1+\frac{n_{ij}}{n_{i0}}}d_{2a}\sqrt{u} - \frac{\sqrt{n_{ij}}\,m_{i,j-0}}{s} \right) j\,(y)dy \Bigg] h(u)du \end{aligned} \tag{4.15}$$

(Proof: similar to the proof of the all-pairs power given above; see also formula (3.22) and (3.23))

Like in the one-sided test situation, the sample sizes n_{ij} are only determined in case of $n_{i1} = n_{i2} = ... = n_{ic} = n_i$. Again, let $l = \frac{n_i}{n_{i0}}$ and denote the unknown number of differences $\left|m_{i,j-0}\right| \ge \Delta$ by k_i, where $g_i \le k_i \le h_i$ for some lower bound integers g_i and upper bound integers h_i with $0 \le g_i \le h_i \le c$ and at least one $g_i \ge 1$. In addition, denote the unknown number of differences $m_{i,j-0} \le -\Delta$ by m_i and the unknown number of differences $m_{i,j-0} \ge \Delta$ by $k_i - m_i$.

Then the configuration $m_{i,1-0} = ... = m_{i,[h_i/2]-0} = -\Delta,\ m_{i,([h_i/2]+1)-0} = ... = m_{i,h_i-0} = \Delta$ (i = 1, …, r) is a LFC of the all-pairs Δ power where $[h_i/2]$ denotes the smallest integer not smaller than $h_i/2$.

Proof:

$$Power_{all-pairs\ \Delta} = \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \prod_{j:m_{i,j-0} \leq -\Delta} \Phi\left(\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} - \frac{\sqrt{n_i}m_{i,j-0}}{s} \right) x \right.$$

$$\left. x \prod_{j:m_{i,j-0} \geq \Delta} \Phi\left(-\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}m_{i,j-0}}{s} \right) j\,(y)dy \right] h(u)du$$

$$\geq \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \Phi^{m_i}\left(\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) x \right.$$

$$\left. x\ \Phi^{k_i - m_i}\left(-\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) j\,(y)dy \right] h(u)du \quad (*)$$

$$\geq \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \Phi^{[k_i/2]}\left(\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) x \right.$$

$$\left. x\ \Phi^{k_i - [k_i/2]}\left(-\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) j\,(y)dy \right] h(u)du$$

$$\geq \int_0^{\infty} \prod_{i=1}^{r} \left[\int_{-\infty}^{\infty} \Phi^{[h_i/2]}\left(\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) x \right.$$

$$\left. x\Phi^{h_i - [h_i/2]}\left(-\sqrt{l}y - \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}\Delta}{s} \right) j\,(y)dy \right] h(u)du$$

$$= P\Big(D_{2i1} < -d_{2a}, \ldots, D_{2i[h_i/2]} < -d_{2a}, D_{2i[h_i/2]+1} > d_{2a}, \ldots, D_{2ih_i} > d_{2a}\ \forall i\ \Big|$$

$$m_{i,1-0} = \ldots = m_{i,[h_i/2]-0} = -\Delta,\ m_{i,([h_i/2]+1)-0} = \ldots = m_{i,h_i-0} = \Delta\ \forall i \Big)$$

(4.16)

$(*)$ It can be shown that the expression attains its minimum at $m_i = [k_i/2]$ and its maximum at $m_i = 0$ or $m_i = k_i$; see Appendix 2 of Horn and Vollandt (1998).

This power probability can also be calculated as the equi-percentage point of a noncentral t-distribution. To see this, rewrite

$$P\Big(D_{2i1} < -d_{2a}, \ldots, D_{2i[h_i/2]} < -d_{2a}, D_{2i[h_i/2]+1} > d_{2a}, \ldots, D_{2ih_i} > d_{2a}\ \forall i\ \Big|$$

$$m_{i,1-0} = \ldots = m_{i,[h_i/2]-0} = -\Delta,\ m_{i,([h_i/2]+1)-0} = \ldots = m_{i,h_i-0} = \Delta\ \forall i \Big)$$ as

$$P\Big(D_{2i1} < -d_{2a}, \ldots, D_{2i[h_i/2]} < -d_{2a}, -D_{2i[h_i/2]+1} < -d_{2a}, \ldots, -D_{2ih_i} < -d_{2a}\ \forall i\ \Big|$$

$$m_{i,1-0} = \ldots = m_{i,[h_i/2]-0} = -\Delta,\ m_{i,([h_i/2]+1)-0} = \ldots = m_{i,h_i-0} = \Delta\ \forall i\Big).$$

Then notice that $\left(D_{211},\ldots,D_{21[h_1/2]},-D_{1[h_1/2]+1},\ldots,-D_{21h_1},\ldots,D_{2r1},\ldots,D_{2r[h_r/2]},-D_{2r[h_r/2]+1},\ldots,-D_{2rh_r}\right)$ follows an *h*-variate noncentral t-distribution with correlation matrix $\mathbf{H}_h$, $n = \sum_{i=1}^{r}\left(n_i l^{-1} + cn_i - c - 1\right)$ degrees of freedom and noncentrality vector $\mathbf{d} = \left(\dfrac{-\Delta\sqrt{n_i}}{s\sqrt{1+l}}\right)$, where $h = \sum_{i=1}^{r} h_i$. The correlation matrix $\mathbf{H}_h$ has a block-diagonal structure, with correlation matrices $\mathbf{H}_{h_i}$ (*i* = 1, …, *r*) on the diagonal and zero's elsewhere.

The correlation matrices $\mathbf{H}_{h_i}$ have coefficients depending on the different signs of the components:

$$r_{i(j_1,j_2)} = \begin{cases} l/(1+l) & j_1 \neq j_2 \leq [h_i/2] \text{ or } j_1 \neq j_2 \geq [h_i/2]+1 \\ 1 & j_1 = j_2 \\ -l/(1+l) & j_1 \leq [h_i/2], j_2 \geq [h_i/2] \text{ or } j_1 \geq [h_i/2], j_2 \leq [h_i/2] \end{cases} \qquad (i = 1,\ldots,r) \quad (4.17)$$

Similarly it can be shown that

$$Power_{any-pair\ \Delta} = 1 - \int_0^{\infty}\prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\prod_{j:m_{i,j-0}\leq-\Delta}\Phi\left(-\sqrt{l}y + \sqrt{1+l}d_{2a}\sqrt{u} + \frac{\sqrt{n_i}m_{i,j-0}}{s}\right)x\right.$$

$$\left. x\prod_{j:m_{i,j-0}\geq\Delta}\Phi\left(\sqrt{l}y + \sqrt{1+l}d_{2a}\sqrt{u} - \frac{\sqrt{n_i}m_{i,j-0}}{s}\right)j(y)dy\right]h(u)du$$

$$\geq 1 - \int_0^{\infty}\prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\Phi^{k_i}\left(\sqrt{l}y + \sqrt{1+l}d_{2a}\sqrt{u} - \frac{\sqrt{n_i}\Delta}{s}\right)j(y)dy\right]h(u)du$$

$$\geq 1 - \int_0^{\infty}\prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\Phi^{g_i}\left(\sqrt{l}y + \sqrt{1+l}d_{2a}\sqrt{u} - \frac{\sqrt{n_i}\Delta}{s}\right)j(y)dy\right]h(u)du$$

$$= 1 - P\left(D_{2i1} < d_{2a},\ldots,D_{2ig_i} < d_{2a}\ \forall i: g_i \geq 1 \,\middle|\, m_{i,1-0} = \ldots = m_{i,g_i-0} = \Delta\ \forall i: g_i \geq 1\right). \quad (4.18)$$

Therefore, $m_{i,1-0} = \ldots = m_{i,g_i-0} = \Delta,\ m_{i,(g_i+1)-0} = \ldots = m_{i,c-0} = 0$ (*i* = 1, …, *r*) is a LFC for the any-pair Δ power in the two-sided testing situation. This is exactly the same LFC as in the one-sided testing situation. (see also formula (3.26))

Under the assumption that the variance s^2 is known, i.e. assuming infinite degrees of freedom, explicit formulas of n_i can again be obtained.

$$Power_{all-pairs\ \Delta} \geq P\Big(D_{2i1} < -d_{2a}, \ldots, D_{2i[h_i/2]} < -d_{2a}, -D_{2i[h_i/2]+1} < -d_{2a}, \ldots, -D_{2ih_i} < -d_{2a}\ \forall i\ \Big|$$

$$m_{i,1-0} = \ldots = m_{i,[h_i/2]-0} = -\Delta,\ m_{i,([h_i/2]+1)-0} = \ldots = m_{i,h_i-0} = \Delta\ \forall i\Big) =$$

$$= P\left(Z_{2ij} < \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_{2a}\ j = 1 \text{ to } [h_i/2],\ -Z_{2ij} < \frac{\Delta\sqrt{n_i}}{s\sqrt{1+l}} - d_{2a}\ j = [h_i/2]+1 \text{ to } h_i;\ \forall i\right) \quad (4.19)$$

where the vector $\left(Z_{211}, \ldots, Z_{21[h_1/2]}, -Z_{1[h_1/2]+1}, \ldots, -Z_{21h_1}, \ldots, Z_{2r1}, \ldots, Z_{2r[h_r/2]}, -Z_{2r[h_r/2]+1}, \ldots, -Z_{2rh_r}\right)$ is distributed as a standardized *h*-variate normal random variable with correlation matrix $\mathbf{H}_h$ as defined above (4.17).

Similar to the one-sided testing situation, it is easy to see that under the condition of equal sample sizes across all strata, i.e. $n_i = n$, and treating the problem symmetrically with regard to all hypotheses with $m_{i,j-0} \geq \Delta$, n is the smallest integer with

$$n \geq (1+l)\left(d_{2a} + x_{h,\mathbf{0},\mathbf{H}_h;1-b}\right)^2 s^2 / \Delta^2 \quad (4.20)$$

where $x_{h,\mathbf{0},\mathbf{R}_h;1-b}$ is the $1-b$ equi-percentage point of an *h*-variate standardized normal distribution with correlation matrix $\mathbf{H}_h$. Notice that d_{2a} is the critical value based on an infinite number of degrees, i.e. $d_{2a} = |t|_{rc,\infty,\mathbf{R};1-a}$; in fact it is the two-sided 1 - a equi-percentage point of the *rc*-variate standard normal distribution with correlation matrix $\mathbf{R}$, denoted as $d_{2a} = |x|_{rc,\mathbf{0},\mathbf{R};1-a}$.

Without any a priori knowledge about the unknown number of differences $m_{i,j-0} \geq \Delta$, *h* has to be replaced by *rc*.

The LFC for the any-pair Δ power in case of the two-sided testing situation is identical to the LFC of the one-sided testing. Therefore the same formula (3.31) applies here to determine the smallest n such that

$$n \ge (1+l)\left(d_{2a} - x_{g,\mathbf{0},\mathbf{R}_g;b}\right)^2 s^{\,2}/\Delta^2 \tag{4.21}$$

However, notice that the percentage point d_a has been replaced by d_{2a}.

Hsu (1989) used a simultaneous confidence interval method to determine the sample sizes. Horn and Vollandt (2000) noted that his requirements are more strict because he required that all simultaneous 1 - a confidence intervals should cover their corresponding parameter differences with probability ≥ 1 - b and are sufficiently narrow to ensure that zero is not included. The power definition used here requires this only for those differences worth detecting, which results in smaller sample sizes.

Step-down procedure

The step-down procedure proposed by Cheung and Holland (1994) and described in Section 3.5 for the one-sided testing situation can easily be adapted to the two-sided test situation. Instead of ordering the observed one-sided test statistics d_{ij}, the observed two-sided test statistics d_{2ij} should be ordered and the upper percentage points $t_{mn,\mathbf{R}_{(m)};1-a}$ have to be replaced by $|t|_{m,n,\mathbf{R}_{(m)};1-a}$.

This results in the following step-down testing procedure for the two-sided testing situation:

- Order all observed test statistics d_{2ij}'s from smallest to largest, say $d_{2(1)} \le d_{2(2)} \le \ldots \le d_{2(rc)}$. Let $H_{0(1)}$, $H_{0(2)}$, …,$H_{0(rc)}$ be the corresponding null-hypotheses and let $E_{(m)}$ be the subset of indices *ij*'s corresponding to the *m* smallest d_{2ij}'s (*m* = 1, …, *rc*). Thus $E_{(rc)}$ is the set of all indices and $E_{(1)}$ refers to the indices corresponding to $d_{2(1)}$.

  Denote with $\mathbf{R}_{(m)}$ the sub-matrix of the correlation matrix $\mathbf{R}$ restricted to $E_{(m)}$ (*m* = 1, …, *rc*).
- Start with testing $H_{0(rc)}$ and reject $H_{0(rc)}$ if $d_{2(rc)} > |t|_{rcn,\mathbf{R}_{(rc)};1-a}$; otherwise retain all sub-hypotheses without further tests.
- The general step *m* is, reject $H_{0(m)}$ if $H_{0(rc)}$, ..., $H_{0(m+1)}$ are rejected and $d_{2(m)} > |t|_{m,n,\mathbf{R}_{(m)};1-a}$.

  If $H_{0(m)}$ is not rejected, then also retain $H_{0(m-1)}$, …,$H_{0(1)}$ without any further testing (*m* = 1, …, *rc*).

The p-value version can also easily be adapted to the two-sided testing situation by defining the adjusted p-value for $H_{0(m)}$ as

$$p_{2(m)} = \max\left\{\tilde{p}_{2(m)}, \tilde{p}_{2(m+1)}, \ldots, \tilde{p}_{2(rc)}\right\} \text{ } (m = 1, \ldots, rc). \qquad (4.22)$$

where

$$\tilde{p}_{2(m)} = P\left(\text{at least one } D_{2ij} > d_{2(m)},\ \{\text{ij}\} \in E_{(m)}\right) = 1 - P\left(D_{2ij} \le d_{2(m)},\ \{\text{ij}\} \in E_{(m)}\right) \quad (4.23)$$

$$(m = 1, \ldots, rc)$$

(Remember the note of controlling both the Type I and Type III FWE given in the last paragraph of Section 4.1.)

Computations of the example

The analysis described in Section 3.2, using the same example, is now illustrated for the two-sided testing situation

The two-sided testing situation for the example is as follows:

$$H_0 : m_{10} = m_{11} = m_{12} \text{ and } m_{20} = m_{21} = m_{22}$$

versus

$$H_1 : m_{11} \ne m_{10} \text{ or } m_{12} \ne m_{10} \text{ or } m_{21} \ne m_{20} \text{ or } m_{22} \ne m_{20}$$

where i = 1 and i = 2 represent the males and females respectively, and where j = 1 and j = 2 represent the low and high dose respectively.

The results of the analysis are presented in the following table.

Table 4.1 Two-sided adjusted p-values and simultaneous confidence intervals

Stratum	Contrast	Estimate	Adjusted p-value $\tilde{p}_{2ij}$	Two-sided 95% Confidence interval
M	Plac-Low	0.864	0.140	(-0.187, 1.914)
	Plac-High	2.163	<0.001	(0.996, 3.330)
F	Plac-Low	0.582	0.516	(-0.519, 1.682)
	Plac-High	1.265	0.029	(0.098, 2.433)

The program code can be found in program Ch4.sas of Appendix 3, which also computes the two sided upper percentage point $d_{2,0.05} = d_2(0.05,2,2,37,\{b_{ij}\}) = |t|_{4,37,\mathbf{R};0.95} = 2.601$.

Basically, the conclusions drawn from this table are identical to the conclusions found treating the testing problem as a one sided testing problem. The high dosage in both genders is statistically significantly (p<0.05) different from placebo. Because $d_{11} = 2.163 > d_{2,0.05}$ and $d_{22} = 1.265 > d_{2,0.05}$ it is allowed to say that high dosage is superior to placebo in both genders while controlling the Type I and Type III FWE (see formula (4.3)).
However, it is obvious that the two-sided adjusted p-values are larger than the one-sided p-values. (See the next section as well.) Also, the lower bounds of the two-sided simultaneous confidence intervals are closer to zero than the lower bounds of the one-sided simultaneous confidence intervals, which is the price to be paid for obtaining upper bounds in addition. But this is expected and well known from the univariate testing situation.

4.3 Two-sided tests considered as one-sided tests

An alternative way to consider the two-sided testing situation is to express the two-sided test as a pair of one-sided tests. In this way the two-sided testing problem can be formulated as a Union Intersection (UI) multiple testing problem.

Consider the test of the global null hypothesis

$$H_0 : m_{ij} = m_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

versus the two-sided alternative hypothesis

$$H_1 : \exists ij : m_{ij} \neq m_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

This is equivalent to the simultaneous testing of the pair of one-sided hypotheses:

$$H_{01} : m_{ij} = m_{i0} \;\; (i = 1,\ldots,r \;\;\; j = 1,\ldots,c) \text{ versus } H_{11} : \exists ij : m_{ij} > m_{i0} \;\; (i = 1,\ldots r \;\;\; j = 1,\ldots,c)$$

and (4.24)

$$H_{02}: m_{ij} = m_{i0} \quad (i = 1,...,r \quad j = 1,...,c) \text{ versus } H_{12}: \exists ij: m_{ij} < m_{i0} \quad (i = 1,...r \quad j = 1,...,c).$$

The two-sided testing problem can be represented as a UI multiple testing problem by writing

$$H_0 = H_{01} \cap H_{02} \text{ versus } H_1 = H_{11} \cup H_{12}.$$

As stated in Section 2.1, the rejection region for H_0 is given by the rejection regions for H_{01} and H_{02}, so H_0 is rejected if and only if at least H_{01} or H_{02} is rejected.

Suppose an error rate a_1 for H_{01} and a separate error rate a_2 for H_{02} has been chosen. If the two alternatives H_{11} and H_{12} were disjoint, i.e. cannot be rejected simultaneously, then the error rate of the two-sided test of H_0 would be equal to $a = a_1 + a_2$. Notice that this is true for a univariate two-sided test written as a pair of one-sided tests. Unfortunately, the alternatives H_{11} and H_{12} are not disjoint. However, using the same error rate for both one-sided tests and applying the Bonferroni method provides a conservative solution for the two-sided situation. Thus multiplying an one-sided adjusted p-value < 0.5 by two results in a conservative two-sided adjusted p-value. See for example Dunnett and Gent (1996) for more details.

This is also illustrated by the example used in the previous section. The adjusted one-sided p-values computed in Section 3.2 multiplied by two are close but larger than the adjusted two-sided p-values.

There are attempt in literature to bridge the gap between one-sided procedures and two-sided procedures. For example Hayter, Miwa and Liu (2000) proposed a procedure that combines the advantages of the one-sided and two-sided procedures for comparing several treatments with a control for the situation of a one-way layout. It has the advantage of the two-sided procedure to provide both upper and lower limits on the differences between each treatment and control. In addition it declares treatments better than control based on the sharper inferences of the one-sided procedure.

This procedure can be extended to the stratified two-way layout, though this won't be discussed here any further.

5 Non-inferiority and equivalence testing

The multiple comparison procedure of Chapter 3 describes the many-to-one comparison situation that at least one of the active treatments is superior to control in any of the strata and Chapter 4 discusses the two-sided testing situation that at least one active treatment is different from control. These so called 'superiority' type of trials are most convincingly to establish efficacy according to the recent International Conference of Harmonization (ICH) guidance document ‘E9 Statistical Principles for Clinical Trials’. The control treatment is usually a real placebo. For serious illnesses, a placebo may be considered unethical if a therapeutic treatment exists which has proven efficacious in relevant superiority trial(s). In that case, the scientifically sound use of an active treatment as a control should be considered. Then the superiority type of trial is not always appropriate or feasible.

Active control trials designed to show that the efficacy of an investigational product is not relevantly worse than that of the active comparator are called 'non-inferiority' trials.

Another type of trial is the ‘equivalence‘ trail, which is designed to confirm the absence of a meaningful difference between the treatments. This kind of trial is very common to investigate the bioavailability and pharmacokinetic properties of the active substance from a pharmaceutical product.

This short chapter illustrates how to perform many-to-one comparisons in a stratified two-way layout for some of these types of trials. Section 5.1 describes the non-inferiority setting and Section 5.2 describes the equivalence setting.

5.1 Non-inferiority

The non-inferiority testing situation looks similar to the one-sided superiority testing situation as described in Chapter 3. Instead of showing that an active treatment is superior to the control treatment, one should show that the active treatment is not relevantly worse or so called non-inferior than the control treatment.

It is assumed that the standard assumptions of Chapter 3 are still valid, i.e. the sample values $\{X_{ijk}\}$ are independently normal distributed with mean m_{ij} and variance s^2, S^2 is the usual pooled variance estimator of s^2 based on n degrees of freedom, there are *c* active treatments within each of the *r* strata and a positive value of the difference between the active and control treatment occurs when the active treatment is superior to the control treatment.

Then the global null hypothesis to be tested is that all treatments are inferior to the control treatment

$$H_0^{\Delta} : m_{ij} - m_{i0} \leq -\Delta \ \ (i = 1,\ldots,r \quad j = 1,\ldots,c) \tag{5.1}$$

versus the alternative hypothesis that at least one of the treatments in any of the strata is non-inferior to control

$$H_1^{\Delta} : \exists ij : m_{ij} - m_{i0} > -\Delta \ \ (i = 1,\ldots,r \quad j = 1,\ldots,c)$$

where $\Delta > 0$ represents the minimum difference considered to be relevant.

The symbol Δ has also been used in the determination of the sample sizes for the one-sided superiority testing problem in Section 3.4. Although both symbols have a similar interpretation, they play a slightly different role. In the sample size calculations it is used to introduce a difference that is worthwhile to detect up-front, i.e. to determine the power. Here it represents a difference that should really be exceeded by the treatment effect compared to placebo.

Notice that the one-sided superiority testing situation can be considered as a special case of the non-inferiority testing problem by taking $\Delta = 0$ (see formula (3.2)).

The proposed test statistic for the non-inferiority many-to-one multiple testing problem is

$$D^{\Delta} = \max_{1 \leq i \leq r; 1 \leq j \leq c} \left\{ D_{ij}^{\Delta} \right\} \text{ with } D_{ij}^{\Delta} = \frac{\overline{X}_{ij} - \overline{X}_{i0} + \Delta}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \ \ (i = 1,\ldots,r \ \ j = 1,\ldots,c) \tag{5.2}$$

It can be shown that the joint distribution of the D_{ij}^{Δ}'s follows a *rc*-variate central t-distribution with n degrees of freedom and correlation matrix $\mathbf{R}$ under the null hypothesis and an *rc*-variate noncentral t-distribution with the same degrees of freedom and correlation matrix $\mathbf{R}$ and noncentrality vector $\mathbf{d}^{\Delta} = \left(d_{ij}^{\Delta} \right)_{1 \leq i \leq r; 1 \leq j \leq c} = \left(\dfrac{m_{ij} - m_{i0} + \Delta}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \right)_{1 \leq i \leq r; 1 \leq j \leq c}$ under the alternative hypothesis. The correlation matrix $\mathbf{R}$ is the same block diagonal correlation matrix defined in formula (3.6) for the one-sided superiority testing problem.

Therefore, results such as the computation of adjusted p-values, simultaneous confidence intervals and sample sizes as discussed in Chapter 3 for the one-sided superiority setting are easily derived for the non-inferiority situation as well. Simply use the statistics D_{ij}^{Δ}'s instead of the statistics D_{ij}'s as defined in formula (3.3). For this reason they are not discussed here any further.

Notice that in contrast to the superiority testing problem the value of Δ is required in the test statistic and therefore should be explicitly known beforehand. In case it is impossible to define Δ a-priori, the use of one-sided simultaneous confidence intervals based on the statistics D_{ij}'s might be an option: the sub-hypothesis $H_{0ij}^{\Delta}: m_{ij} - m_{i0} \leq -\Delta$ is rejected in favor of $H_{1ij}^{\Delta}: m_{ij} - m_{i0} > -\Delta$ at level a if and only if the lower bound of the 100(1-a)% simultaneous confidence interval for $m_{ij} - m_{i0}$ exceeds $-\Delta$.

So it is clear that there exists a close relationship between non-inferiority testing problems and superiority testing problems. Dunnett and Gent (1996) utilized this relationship as well. They illustrated that in case of comparing one active treatment against a control treatment, when non-inferiority of the active treatment was shown, a conditional analysis could be performed to establish superiority of the active treatment without any multiple comparison adjustment. This can also be demonstrated by making use of a confidence interval for the true difference between the active and control treatment means. When the lower bound exceeds $-\Delta$, non-inferiority can be concluded and when the lower bound also exceeds 0, superiority can be concluded.

Kieser (1995) and Bauer and Kieser (1996) investigated this relationship in a much broader context. Among other families, they considered the family of null hypotheses with elements H_0^{Δ}, where Δ falls in a relevant interval, say between -*L* < 0 and -*U* > 0, and defined a multiple testing procedure for this family of hypotheses based on the closed testing principle (Marcus et. al. (1976); Section 3.5) using the test statistics D^{Δ}.

For practical details about superiority and non-inferiority see for example the document 'Points to Consider on Switching between Superiority and Non-inferiority' of the Committee for Proprietary Medicinal Products (CPMP).

5.2 Equivalence

In equivalence trials the issue is no longer to detect a difference between the treatments but to demonstrate that the treatments are equivalent within an a priori stipulated equivalence range defining acceptance values for the differences between the treatments. Keep in mind that failing to reject the null hypothesis in a trial designed to detect significant differences doesn't show equivalence; '*absence of evidence is not evidence of absence*' as expressed by Altman and Bland (1995). See for example Chapters 15 and 22 of Senn (1997) for an introduction concerning other issues of equivalence studies.

Hauschke (1999) described the many-to-one comparison for global equivalence testing in the situation of a one-way layout. Global means that all active treatments should be equivalent to the control treatment. This section demonstrates how to proceed in the stratified two-way layout. Such a testing problem is not purely hypothetical but may occur for example in a dose ranging carcinogenicity study conducted in both male and female animals were one should proof that all experimental dosages are safe, i.e. equivalent to the standard/control treatment.

The global null hypothesis to be tested is

$$H_0^{\Delta_{12}} : \exists ij : m_{ij} - m_{i0} \leq \Delta_1 \text{ or } m_{ij} - m_{i0} \geq \Delta_2 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c) \qquad (5.3)$$

against the alternative hypothesis that all active treatments in any of the strata are equivalent to control

$$H_1^{\Delta_{12}} : \Delta_1 < m_{ij} - m_{i0} < \Delta_2 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

where $\left(\Delta_1, \Delta_2\right)$, $\Delta_1 < 0 < \Delta_2$ describes the area of irrelevant differences.

Notice that any of the *rc* two-sided sub-hypotheses

$$H_{0ij}^{\Delta_{12}} : m_{ij} - m_{i0} \leq \Delta_1 \text{ or } m_{ij} - m_{i0} \geq \Delta_2$$

against (5.4)

$$H_{1ij}^{\Delta_{12}} : \Delta_1 < m_{ij} - m_{i0} < \Delta_2$$

can be expressed by two one-sided sub-hypotheses:

$$H_{0ij}^{\Delta_1}: m_{ij} - m_{i0} \le \Delta_1 \qquad \text{against} \qquad H_{1ij}^{\Delta_1}: m_{ij} - m_{i0} > \Delta_1$$

$$\text{and} \tag{5.6}$$

$$H_{0ij}^{\Delta_2}: m_{ij} - m_{i0} \ge \Delta_2 \qquad \text{against} \qquad H_{1ij}^{\Delta_2}: m_{ij} - m_{i0} < \Delta_2$$

Note that the global null hypothesis $H_0^{\Delta_{12}}$ is the hypothesis that there exists at least an active treatment in any of the strata, which is superior to the control treatment by at least Δ_2 or inferior to the control treatment by at least Δ_1. Thus the global null hypothesis $H_0^{\Delta_{12}}$ is the union of all one-sided sub-hypotheses $H_{0ij}^{\Delta_1}$ and $H_{0ij}^{\Delta_2}$, i.e.

$$H_0^{\Delta_{12}} = \bigcup_{i=1}^{r}\bigcup_{j=1}^{c}\bigcup_{k=1}^{2} H_{0ij}^{\Delta_k} \ . \tag{5.7}$$

Similarly, the alternative hypothesis $H_1^{\Delta_{12}}$ is the intersection of all one-sided sub-hypotheses $H_{1ij}^{\Delta_1}$ and $H_{1ij}^{\Delta_2}$, i.e.

$$H_1^{\Delta_{12}} = \bigcap_{i=1}^{r}\bigcap_{j=1}^{c}\bigcap_{k=1}^{2} H_{1ij}^{\Delta_k} \ . \tag{5.8}$$

The one-sided null sub-hypothesis $H_{0ij}^{\Delta_1}$ and $H_{0ij}^{\Delta_2}$ can be tested using test statistics (5.2) as proposed for the non-inferiority testing problem in the previous section.

The null sub-hypothesis $H_{0ij}^{\Delta_1}$ is rejected at level a if

$$D_{ij}^{\Delta_1} = \frac{\bar{X}_{ij} - \bar{X}_{i0} - \Delta_1}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \ge t_{1-a,n} \tag{5.9}$$

and the null sub-hypothesis $H_{0ij}^{\Delta_2}$ is rejected at level a if

$$D_{ij}^{\Delta_2} = \frac{\bar{X}_{ij} - \bar{X}_{i0} - \Delta_2}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \leq -t_{1-a\,n} \,. \tag{5.10}$$

According to the Intersection-Union (IU) principle described by Berger (1982), the global null hypothesis $H_0^{\Delta_{12}}$ can be rejected at level a in favor of the alternative hypothesis $H_1^{\Delta_{12}}$ if and only if all one-sided sub-hypotheses $H_{0ij}^{\Delta_1}$ and $H_{0ij}^{\Delta_2}$ are rejected at level a .

Thus the global hypothesis $H_0^{\Delta_{12}}$ is rejected if and only if $D_{ij}^{\Delta_1} \geq t_{1-a,n}$ and $D_{ij}^{\Delta_2} \leq -t_{1-a,n}$ for all *ij*'s.

Equivalently, one can make use of simultaneous confidence intervals. As mentioned in the previous section, it avoids the use of $D_{ij}^{\Delta_1}$ and $D_{ij}^{\Delta_2}$ that requires the specification of the Δ_1 and Δ_2 beforehand.

The global null hypothesis $H_0^{\Delta_{12}}$ is rejected in favor of the alternative hypothesis $H_1^{\Delta_{12}}$ at level a if all two-sided 100(1-2a)% confidence intervals for the differences $m_{ij} - m_{i0}$ are contained in the equivalence interval (Δ_1, Δ_2):

$$\left(\bar{X}_{ij} - \bar{X}_{i0} - t_{1-a\,n} s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}, \bar{X}_{ij} - \bar{X}_{i0} + t_{1-a,n} s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}\right) \subset (\Delta_1, \Delta_2) \tag{5.11}$$

for all *i* = 1, …, *r* and *j* = 1, …, *c*.

6 Ratio testing situation

In the previous chapters the testing situations are all formulated in terms of differences between the population means of the active treatment and the control treatment. However there are testing situations where it is more appropriate to express the testing problem in terms of proportions of the population means rather than differences. In particular in equivalence testing situations it is more common to express the equivalence limits as proportions of the population means. If the observations follow a lognormal distribution, then there is international consensus that equivalence should be assessed on the logarithmic scale. Taking logarithms transforms the ratio testing situation back to the testing situation expressed in differences.

Nevertheless, there are also many situations for which the normality assumption for the original variable is justified. Hauschke, Kieser et al. (1999) showed the example of the assessment of therapeutic equivalence for two inhalers applied for the relief of asthma attacks using the morning peak expiratory flow rate as a measure of airflow obstruction and the example of pharmacokinetic characteristic AUC for topical dermatological corticosteriods where the assumption of normality is acceptable without log-transforming the original data.

Hauschke, Kieser et al. (1999) and Kieser and Hauschke (1999) described the problem of equivalence testing based on the ratio of two means and Hauschke (1999) described the equivalence testing situation of many-to-one comparisons for a one-way layout.

This chapter describes the many-to-one comparisons for the stratified two-way layout based on ratios. The one-sided testing situation is discussed in the first section and the two-sided testing situation is discussed in the second section.

6.1 One-sided testing situation

The standard assumptions are maintained: there are c active treatments within each of the r strata, $X_{ijk} \sim N(m_{ij}, s^2)$ and S^2 is the usual pooled variance estimator of s^2 based on n degrees of freedom.

Then the many-to-one comparisons for the stratified situation in terms of ratios can be formulated for the one-sided testing problems as:

$$H_0^q : \frac{m_{ij}}{m_{i0}} = q \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

versus (6.1)

$$H_1^q : \exists ij : \frac{m_{ij}}{m_{i0}} > q \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

where $q > 0$.

Notice that this formulation covers both non-inferiority testing as well as superiority testing. Assume that a higher value of the ratio $\frac{m_{ij}}{m_{i0}}$ occurs when the active treatment is better then the control treatment. Then for non-inferiority testing, the value $q < 1$ represents the smallest value of the ratio still to be considered as relevant. (See also Section 5.1) To perform superiority testing, the value $q = 1$ can be taken. Although Kieser and Hauschke (1999) mentioned that it is suggested not to perform a test against perfect equality but to use a threshold value $q > 1$ expressing a relevant improvement of the active treatment over the control treatment.

To see the similarity between the testing problem phrased in terms of ratios and the testing problem phrased in terms of differences, the above testing problem can also be written as:

$$H_0^q : m_{ij} - q m_{i0} = 0 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

versus (6.2)

$$H_1^q : \exists i, j : m_{ij} - q m_{i0} > 0 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c),$$

under the assumption that $m_{i0} > 0$.

The sign symbol '>' should be changed into '<' if $m_{i0} < 0$ is assumed. Without such a restriction on m_{i0} it is not possible. In the remainder of this section it is assumed that $m_{i0} > 0$. This assumption is not a real burden in practical problems, because if it is unclear whether $m_{i0} > 0$ or $m_{i0} < 0$, then is seems more logical to test against a two-sided alternative hypothesis. This will be described in Section 6.2.

Sasabuchi (1988) proposed to define test statistics T_{ij}^q, which are in a way similar way to the test statistics D_{ij} for the testing problem defined in terms of differences:

$$T_{ij}^q = \frac{\overline{X}_{ij} - q\overline{X}_{i0}}{s\sqrt{n_{ij}^{-1} + q^2 n_{i0}^{-1}}} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c) \qquad (6.3)$$

Notice that in the situation of $q = 1$, the test statistics T_{ij}^q's are identical to the test statistics D_{ij}'s (3.3).

Then to test the global null-hypothesis H_0^q the test statistic T^q can be used, where T^q is defined as:

$$T^q = \max_{1 \le i \le r; 1 \le j \le c} \{T_{ij}^q\}. \qquad (6.4)$$

Sasabuchi (1988) showed that the test statistic T_{ij}^q follows a Student t-distribution under the null-hypothesis.

Similar to showing that the D_{ij}'s are jointly multivariate t distributed (see Sections 2.1 and 3.1) it can be shown that the vector $\left(T_{11}^q, \ldots, T_{rc}^q\right)^`$ follows under the null-hypothesis H_0^q a central *rc*-variate t-distribution with correlation matrix $\mathbf{R}^q$ and n degrees of freedom.

The correlation matrix $\mathbf{R}^q$ is a block diagonal matrix with the following structure:

$$\mathbf{R}^q = \begin{pmatrix} \mathbf{R}_1^q & .. & 0 \\ .. & .. & .. \\ 0 & .. & \mathbf{R}_r^q \end{pmatrix} \text{ where } \mathbf{R}_i^q = \begin{pmatrix} 1 & r_{i(1,2)}^q & .. & r_{i(1,c)}^q \\ r_{i(2,1)}^q & 1 & .. & .. \\ .. & .. & 1 & r_{i(c-1,c)}^q \\ r_{i(c,1)}^q & .. & r_{i(c,c-1)}^q & 1 \end{pmatrix} \quad (i = 1, \ldots, r) \qquad (6.5)$$

The correlation coefficients between each pair of $T_{ij_1}^q$ and $T_{ij_2}^q$ within the same stratum *i* is given by

$$r^q_{i(j_1,j_2)} = b^q_{ij_1} b^q_{ij_2} \quad (1 \le \mathrm{j}_1 \ne j_2 \le c) \text{ where } b^q_{ij} = \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij}} + q^2}}. \tag{6.6}$$

Thus each matrix $\mathbf{R}^q_i$ satisfies the product correlation structure.

These factors b^q_{ij} can be derived as:

$$r^q_{i(j_1,j_2)} = \frac{\operatorname{cov}(T^q_{ij_1}, T^q_{ij_{12}})}{\sqrt{\operatorname{var}(T^q_{ij_1})}\sqrt{\operatorname{var}(T^q_{ij_2})}} = \frac{\operatorname{cov}(\bar{X}_{ij_1} - q\bar{X}_{i0}, \bar{X}_{ij_2} - q\bar{X}_{i0})}{\sqrt{\operatorname{var}(\bar{X}_{ij_1} - q\bar{X}_{i0})}\sqrt{\operatorname{var}(\bar{X}_{ij_2} - q\bar{X}_{i0})}} =$$

$$= \frac{q^2 \frac{s^2}{n_{i0}}}{\sqrt{\frac{s^2}{n_{ij_1}} + q^2 \frac{s^2}{n_{i0}}}\sqrt{\frac{s^2}{n_{ij_2}} + q^2 \frac{s^2}{n_{i0}}}} = \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij_1}} + q^2}} \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij_2}} + q^2}} = b^q_{ij_1} b^q_{ij_2}$$

Notice that in the situation of $q = 1$, the correlation coefficients $r^q_{i(j_1,j_2)}$ coincide with the correlation coefficients $r_{i(j_1,j_2)}$ defined for the testing problem in terms of differences (see formula (3.5)).

<u>Percentage points</u>

The upper percentage point $d(a, r, c, n, \{b^q_{ij}\})$ such that $P_{H_0}\left(T^q \le d(a, r, c, n, \{b^q_{ij}\})\right) = 1 - a$ is the $1 - a$ percentage point of the central *rc*-variate t-distribution with correlation matrix $\mathbf{R}^q$ and n degrees of freedom which is denoted as $t_{rc,n,\mathbf{R}^q;1-a}$.

The same algorithms as described for the testing problem defined in terms of differences can be used to calculate these upper percentage points by simply replacing $\mathbf{R}$ by $\mathbf{R}^q$ or equivalently by replacing $b_{ij} = \sqrt{\frac{n_{ij}}{n_{i0} + n_{ij}}}$ by $b^q_{ij} = \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij}} + q^2}}$.

And thus, the upper percentage points can be easily computed within SAS using the PROBMC function and the subroutine QUAD within PROC IML. (See Section 3.2).

Notice for example that $P\left(T^q \le t\right)$ can be written as:

$$P\left(T^q \le t\right) = P\left(\max_{1\le i\le r;1\le j\le c}\left\{T_{ij}^q\right\} \le t\right) = \int_0^\infty P\left(\max_{1\le i\le r;1\le j\le c}\left\{\frac{\bar{X}_{ij} - q\,\bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + q^2 n_{i0}^{-1}}}\right\} \le t\sqrt{u}\right) h(u)du =$$

$$= \int_0^\infty \prod_{i=1}^r P\left(\max_{1\le j\le c}\left\{\frac{\bar{X}_{ij} - q\,\bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + q^2 n_{i0}^{-1}}}\right\} \le t\sqrt{u}\right) h(u)du = \int_0^\infty \prod_{i=1}^r P\left(\frac{\bar{X}_{ij} - q\,\bar{X}_{i0}}{s\,\sqrt{n_{ij}^{-1} + q^2 n_{i0}^{-1}}} \le t\sqrt{u}\ ;\ \forall j\right) h(u)du =$$

$$= \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j=1}^c \Phi\left(\frac{b_{ij}^q y + t\sqrt{u}}{\sqrt{1-\left(b_{ij}^q\right)^2}}\right) j(y)dy\right] h(u)du \qquad (6.7)$$

where *h(u)* is the density function of a c_n^2/n distributed variable and $\Phi(y)$ and $j(y)$ are the standard cumulative distribution function and probability density function respectively.
This is identical to formula (3.7) using the b_{ij}^q's instead of the b_{ij}'s.

Adjusted p-values

Define the null sub-hypothesis H_{0ij}^q as

$$H_{0ij}^q : \frac{m_{ij}}{m_{i0}} = q$$

and the alternative sub-hypothesis H_{1ij}^q as (6.8)

$$H_{1ij}^q : \frac{m_{ij}}{m_{i0}} > q\ .$$

Obviously, the global null hypothesis H_0^q can be written as the intersection of the sub null hypotheses $H_0^q = \bigcap_{ij} H_{0ij}^q$ and the global alternative hypothesis H_1^q can be written as the union of the sub alternative hypotheses $H_1^q = \bigcup_{ij} H_{1ij}^q$. Thus the problem of testing the global hypothesis can also be seen as a Union-Intersection multiple testing problem.

Then analogue to the testing problem in terms of differences, adjusted p-values $\tilde{p}_{ij}^{q}$ corresponding to the sub-hypothesis H_{0ij}^{q} can be computed as:

$$\tilde{p}_{ij}^{q} = \min\left\{a \;\middle|\; H_{0ij}^{q} \text{ is rejected at FWE} = a\right\} = P_{H_0}\left(T^{q} > t_{ij}^{q}\right) = 1 - T_{rc}\left(-\infty, \mathbf{t}_{ij}^{q}; n, \mathbf{R}^{q}\right) \quad (6.9)$$

where t_{ij}^{q} is the observed value of the test statistic T_{ij}^{q} (i = 1, …, r and j = 1, …, c) and $T_{rc}\left(-\infty, \mathbf{t}_{ij}^{q}; n, \mathbf{R}^{q}\right)$ is the cumulative density function of a central *rc*-variate t-distribution with correlation matrix $\mathbf{R}^{q}$ and n degrees of freedom.

In case of $q = 1$, the test statistic T_{ij}^{q} and correlation matrix $\mathbf{R}^{q}$ coincide with the test statistic and correlation matrix defined for the testing problem in terms of differences. Hence the adjusted p-values $\tilde{p}_{ij}^{q}$ are identical to the adjusted p-values $\tilde{p}_{ij}$ defined for the testing problem in terms of differences (see formula (3.9)).

<u>100(1-α)% simultaneous confidence intervals</u>

The derivation of simultaneous confidence intervals for the ratios $\dfrac{m_{ij}}{m_{i0}}$ is more complicated than expected in first instance. The reason why is illustrated below.

Fieller (1954) derived a two-sided confidence interval for the ratio $q = \dfrac{m_X}{m_Y}$ of mean values of two independent normal distributed random variables X and Y with a common variance s^2. Let $\bar{X}$ and $\bar{Y}$ denote the observed means based on n_X and n_Y observations respectively and let s^2 be the estimate of s^2 based on n degrees of freedom. Then Fieller showed that the bounds of the two-sided 100(1-α)% confidence interval for $\dfrac{m_X}{m_Y}$ are given by

$$\frac{\bar{X}\bar{Y} \pm \sqrt{t_{n;1-a/2}^{2} s^{2}\left(n_Y^{-1}\bar{X}^{2} + n_X^{-1}\bar{Y}^{2} - t_{n;1-a/2}^{2} s^{2} n_X^{-1} n_Y^{-1}\right)}}{\bar{Y}^{2} - t_{n;1-a/2}^{2} s^{2} n_Y^{-1}}, \quad (6.10)$$

under the restriction that $d = \bar{Y}^2 - t^2_{n,1-a} s^2 n_Y^{-1} > 0$.

In his proof Fieller made use of the theorem that the set of values of q_0, which are not rejected testing the null hypothesis $H_0 : \frac{m_X}{m_Y} = q_0$ against the two-sided alternative hypothesis $H_1 : \frac{m_X}{m_Y} \neq q_0$ at level a, provides just a two-sided 100(1-α)% confidence interval for the ratio $q = \frac{m_X}{m_Y}$.

The intention is to use this approach for the multiple testing situation as well in order to compute one-sided simultaneous confidence intervals, i.e. the set of values of $\mathbf{?_0} = \left(q_{011},\ldots,q_{0rc}\right)$ that are not rejected testing $H^q_{0ij} : \frac{m_{ij}}{m_{i0}} = q_{0ij}$ against $H^q_{1ij} : \frac{m_{ij}}{m_{i0}} > q_{0ij}$ simultaneously at level a should provide a one-sided 100(1-α)% confidence interval for the ratios $\frac{m_{ij}}{m_{i0}}$ (i = 1, …, r j = 1, …, c). This would translate into the problem to compute upper-percentage points of the joint distribution of the test statistics $T_{ij}^{q_{0ij}}$ and although it is a multivariate t-distribution, its correlation matrix depends on $\mathbf{?_0}$, and thus the percentage points depend on $\mathbf{?_0}$ as well. The vector $\mathbf{?_0}$ is not necessarily restricted to a subspace of $\square^{rc}$ and therefore even no worst-case situation can be found which could be used.

This problem hasn't received much attention in literature and therefore might be an interesting topic for further research in future.

The proposal at the moment is to use the conservative Bonferroni adjustment procedure.
(See Hauschke (1999) p.70 and also Jensen (1989) who used the Sidák inequality to provide simultaneous confidence intervals for the two-sided testing situation.)
Doing so, conservative approximated one-sided 100(1-a)% simultaneous confidence intervals for $\frac{m_{ij}}{m_{i0}}$ (i = 1, …, r j = 1, …, c) are given by

$$\left(\frac{\bar{X}_{ij}\bar{X}_{i0} - \sqrt{a_{i0}\bar{X}^2_{ij} + a_{ij}\bar{X}^2_{i0} - a_{i0}a_{ij}}}{\bar{X}^2_{i0} - a_{i0}}, \infty \right) \qquad \text{(i = 1, …, r j = 1, …, c)} \qquad (6.11)$$

where $a_{i0} = s^2 n_{i0}^{-1} t_{1-a/rc,n}^2$ and $a_{ij} = s^2 n_{ij}^{-1} t_{1-a/rc,n}^2$, under the condition that $\bar{X}_{i0}^2 > a_{i0}$.

Power

Similar to the problem phrased in terms of differences (Section 3.4) it can be shown that under the alternative hypothesis H_1^q the joint distribution of the T_{ij}^q's is an *rc*-variate noncentral t-distribution with correlation matrix $\mathbf{R}^q$, n degrees of freedom and noncentrality vector

$$J = \left(J_{ij}\right)_{1\le i\le r;1\le j\le c} = \left(\frac{m_{ij} - q m_{i0}}{s\sqrt{n_{ij}^{-1} + q^2 n_{i0}^{-1}}} \right)_{1\le i\le r;1\le j\le c}.$$

Therefore the global power, the all-pairs power and the any-pair power can be expressed directly in terms of probabilities of multivariate noncentral t-distributions with n degrees of freedom and noncentrality vectors characterized by the appropriate J_{ij}'s. The per-pair power can be expressed in terms of probabilities of a univariate noncentral Student t-distribution with n degrees of freedom and noncentrality parameter J_{ij}.

On the other hand, the different kind of powers can also be expressed in terms of univariate normal distribution functions.

Let S be the subset of ij's such that the null hypotheses H_{0ij}^q are false when $ij \in S$ and all remaining null hypotheses are true. Suppose there are k false null hypotheses H_{0ij}^q, i.e. the dimension of S is equal to *k*.

Then the following expressions can be derived for the all-pairs power and the any-pair power:

$$Power_{all-pairs} = P\left(T_{ij}^q > d_a^q \ \forall ij \in S\right) = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j\,ij\in S} \Phi\left(\frac{-b_{ij}^q y - d_a^q \sqrt{u} + J_{ij}}{\sqrt{1-(b_{ij}^q)^2}} \right) j(y)dy \right] h(u)du \qquad (6.12)$$

$$Power_{any-pair} = P\left(T_{ij}^q > d_a^q \ \exists ij \in S\right) = 1 - \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j\,ij\in S} \Phi\left(\frac{b_{ij}^q y + d_a^q \sqrt{u} - d_{ij}}{\sqrt{1-(b_{ij}^q)^2}} \right) j(y)dy \right] h(u)du \qquad (6.13)$$

where $\Phi(.)$ and $j(.)$ are the univariate standard cumulative distribution function and probability density function and *h(u)* is the density function of a c_n^2/n distributed variable.

Notice that these expressions are identical to the expressions derived for the testing problem defined in terms of differences (formulas (3.17) and (3.19)) if one replace d_a^q by d_a, $b_{ij}^q = \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij}} + q^2}}$ by $b_{ij} = \sqrt{\frac{n_{ij}}{n_{i0} + n_{ij}}}$ and J_{ij} by d_{ij}.

The global power and the per-pair power are not considered here because they can be seen as special cases of the any-pair power and all-pairs power.

Sample size calculations

To perform the sample size calculation in a ratio testing problem, requirs the determination of a pre-assigned minimum ratio, say Θ (with $\Theta > q$), that expresses a relevant improvement of the active treatment over the control treatment.

For example, in the 'classical' superiority testing situation with $q = 1$, a value of $\Theta = 1.2$ expresses that an improvement of 20% is considered to be a relevant improvement.

The role that Θ plays, is identical to the role of Δ, which indicates the minimum relevant difference, in the situation of a testing problem in terms of differences.

Analogue to the definition of the all-pairs Δ power and any-pair Δ power given in Section 3.4 define the all-pairs Θ power and any-pair Θ power as:

$$\text{All-pairs } \Theta \text{ power} = \text{P(reject all } H_{0ij}^q \text{ with } \frac{m_{ij}}{m_{i0}} \geq \Theta), \qquad (6.14)$$

$$\text{Any-pair } \Theta \text{ power} = \text{P(reject at least one } H_{0ij}^q \text{ with } \frac{m_{ij}}{m_{i0}} \geq \Theta).$$

Like before the sample sizes n_{ij} are only determined in the situation that all active treatment arms have the same sample size within each stratum, i.e. $n_{i1} = n_{i2} = ... = n_{ic} = n_i$, although n_{i0} may be different from n_i but the ratio is supposed to be constant for all strata, i.e. $\frac{n_i}{n_{i0}} = l$.

Denote the unknown number of ratios $\frac{m_{ij}}{m_{i0}} \geq \Theta$ by k_i and assume that a priori knowledge learns that $g_i \le k_i \le h_i$ for some lower bound integers g_i and upper bound integers h_i with $0 \le g_i \le h_i \le c$ and at least one $g_i \ge 1$. The situation of no a priori information at all is represented as the special case $g = \sum_{i=1}^{r} g_i = 1$ and $h_i = c$.

Least favorable configuration (LFC) can be determined by deriving the following lower bounds for the all-pairs Θ power and any-pair Θ power respectively:

$$Power_{all-pairs\ \Theta} \ge P\left(T_{i1}^q > d_a^q, ..., T_{ih_i}^q > d_a^q \ \forall i \,\middle|\, \frac{m_{i1}}{m_{i0}} = ... = \frac{m_{ih_i}}{m_{i0}} = \Theta \ \forall i \right) \tag{6.15}$$

$$Power_{any-pair\ \Theta} \ge 1 - P\left(T_{i1}^q < d_a^q, ..., T_{ig_i}^q < d_a^q \ \forall i : g_i \ge 1 \,\middle|\, \frac{m_{i1}}{m_{i0}} = ... = \frac{m_{ig_i}}{m_{i0}} = \Theta \ \forall i : g_i \ge 1 \right) \tag{6.16}$$

Proof:

Remember that $b_{ij}^q = \dfrac{q}{\sqrt{\dfrac{n_{i0}}{n_i} + q^2}}$, $J_{ij} = \dfrac{m_{ij}/m_{i0} - q}{s/m_{i0}\sqrt{n_i^{-1} + q^2 n_{i0}^{-1}}}$ and $l = \dfrac{n_i}{n_{i0}}$, then it follows that:

$$Power_{all-pairs\ \Theta} = \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{ij}/m_{i0} \ge \Theta} \Phi\left(\frac{-b_{ij}^q y - d_a^q \sqrt{u} + J_{ij}}{\sqrt{1-(b_{ij}^q)^2}} \right) j(y)dy \right] h(u)du =$$

$$= \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{ij}/m_{i0} \ge \Theta} \Phi\left(-\sqrt{l}qy - \sqrt{1+q^2 l}\, d_a^q \sqrt{u} + \frac{\sqrt{n_i}\left(m_{ij}/m_{i0} - q \right)}{s/m_{i0}} \right) j(y)dy \right] h(u)du \ge$$

$$\ge \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \prod_{j:m_{ij}/m_{i0} \ge \Theta} \Phi\left(-\sqrt{l}qy - \sqrt{1+q^2 l}\, d_a^q \sqrt{u} + \frac{\sqrt{n_i}\left(\Theta - q \right)}{s/m_{i0}} \right) j(y)dy \right] h(u)du \ge$$

$$\ge \int_0^\infty \prod_{i=1}^r \left[\int_{-\infty}^\infty \Phi^{h_i}\left(-\sqrt{l}qy - \sqrt{1+q^2 l}\, d_a^q \sqrt{u} + \frac{\sqrt{n_i}\left(\Theta - q \right)}{s/m_{i0}} \right) j(y)dy \right] h(u)du =$$

$$= P\left(T_{i1}^q > d_a^q, ..., T_{ih_i}^q > d_a^q \ \forall i \,\middle|\, \frac{m_{i1}}{m_{i0}} = ... = \frac{m_{ih_i}}{m_{i0}} = \Theta \ \forall i \right)$$

The proof of the any-pair Θ power is very similar and is not given here.

These expressions imply that $\frac{m_{i1}}{m_{i0}} = \ldots = \frac{m_{ih_i}}{m_{i0}} = \Theta,\ \frac{m_{i,h_i+1}}{m_{i0}} = \ldots = \frac{m_{ic}}{m_{i0}} = 0$ (*i* = 1, …, *r*) is a LFC for the all-pairs Θ power and that $\frac{m_{i1}}{m_{i0}} = \ldots = \frac{m_{ig_i}}{m_{i0}} = \Theta,\ \frac{m_{i,g_i+1}}{m_{i0}} = \ldots = \frac{m_{ic}}{m_{i0}} = 0$ (*i* = 1, …, *r*) is a LFC for the any-pair Θ power, if $g_i \le k_i \le h_i$ (*i* = 1, …, *r*).

Again, no explicit expression of n_i can be obtained if the variance s^2 is unknown and the determination of n_i should be performed iteratively. Assuming that the variance s^2 is known a priori, the formulas of n_i can be simplified. The probabilities of the LFC's can then be expressed in terms of standardized multivariate normal distributions, e.g.

$$P\left(T_{i1}^q > d_a^q, \ldots, T_{ih_i}^q > d_a^q \ \forall i \,\middle|\, \frac{m_{i1}}{m_{i0}} = \ldots = \frac{m_{ih_i}}{m_{i0}} = \Theta \ \forall i \right) = \tag{6.17}$$

$$= P\left(Z_{i1}^q > d_a^q - \frac{(\Theta - q)\sqrt{n_i}}{CV_{i0}\sqrt{1+q^2 l}}, \ldots, Z_{ih_i}^q > d_a^q - \frac{(\Theta - q)\sqrt{n_i}}{CV_{i0}\sqrt{1+q^2 l}} \ \forall i \right)$$

where the Z_{ij}^q's are jointly distributed as a standardized $h = \sum_{i=1}^{r} h_i$ -variate normal random variable with correlation matrix $\mathbf{R}_h^q$, $d_a = x_{rc,\mathbf{0},\mathbf{R}_{rc}^q;1-a}$ is the $1-a$ percentage point of this distribution and $CV_{i0} = s / m_{i0}$ is the coefficient of variation of the control treatment in stratum *i*.

There is no unique solution because there are many *h*-vectors $\mathbf{b} = (b_1, \ldots, b_1, \ \ldots \ , b_r, \ldots, b_r)'$, with $b_i = d_a^q - \frac{(\Theta - q)\sqrt{n_i}}{CV_{i0}\sqrt{1+q^2 l}}$ (*i* = 1, …, *r*) such that $\Phi_h(\mathbf{b}, \infty, \mathbf{0}, \mathbf{R}_h^q) \ge 1 - b$. But assuming a certain relation between the b_i's, for example assuming that all sample sizes are equal across all strata, i.e. $n_i = n$, the solution is unique and can easily be found iteratively, because all parameters are known except n. The sample size n is the smallest n such that $\Phi_h(\mathbf{b}(n), \infty, \mathbf{0}, \mathbf{R}_h^q) \ge 1 - b$.

Although the assumption of equal sample sizes seems to be reasonable at first instance, it doesn't treat all alternative hypotheses equally. Assume that the mean control level in stratum i_1 is lower than the mean control level in stratum i_2, i.e. $m_{i_1} < m_{i_1}$. Then the coefficient of variation of the control treatment in stratum i_1 is larger than in stratum i_2, which results in a higher value of b_{i_1} compared to b_{i_2}. But then if follows from formula (6.17) that the hypotheses in stratum i_1 are less likely to be rejected than in stratum i_2.

Under the assumption that $\frac{\sqrt{n_i}}{CV_{i0}}$ should be constant, explicit formulas for the sample size can be derived. Then the sample sizes are the smallest integers n_i that fulfill the inequality:

$$n_i \geq \left(1+q^2 l\right)\left(x_{rc,\mathbf{0},\mathbf{R}^q_{rc};1-a} + x_{h,\mathbf{0},\mathbf{R}^q_h;1-b}\right)^2 CV_{i0}{}^2 / \left(\Theta - q\right)^2 \quad (i = 1, .., r) \qquad (6.18)$$

if a minimal all-pairs Θ power of $1-b$ is required and the smallest integers n_i that fulfill the inequality:

$$n_i \geq \left(1+q^2 l\right)\left(x_{rc,\mathbf{0},\mathbf{R}^q_{rc};1-a} - x_{h,\mathbf{0},\mathbf{R}^q_h;b}\right)^2 CV_{i0}{}^2 / \left(\Theta - q\right)^2 \quad (i = 1, .., r) \qquad (6.19)$$

if a minimal any-pair Θ power of $1-b$ is required.

Step-down procedure

The step-down procedure based on the closed testing procedure as discussed in Section 3.5 for the testing situation formulated in terms of differences can also be applied for the testing problem phrased in terms of ratios.

Consider the finite family of *rc* individual sub-hypotheses as defined in formula (6.8) with corresponding test statistics T^q_{ij} and observed values t^q_{ij} (i = 1, …, r and j = 1, …, c).

Then the step-down procedure to test the hypotheses is as follows:

- Order all observed test statistics t^q_{ij}'s from smallest to largest, say $t^q_{(1)} \leq t^q_{(2)} \leq \ldots \leq t^q_{(rc)}$. Let $H^q_{0(1)}$, $H^q_{0(2)}$, …, $H^q_{0(rc)}$ be the corresponding null-hypotheses and let $E_{(m)}$ be the subset of indices *ij*'s corresponding to the *m* smallest t^q_{ij}'s (m = 1, …, rc). Thus $E_{(rc)}$ is the set of all indices and $E_{(1)}$ refers to the indices corresponding to $t^q_{(1)}$.

Denote with $\mathbf{R}_{(m)}$ the sub-matrix of the correlation matrix $\mathbf{R}$ restricted to $E_{(m)}$ (m = 1, …, rc).

- Start with testing $H^q_{0(rc)}$ and reject $H^q_{0(rc)}$ if $t^q_{(rc)} > t_{rc,n,\mathbf{R}_{(rc)};1-a}$; otherwise retain all sub-hypotheses without further tests.
- The general step m is, reject $H^q_{0(m)}$ if $H^q_{0(rc)}$, …, $H^q_{0(m+1)}$ are rejected and $t^q_{(m)} > t_{m,n,\mathbf{R}_{(m)};1-a}$. If $H^q_{0(m)}$ is not rejected, then also retain $H^q_{0(m-1)}$, …, $H^q_{0(1)}$ without any further testing (m = 1, …, rc).

An adjusted p-value for $H^q_{0(m)}$ can be computed as

$$p^q_{(m)} = \max\left\{\tilde{p}^q_{(m)}, \tilde{p}^q_{(m+1)}, \dots, \tilde{p}^q_{(rc)}\right\} \tag{6.20}$$

where

$$\tilde{p}^q_{(m)} = P\left(\text{at least one } T^q_{ij} > t^q_{(m)},\ \{ij\} \in E_{(m)}\right) = 1 - P\left(T^q_{ij} \le t^q_{(m)},\ \{ij\} \in E_{(m)}\right). \tag{6.21}$$

6.2 Two-sided testing situation

Chapter 4 discusses the required adaptations to transform the results from the one-sided testing problem to the two-sided testing problem in case the problem is formulated in terms of differences. Similar adaptations can be applied to the results of Section 6.1 to derive results for the two-sided testing problem in terms of ratios. These adaptations are rather straightforward.

For example, the global null hypothesis

$$H^q_0 : \frac{m_{ij}}{m_{i0}} = q \qquad (i = 1, \dots, r\ \ j = 1, \dots, c)$$

is tested against the two-sided alternative hypothesis (6.22)

$$H^q_1 : \exists ij : \frac{m_{ij}}{m_{i0}} \neq q \qquad (i = 1, \dots, r\ \ j = 1, \dots, c)$$

where $q > 0$, by using the test statistic

$$T_2^q = \max_{1\le i\le r;1\le j\le c}\{T_{2ij}^q\} \text{ where } T_{2ij}^q = \frac{\left|\bar{X}_{ij} - q\,\bar{X}_{i0}\right|}{s\sqrt{n_{ij}^{-1} + q^{2} n_{i0}^{-1}}} \quad (i = 1, \ldots, r\ \ j = 1, \ldots, c). \qquad (6.23)$$

See also Sasabuchi (1988).

Percentage points can be calculated by using the following formula for the probability $P\left(T_2^q \le t\right)$:

$$P\left(T_2^q \le t\right) = \int_0^{\infty}\prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\prod_{j=1}^{c}\left[\Phi\left(\frac{b_{ij}^q y + t\sqrt{u}}{\sqrt{1-\left(b_{ij}^q\right)^2}}\right) - \Phi\left(\frac{b_{ij}^q y - t\sqrt{u}}{\sqrt{1-\left(b_{ij}^q\right)^2}}\right)\right] j\,(y)dy\right] h(u)du \qquad (6.24)$$

where $b_{ij}^q = \frac{q}{\sqrt{\frac{n_{i0}}{n_{ij}} + q^{2}}}$, *h(u)* is the density function of a c_n^2 / n distributed variable and $\Phi(y)$ and $j\,(y)$ are the standard cumulative distribution function and probability density function respectively.

Adjusted p-values $\tilde{p}_{2ij}^q$ corresponding to the sub-hypothesis $H_{0ij}^q : \frac{m_{ij}}{m_{i0}} = q$ can be computed as:

$$\tilde{p}_{2ij}^q = \min\left\{a \;\middle|\; H_{0ij}^q \text{ is rejected at FWE} = a\right\} = P_{H_0}\left(T_2^q > t_{2ij}^q\right) = \qquad (6.25)$$

$$= 1 - T_{rc}\left(\mathbf{-t}_{2ij}^q, \mathbf{t}_{2ij}^q; n, \mathbf{R}^q\right)$$

where t_{2ij}^q (> 0) is the observed value of the test statistic T_{2ij}^q (*i* = 1, …, *r* and *j* = 1, …, *c*).

The other results can also be derived rather easily but the adaptations are not discussed any further in this section.

The testing problem of showing global equivalence as discussed in Section 5.2 in terms of differences can also be formulated for ratios. (See also Hauschke (1999))

The global null hypothesis to be tested is

$$H_0^{q_{12}} : \exists ij : \frac{m_{ij}}{m_{i0}} \le q_1 \text{ or } \frac{m_{ij}}{m_{i0}} \ge q_2 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

versus (6.26)

$$H_1^{q_{12}} : q_1 < \frac{m_{ij}}{m_{i0}} < q_2 \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

where (q_1, q_2), $q_1 < 1 < q_2$ describes the area of irrelevant proportions.

This situation can be handled by considering the following:

- the null sub-hypothesis $H_{0ij}^{q_1} : \frac{m_{ij}}{m_{i0}} \le q_1$ is rejected at level a in favor of the alternative sub-hypothesis $H_{1ij}^{q_1} \frac{m_{ij}}{m_{i0}} > q_1$ if $T_{ij}^{q_1} = \frac{\bar{X}_{ij} - q_1 \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + q_1^{\,2} n_{i0}^{-1}}} \ge t_{1-a,n}$ and
- the null sub-hypothesis $H_{0ij}^{q_2} : \frac{m_{ij}}{m_{i0}} \ge q_2$ is rejected at level a in favor of the alternative sub-hypothesis $H_{1ij}^{q_2} \frac{m_{ij}}{m_{i0}} < q_2$ if $T_{ij}^{q_2} = \frac{\bar{X}_{ij} - q_2 \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + q_2^{\,2} n_{i0}^{-1}}} \le -t_{1-a,n}$ and
- the Intersection-Union principle can be applied by writing the global hypotheses as $H_0^{q_{12}} : \bigcup_{i=1}^{r} \bigcup_{j=1}^{c} \bigcup_{k=1}^{2} H_{0ij}^{q_k}$ and $H_1^{q_{12}} : \bigcap_{i=1}^{r} \bigcap_{j=1}^{c} \bigcap_{k=1}^{2} H_{1ij}^{q_k}$ respectively.

Computations of the example

The same example as used in Chapter 3 and 4 will be used to illustrate the many-to-one comparisons in a stratified two-way layout in case the problem is phrased in terms of ratios rather than in differences.

In case of testing equality, the two-sided testing situation is as follows:

$$H_0: \frac{m_{11}}{m_{10}} = \frac{m_{12}}{m_{10}} = \frac{m_{21}}{m_{20}} = \frac{m_{22}}{m_{20}} = 1$$

versus

$$H_1: \frac{m_{11}}{m_{10}} \neq 1 \text{ or } \frac{m_{12}}{m_{10}} \neq 1 \text{ or } \frac{m_{21}}{m_{20}} \neq 1 \text{ or } \frac{m_{22}}{m_{20}} \neq 1$$

where *i* = 1 and *i* = 2 represent the males and females and *j* = 1 and *j* = 2 represent the low and high dose respectively.

The results of the analysis are presented in the following table.

Table 6.1 Analysis of example expressed as ratio testing problem

Stratum	Contrast	Estimate	Adjusted p-value $\tilde{p}^q_{2ij}$
M	Plac-Low	1.084	0.140
	Plac-High	1.210	<0.001
F	Plac-Low	1.040	0.516
	Plac-High	1.086	0.029

The program code can be found in program Ch6.sas of Appendix 3, which also computes the two sided upper percentage point $d_{2,0.05} = d_2(0.05,2,2,37,\{b^q_{ij}\}) = |t|_{4,37,\mathbf{R}^q;0.95} = 2.601$.

This table shows for example that the best improvement is seen for the high dosage in the males. The estimated improvement is 21% and with a p-value of p<001, the effect is highly significant.

Notice that because of testing $q = 1$, the adjusted p-values are identical to the adjusted p-values computed for the two-sided testing situation in terms of differences as shown in Chapter 4.

The estimated effect of the low dose in males (8.4%) is almost equal to the estimated effect of the high dose in females (8.6%). Hence, at first instance, it seems strange that the adjusted p-values are quite different, while the sample sizes are similar. The explanation is that the level of the group means for the females is higher than for the males, which results in higher observed values for the test statistics in the females.

7 Nonparametric Procedure

So far the data are assumed to be normally distributed. In practice, there are situations where this assumption is suspect. Then the use of a distribution-free or so-called nonparametric approach might be appropriate. For example the use of a general nonparametric approach based on pairwise rankings. Many-to-one comparisons based on pairwise rankings was already published by Steel (1959), although this required a continuous distribution. Joint ranking procedures are not discussed because they do not control the Type I FWE as shown by Oude Voshaar (1980) and Fligner (1984). See also the discussion in Chapter 9 of Hochberg and Tamhane (1987) and Chapter 3 of Hsu (1996).
Akritas and Brunner (1997) derived an asymptotic approach to rank tests for continuous as well as tied data and Munzel and Hothorn (2001) applied their approach to describe as a special case the many-to-one comparisons for the one-way layout based on a pairwise ranking procedure. A nice overview of rank procedures in factorial designs can be found in Brunner and Puri (1996).

This chapter describes a single-step asymptotic test procedure based on pairwise rankings to perform many-to-one comparisons for a stratified two-way layout. This procedure is an extension of the method proposed by Munzel and Hothorn (2001). Although it will not be discussed any further, a step-down asymptotic test procedure can be derived in analogue to the derivation of the step-down test procedure assuming normal distributed data as described in Section 3.5 and Section 4.2.
Section 7.1 describes the general setting of the testing problem. Characteristics of the asymptotic distribution of the relative pairwise effects are presented in Section 7.2 and the test procedures are derived in Section 7.3.

7.1 Distribution functions, relative effects and hypotheses

This section introduces the hypotheses to be tested in terms of arbitrary distribution functions instead of normal distribution functions only.

Let the random variable X_{ijk} denotes the *k*-th observation on treatment *j* (where again *j* = 0 denotes the control treatment) in stratum *i*, and let X_{ijk} be independently distributed according to an arbitrary distribution function $F_{ij}(x)$, i.e.

$$X_{ijk} \sim F_{ij}(x) = 0.5\left[P(X_{ij} \le x) + P(X_{ij} < x)\right] \qquad (7.1)$$

$i = 1, \ldots, r$, $j = 0, 1, \ldots, c$ and $k = 1, \ldots, n_{ij}$

This definition of the distribution function is the so called normalized version (Ruymgaart (1980)) that includes both continuous as well as discontinuous data as long as the scale level of the observations is at least ordinal. It only excludes the trivial case of a one-point distribution.

The relative (pairwise) effect of active treatment *j* with respect to the control treatment within each of the *r* groups can be expressed as:

$$p_{ij} = \int F_{ij} dF_{i0} = P(X_{ijk} < X_{i0k}) + 0.5P(X_{ijk} = X_{i0k}) \ \ (i = 1, \ldots, r\ ,\ j = 1, \ldots, c) \qquad (7.2)$$

Notice that this is a generalization of the effect of the Wilcoxon-Mann-Whitney (1947) rank test in case of ties.

If the space of possible distribution functions is reduced to a certain one-dimensional subspace, e.g. in shift models where $F(x) = F(x - m)$ or by assuming non-crossing distribution functions, the relative effect p_{ij} defines a stochastic order by $F_{ij} < F_{i0}$, $F_{ij} = F_{i0}$ or $F_{ij} > F_{i0}$, according to $p_{ij} <$ 0.5, $p_{ij} =$ 0.5 or $p_{ij} >$0.5.

In the general situation that only the distribution functions are specified, such an ordering does not exist and no natural parameters are available that measure differences between the treatment groups. Therefore the hypothesis of no treatment effect may be formulated either in terms of the distribution functions $F_{ij}(x)$ or in terms of the relative treatment effects p_{ij}.

Consider the family of *rc* sub-hypotheses

$$H_{0ij}^{F} : F_{ij} = F_{i0}$$

against the two-sided alternatives (7.3)

$$H_{1ij}^{F} : F_{ij} \ne F_{i0}$$

with the corresponding global null hypothesis

$$H_0^F = \bigcap_{i=1}^{r}\bigcap_{j=1}^{c} H_{0ij}^F \;. \qquad (7.4)$$

Each of these hypotheses is expressed in terms of distribution functions and tests whether the distribution function of the active treatment is equal to the distribution function of the control treatment. It can be considered as a test of homogeneity.

The family of sub-hypotheses expressed in terms of the relative treatment effects

$$H_{0ij}^p : p_{ij} = 0.5$$

against the two-sided alternatives (7.5)

$$H_{1ij}^p : p_{ij} \neq 0.5$$

and corresponding global null hypothesis

$$H_0^p = \bigcap_{i=1}^{r}\bigcap_{j=1}^{c} H_{0ij}^p \qquad (7.6)$$

tests whether the active treatment effects are equal to the effect of the control treatment for each of the *r* strata. It tests whether the relative treatment effects are equal to 0.5.

Notice that $H_{0ij}^F : F_{ij} = F_{i0}$ implies $H_{0ij}^p : p_{ij} = 0.5$, whereas the other way around is not true in general. However, if it is assumed, e.g. that the possible distribution functions are non-crossing, i.e. $F_{ij}(x) \leq F_{i0}(x)$ or $F_{ij}(x) \geq F_{i0}(x)$ for all x, the hypotheses H_{0ij}^F and H_{0ij}^p coincide. Moreover, in a shift model are both hypotheses H_{0ij}^F and H_{0ij}^p equivalent to the hypothesis that tests equality of the location parameters $H_{0ij}^m : m_{ij} = m_{i0}$.

The problem of testing H_{0ij}^p is sometimes called the multiple nonparametric Behrens-Fisher testing problem because H_{0ij}^p is equivalent to testing the location parameters $H_{0ij}^m : m_{ij} = m_{i0}$ in case the expectation of two heterogeneous normal distributions $N(m_{ij}, s_{ij}^{2})$ and $N(m_{i0}, s_{i0}^{2})$ are compared, whereas H_{0ij}^{F} requires equality of the variances.

The testing problems are formulated as two-sided testing problems. Notice that the one-sided testing problems do not have reasonable interpretations in general. Therefore the one-sided testing situation is not described in this chapter.
However, if the space of possible distribution functions is reduced to a certain one-dimensional subspace it might be useful to test against one-sided alternatives. In that case, the procedures that will be derived later in this chapter to deal with the two-sided testing situation can be easily adapted to suit the one-sided testing situation.

7.2 Estimators of relative effects and asymptotic covariance matrix

The asymptotic distribution of a consistent estimate of the relative pairwise treatment effects as well as an estimator of the asymptotic covariance matrix are derived in this section in order to derive test statistics for H_0^F and H_0^p in Section 7.3.

The distribution functions $F_{ij}(x)$ are unknown but can be estimated by their empirical counterparts $\hat{F}_{ij}(x)$. The empirical distribution function of $F(x)$ is denoted by

$$\hat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} c\left(x - X_k\right) \tag{7.7}$$

where $c(u)$ denotes the counting function and $c(u) = \begin{cases} 0 & \text{if } u < 0 \\ 0.5 & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases}$

Therefore the relative pairwise effects p_{ij} can be estimated by

$$\hat{p}_{ij} = \int \hat{F}_{ij} d\hat{F}_{i0} \qquad (i = 1, \dots, r \quad j = 1, \dots, c). \tag{7.8}$$

It can be shown, e.g. by using the ideas in Brunner, Puri and Sun (1995) that $\hat{p}_{ij}$ is an unbiased and consistent estimate of p_{ij} .

The estimate $\hat{p}_{ij}$ can also be expressed in terms of mid-ranks

$$\hat{p}_{ij} = \int \hat{F}_{ij} d\hat{F}_{i0} = \frac{1}{n_{ij}} \left(\bar{R}_{i0.}^{(j0)} - \frac{n_{i0}+1}{2} \right) \tag{7.9}$$

where $\bar{R}_{i0.}^{(j0)} = \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} R_{i0k}^{(j0)}$ and

$R_{i0k}^{(j0)}$ is the mid-rank of the random variable X_{i0k} among all observations within the *j*-th active treatment group and control group in the *i*-th stratum, i.e. $X_{i01},\ldots,X_{i0n_0}$ and $X_{ij1},\ldots,X_{ijn_j}$, which is defined as $R_{i0k}^{(j0)} = 0.5 + \sum_{l=1}^{n_{i0}} c\left(X_{i0k} - X_{i0l}\right) + \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right)$. In order to get the position numbers of the ordered observations in case of no ties, 0.5 has to be added since $c(0) =$ 0.5.

This can easily be seen by noticing that

$$\hat{p}_{ij} = \int \hat{F}_{ij} d\hat{F}_{i0} = \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} \hat{F}_{ij}(X_{i0k}) = \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} \frac{1}{n_{ij}} \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right) = \frac{1}{n_{ij}} \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right) \quad \text{and}$$

$$\bar{R}_{i0.}^{(j0)} = \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} R_{i0k}^{(j0)} = \frac{1}{n_{i0}} \left\{ \frac{n_{i0}}{2} + \sum_{k=1}^{n_{i0}} \sum_{l=1}^{n_{i0}} c\left(X_{i0k} - X_{i0l}\right) + \sum_{k=1}^{n_{i0}} \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right) \right\} =$$

$$= \frac{1}{n_{i0}} \left\{ \frac{n_{i0}}{2} + \frac{n_{i0} n_{i0}}{2} + \sum_{k=1}^{n_{i0}} \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right) \right\} = \frac{n_{i0}+1}{2} + \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} \sum_{l=1}^{n_{ij}} c\left(X_{i0k} - X_{ijl}\right).$$

In order to obtain asymptotic distribution results it is useful to consider the following asymptotic result that follows from general nonparametric theory (see e.g. Brunner and Puri (1996)).

Under the assumptions that

$$n_{ij} \to \infty \text{ and } 0 < l < \frac{n_{ij}}{\sum_{ij} n_{ij}} < 1 - l < 1 \qquad \text{for all } i = 1, \ldots, r \quad j = 1, \ldots, c \tag{7.10}$$

it follows that

$$\left\| \sqrt{\mathbf{N}}(\hat{\mathbf{p}} - \mathbf{p}) - \sqrt{\mathbf{N}}\mathbf{B} \right\|_2 \to 0 \tag{7.11}$$

where

$\mathbf{N} = diag\left(N_{11},...,N_{rc}\right)$ is a diagonal matrix with the pooled sample sizes $N_{ij} = n_{ij} + n_{i0}$ on the diagonal, $\mathbf{B} = \left(B_{11},...,B_{rc}\right)'$ is a vector with elements $B_{ij} = \int F_{ij} d\hat{F}_{i0} - \int F_{i0} d\hat{F}_{ij} + 1 - 2p_{ij}$ (i = 1, …, r and j = 1, …, c) and $\left\| X \right\|_2 = \sqrt{E(X^2)}$ denotes the L_2-norm.

Under the assumptions (7.10) and assuming that

$$s^2_{i0,j} = Var(Y_{i0,j1}) > 0 \text{ and } s^2_{ij,0} = Var(Y_{ij,01}) > 0 \tag{7.12}$$

the asymptotic distribution of $\sqrt{\mathbf{N}}\left(\hat{\mathbf{p}} - \mathbf{p}\right)$ can be obtained:

$$\sqrt{\mathbf{N}}\left(\hat{\mathbf{p}} - \mathbf{p}\right) \to N(\mathbf{0}, \mathbf{V}) \tag{7.13}$$

i.e. $\sqrt{\mathbf{N}}\left(\hat{\mathbf{p}} - \mathbf{p}\right)$ has asymptotically an *rc*-variate normal distribution with expectation **0** and covariance matrix $\mathbf{V} = Cov(\sqrt{\mathbf{N}}\mathbf{B})$.

No complete proof is provided but the outline is as follows.
According the asymptotic result above (7.11) it is sufficient to look at the asymptotic distribution of $\sqrt{\mathbf{N}}\mathbf{B}$. Notice that B_{ij} is the sum of independent, uniformly bounded (≤ 1) and unobservable random variables $Y_{i0,jk} = F_{i0}(X_{ijk})$ and $Y_{ij,0k} = F_{ij}(X_{i0k})$:

$$B_{ij} = \int F_{ij} d\hat{F}_{i0} - \int F_{i0} d\hat{F}_{ij} + 1 - 2p_{ij} = \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} F_{ij}(X_{i0k}) - \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} F_{i0}(X_{ijk}) + 1 - 2p_{ij} =$$

$$= \frac{1}{n_{i0}} \sum_{k=1}^{n_{i0}} Y_{ij,0k} - \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{i0,jk} + 1 - 2p_{ij} .$$

Show that the Lindeberg condition is fulfilled. For more details see e.g. Munzel and Hothorn (2001).

Introduce the following notation to describe the elements of the covariance matrix $\mathbf{V}$:

$$s_{ij_1j_2} = Cov(Y_{ij_1,01}, Y_{ij_2,01})\,, \tag{7.14}$$

$$\bar{Y}_{ij,0.} = \frac{1}{n_{i0}}\sum_{k=1}^{n_{i0}} Y_{ij,0k} \text{ and } \bar{Y}_{i0,j.} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} Y_{i0,jk}\,.$$

Then the elements $v_{(i_1j_1),(i_2j_2)}$ (i_1, i_2 = 1, …, r and j_1, j_2 = 1, …, c) can be expressed as

$$v_{(i_1j_1),(i_2j_2)} = \sqrt{N_{i_1j_1}N_{i_2j_2}}\,C\text{ov}\left(\bar{Y}_{i_1j_1,0.} - \bar{Y}_{i_10,j_1.}, \bar{Y}_{i_2j_2,0.} - \bar{Y}_{i_20,j_2.}\right) \tag{7.15}$$

such that

$$\begin{aligned}
v_{(i_1j_1),(i_2j_2)} &= 0 && \text{if } i_1 \neq i_2 \\
v_{(i_1j_1),(i_2j_2)} &= \sqrt{N_{ij_1}N_{ij_2}}\,\frac{1}{n_{i0}}s_{ij_1j_2} && \text{if } i_1 = i_2 = i \text{ and } j_1 \neq j_2 \\
v_{(i_1j_1),(i_2j_2)} &= N_{ij}\left(\frac{1}{n_{i0}}s^2_{ij,0} + \frac{1}{n_{ij}}s^2_{i0,j}\right) && \text{if } i_1 = i_2 = i \text{ and } j_1 = j_2 = j
\end{aligned} \tag{7.16}$$

and thus the covariance matrix $\mathbf{V}$ has a block diagonal structure.

A consistent estimator $\hat{\mathbf{V}}$ for the covariance matrix $\mathbf{V}$ can be obtained under the assumptions (7.10) and (7.12).

The elements $\hat{v}_{(i_1j_1),(i_2j_2)}$ (i_1, i_2 = 1, …, r and j_1, j_2 = 1, …, c) of the covariance matrix $\hat{\mathbf{V}}$ are given by

$$\begin{aligned}
\hat{v}_{(i_1j_1),(i_2j_2)} &= 0 && \text{if } i_1 \neq i_2 \\
\hat{v}_{(i_1j_1),(i_2j_2)} &= \sqrt{N_{ij_1}N_{ij_2}}\,\frac{1}{n_{i0}}\hat{s}_{ij_1j_2} && \text{if } i_1 = i_2 = i \text{ and } j_1 \neq j_2 \\
v_{(i_1j_1),(i_2j_2)} &= N_{ij}\left(\frac{1}{n_{i0}}\hat{s}^2_{ij,0} + \frac{1}{n_{ij}}\hat{s}^2_{i0,j}\right) && \text{if } i_1 = i_2 = i \text{ and } j_1 = j_2 = j
\end{aligned} \tag{7.17}$$

where

$$\hat{s}^2_{ij,0} = \frac{1}{n_{ij}^2(n_{i0}-1)}\sum_{k=1}^{n_{i0}}\left(R^{(j0)}_{i0k} - R^{(0)}_{i0k} - \bar{R}^{(j0)}_{i0.} + \frac{n_{i0}+1}{2}\right)^2,$$

$$\hat{s}^2_{i0,j} = \frac{1}{n_{i0}^2(n_{ij}-1)} \sum_{k=1}^{n_{ij}} \left(R_{ijk}^{(j0)} - R_{ijk}^{(j)} - \bar{R}_{ij.}^{(j0)} + \frac{n_{ij}+1}{2} \right)^2 \text{ and } \tag{7.18}$$

$$\hat{s}_{ij_1 j_2} = \frac{1}{(n_{i0}-1)} \sum_{k=1}^{n_{i0}} D_{ij_1,0k} D_{ij_2,0k} \text{ with } D_{ij,0k} = \frac{1}{n_{ij}} \left(R_{i0k}^{(j0)} - R_{i0k}^{(0)} - \bar{R}_{i0.}^{(j0)} + \frac{n_{i0}+1}{2} \right)$$

No complete proof is provided but the idea how to proof e.g. that $\hat{s}^2_{ij,0}$ is a consistent estimator for $s^2_{ij,0}$ is as follows.

$s^2_{ij,0} = Var(Y_{ij,01})$ is the variance of the unobservable random variables $Y_{ij,01}, \ldots, Y_{ij,0n_{i0}}$ which are independently and identically distributed. A consistent estimator for $s^2_{ij,0}$ would be $\frac{1}{(n_{i0}-1)} \sum_{k=1}^{n_{i0}} \left(Y_{ij,0k} - \bar{Y}_{ij,0.} \right)^2$ in case the variables $Y_{ij,0k}$ were observable. Now replace the unobservable variables $Y_{ij,0k} = F_{ij}(X_{iok})$ by their empirical quantities $\hat{Y}_{ij,0k} = \hat{F}_{ij}(X_{i0k}) = \frac{1}{n_{ij}} \left(R_{i0k}^{(j0)} - R_{i0k}^{(0)} \right)$. More details can be found in e.g. Munzel and Hothorn (2001)

7.3 Test procedures

Given the asymptotic results of the previous section, test statistics can be defined to test the global null hypothesis H_0^p expressed in terms of the relative treatment effects or to test the global null hypothesis H_0^F expressed in terms of distribution functions.

First, consider the global null hypothesis expressed in terms of the relative treatment effects

$H_0^p : p_{ij} = 0.5$ (*i* = 1, …, *r* *j* = 1, …, *c*)

versus

$H_1^p : \exists ij : p_{ij} \neq 0.5$ (*i* = 1, …, *r* *j* = 1, …, *c*)

as already defined in Section 7.1.

Define the test statistics

$$T_{ij}^{p} = \frac{\sqrt{N_{ij}}\left(\hat{p}_{ij} - \frac{1}{2}\right)}{\sqrt{\hat{v}_{(ij),(ij)}}} = \frac{\hat{p}_{ij} - \frac{1}{2}}{\sqrt{\frac{\hat{s}_{ij,0}^{2}}{n_{i0}} + \frac{\hat{s}_{i0,j}^{2}}{n_{ij}}}} \qquad (i = 1, \ldots, r \quad j = 1, \ldots, c) \tag{7.19}$$

and

$$T^{p} = \max_{1 \le i \le r; 1 \le j \le c} \left\{ \left| T_{ij}^{p} \right| \right\}. \tag{7.20}$$

Notice that for example assuming that observations in the control group have lower values than in the active treatment group j within stratum i, $\hat{p}_{ij}$ is smaller than 0.5, which results in a negative value of T_{ij}^{p}.

The results of the previous section show that under the global null hypothesis H_0^p, the joint distribution of the T_{ij}^{p}'s is asymptotically an *rc*-variate normal distribution with expectation **0** and correlation matrix $\mathbf{R}^p = \left\{ r_{i(j_1,j_2)}^{p} \right\}$.

The correlation matrix $\mathbf{R}^p$ has the following block diagonal structure

$$\mathbf{R}^p = \begin{pmatrix} \mathbf{R}_1^p & .. & 0 \\ .. & .. & .. \\ 0 & .. & \mathbf{R}_r^p \end{pmatrix} \text{ with } \mathbf{R}_i^p = \begin{pmatrix} 1 & r_{i(1,2)}^{p} & .. & r_{i(1,c)}^{p} \\ r_{i(2,1)}^{p} & 1 & .. & .. \\ .. & .. & 1 & r_{i(c-1,c)}^{p} \\ r_{i(c,1)}^{p} & .. & r_{i(c,c-1)}^{p} & 1 \end{pmatrix} \tag{7.21}$$

and correlation coefficients

$$r_{i(j_1,j_2)}^{p} = \frac{v_{(ij_1),(ij_2)}}{\sqrt{v_{(ij_1),(ij_1)} v_{(ij_2),(ij_2)}}} = \frac{s_{ij_1 j_2}}{\sqrt{s_{ij_1,0}^{2} + \frac{n_{i0}}{n_{ij_1}} s_{i0,j_1}^{2}} \sqrt{s_{ij_2,0}^{2} + \frac{n_{i0}}{n_{ij_2}} s_{i0,j_2}^{2}}} \tag{7.22}$$

$(i = 1, \ldots, r \quad j_1, j_2 = 1, \ldots, c)$

which can unfortunately not be written as a product.

The test procedure that rejects the global null hypothesis H_0^p in favor of the two-sided alternative hypothesis H_1^p if $T^p > d_{2a}^p$, asymptotically controls the FWE if d_{2a}^p is chosen such that $P_{H_0^p}\left(T^p > d_{2a}^p\right) = a$. This can be shown similar to the statements given in Section 2.1 for the original Dunnett procedure.

The percentage point $d_{2a}^p = d_2^p(a, r, c, \left\{ r_{i(j_1, j_2)}^p \right\})$ is the two-sided 1 - a percentage point of the *rc*-variate normal distribution with correlation matrix $\mathbf{R}^p$, which is denoted as $|x|_{rc,\mathbf{0},\mathbf{R}^p;1-a}$.

This is analogue to the two-sided test situation described in Section 4.2 that makes use of the multivariate t-distribution. As described earlier, nowadays computer algorithms are available to compute these percentage points. See e.g. Genz (1992) and Genz and Bretz (1999) and SAS/IML code is available at the homepage of Bretz (the website with URL http://www.bioinf.uni-hannover.de/~bretz/).

Notice that the multivariate t-distribution converges to the multivariate normal distribution for increasing degrees of freedom (see Appendix 1). Brunner and Munzel (2000) demonstrated that the accuracy of the normal approximation discussed in Section 7.2 could be improved for small sample sizes by using a multivariate t-distribution, with a Satterthwaite approximation to calculate the degrees of freedom. Munzel and Hothorn (2001) recommend to use the conservative approximation for the degrees of freedom $n = \max\left\{1, \min\left\{\hat{f}_{ij}\right\}\right\}$, where

$$\hat{f}_{ij} = \frac{\left(n_{ij}\hat{s}_{ij,0}^2 + n_{i0}\hat{s}_{i0,j}^2\right)^2}{\left(n_{i0}-1\right)^{-1}\left(n_{ij}\hat{s}_{ij,0}^2\right)^2 + \left(n_{ij}-1\right)^{-1}\left(n_{i0}\hat{s}_{i0,j}^2\right)^2}.$$

In analogy to the two-sided test situation assuming normal distributed data as described in Section 4.2, adjusted p-values $\tilde{p}_{ij}^p$'s can be computed as:

$$\tilde{p}_{ij}^p = \min\left\{a \;\middle|\; H_{0ij}^p \text{ is rejected at FWE} = a\right\} = P_{H_0}\left(T^p > \left|t_{ij}^p\right|\right) = \\ = 1 - \Phi_{rc}\left(-\left|t_{ij}^p\right|, \left|t_{ij}^p\right|; \mathbf{0}, \mathbf{R}^p\right) \tag{7.23}$$

where $\left|t_{ij}^p\right|$ is the observed value of the test statistic $\left|T_{ij}^p\right|$ (*i* = 1, …, *r* and *j* = 1, …, *c*).

Two-sided 100(1 - a)% simultaneous confidence intervals for p_{ij} can also be calculated using the expression:

$$\hat{p}_{ij} \pm |x|_{rc,\mathbf{0},\mathbf{R}^p;1-a} \sqrt{\frac{\hat{s}^{2}_{ij,0}}{n_{i0}} + \frac{\hat{s}^{2}_{i0,j}}{n_{ij}}} \qquad (i = 1, \ldots, r \text{ and } j = 1, \ldots, c). \qquad (7.24)$$

where $|x|_{rc,\mathbf{0},\mathbf{R}^p;1-a}$ is the two-sided 1 - a percentage point of the *rc*-variate normal distribution with correlation matrix $\mathbf{R}^p$ and both $\hat{s}^{2}_{ij,0}$ and $\hat{s}^{2}_{i0,j}$ are given in formula (7.18).

Let us now consider the other testing problem phrased in terms of distribution functions, i.e. consider

$$H_0^F : F_{ij} = F_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c)$$

versus

$$H_{1ij}^F : \exists ij : F_{ij} \neq F_{i0} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c).$$

Notice that the variances $s^{2}_{i0,j} = Var(F_{i0}(X_{ij1}))$ and $s^{2}_{ij,0} = Var(F_{ij}(X_{i01}))$ are equal under the null hypothesis H_0^F.

Hence the variance $v_{(ij),(ij)}$ reduces to $v_{(ij),(ij)} = N_{ij}\left(\frac{1}{n_{i0}} s^{2}_{ij0} + \frac{1}{n_{ij}} s^{2}_{ij0}\right) = \frac{N_{ij}^2}{n_{i0}n_{ij}} s^{2}_{ij0}$, where $s^{2}_{ij0} = s^{2}_{ij,0} = s^{2}_{i0,j} = Var(F_{ij0}(X_{ij1}))$ and $F_{ij0} = F_{ij} = F_{i0}$.

Then a consistent estimator for s^{2}_{ij0} is given by

$$\hat{s}^{2}_{ij0} = \frac{1}{N_{ij}^2(N_{ij}-1)}\left[\sum_{k=1}^{n_{i0}}\left(R^{(j0)}_{i0k} - \frac{N_{ij}+1}{2}\right)^2 + \sum_{k=1}^{n_{ij}}\left(R^{(j0)}_{ijk} - \frac{N_{ij}+1}{2}\right)^2\right] \qquad (7.25)$$

assuming that the assumptions (7.10) and (7.12) are fulfilled.

(The proof is similar to the proof given in Section 7.2. Notice that in this situation $\frac{1}{(N_{ij}-1)}\left[\sum_{k=1}^{n_{i0}}\left(F_{ij0}(X_{iok})-F_{ij0}(\bar{X}_{i..})\right)^2+\sum_{k=1}^{n_{ij}}\left(F_{ij0}(X_{ijk})-F_{ij0}(\bar{X}_{i..})\right)^2\right]$ would be a consistent estimator for s^2_{ij0} in case the variables $F_{ij0}(X_{ijk})$ and $F_{ij0}(X_{i0k})$ were observable. Again replace the unobservable variables by their empirical quantities.)

The test statistics T^p_{ij}'s defined in (7.19) for the testing problem in terms of relative effects reduces therefore into

$$T^F_{ij} = \frac{\sqrt{N_{ij}}\left(\hat{p}_{ij}-\frac{1}{2}\right)}{\sqrt{\hat{v}_{(ij),(ij)}}} = \frac{\sqrt{n_{i0}n_{ij}}\left(\hat{p}_{ij}-\frac{1}{2}\right)}{\sqrt{N_{ij}\hat{s}^2_{ij0}}} \qquad (i = 1, \ldots, r \;\; j = 1, \ldots, c). \qquad (7.26)$$

Asymptotically, under the global null hypothesis H^F_0, the T^F_{ij}'s follow an *rc*-variate normal distribution with expectation **0** and block diagonal correlation matrix $\mathbf{R}^F = \left\{ r^F_{i(j_1,j_2)} \right\}$.

In this situation, the correlation coefficients $r^F_{i(j_1,j_2)}$ can be written as a product

$$r^F_{i(j_1,j_2)} = \frac{v_{(ij_1),(ij_2)}}{\sqrt{v_{(ij_1),(ij_1)}v_{(ij_2),(ij_2)}}} = \frac{\sqrt{N_{ij_1}N_{ij_2}}\,n_{i0}^{-1}s_{ij_1j_2}}{\sqrt{\frac{N^2_{ij_1}}{n_{i0}n_{ij_1}}s^2_{ij_10}}\sqrt{\frac{N^2_{ij_2}}{n_{i0}n_{ij_2}}s^2_{ij_20}}} = b^F_{ij_1}b^F_{ij_2} \qquad (7.27)$$

where $b^F_{ij} = \sqrt{\frac{n_{ij}}{N_{ij}}}$, by noticing that $s_{ij_1j_2} = Cov(F_{ij_1}(X_{i01}), F_{ij_2}(X_{i01})) = Var(F_{ij_1}(X_{i01})) =$ $= Var(F_{ij_2}(X_{i01}))$ under H^F_0.

It follows that the test procedure $T^F = \max_{1\le i\le r;1\le j\le c}\left\{\left|T^F_{ij}\right|\right\}$, which rejects the global null hypothesis H^F_0 in favor of the two-sided alternative hypothesis H^F_1 if $T^F > d^F_{2a}$, asymptotically controls the FWE, provided that $P_{H^F_0}\left(T^F > d^F_{2a}\right) = a$. d^F_{2a} is the two-sided 1 - a percentage point of the *rc*-variate normal distribution with correlation matrix $\mathbf{R}^F$, i.e. $d^F_{2a} = |x|_{rc,\mathbf{R}^F;1-a}$.

The block diagonal correlation matrix $\mathbf{R}^F$ partially satisfies the product correlation structure and hence the computation of d^F_{2a} doesn't involve an *rc*-variate integral but can be expressed by univariate integrals as already shown in Chapters 3 and 4.

For example, the probability $P_{H_0^F}\left(T^F \le t\right)$ can be expressed as:

$$P\left(T^F \le t\right) = P\left(-t \le T^F_{ij} \le t \;\; \forall ij\right) = \prod_{i=1}^{r} \Phi_c\left(\mathbf{-t},\mathbf{t};\mathbf{0},\left\{r^F_{i(j_1,j_2)}\right\}\right) = \tag{7.28}$$

$$= \prod_{i=1}^{r}\left[\int_{-\infty}^{\infty}\prod_{j=1}^{c}\left(\Phi\left(\frac{b^F_{ij}y+t}{\sqrt{1-\left(b^F_{ij}\right)^2}}\right)-\Phi\left(\frac{b^F_{ij}y-t}{\sqrt{1-\left(b^F_{ij}\right)^2}}\right)\right)j\left(y\right)dy\right]$$

where $\Phi_c\left(\mathbf{-t},\mathbf{t};\mathbf{0},\left\{r^F_{i(j_1,j_2)}\right\}\right)$ is the *c*-variate normal integral with expectation $\mathbf{0}$ and correlation matrix characterized by the $r^F_{i(j_1,j_2)}$'s over the rectangular region with lower and upper integration bounds $-t$ and t respectively.

The probability expression within square brackets can directly be computed within the SAS system using the statement:

PROBMC('DUNNETT2', t, ., ., c, b^F_{i1}, b^F_{i2}, . . ., b^F_{ic})

as already introduced in Section 4.2.

Munzel and Hothorn (2001) mentioned that simulation studies showed that the accuracy of the approximation could be slightly improved by using the 'Steel'-factors $b^S_{ij} = \sqrt{\frac{n_{ij}}{N_{ij}+1}}$ instead of the 'Dunnett'-factors $b^F_{ij} = \sqrt{\frac{n_{ij}}{N_{ij}}}$. Using these factors in the unstratified situation would result in the well-known asymptotic Steel test (1959).

Adjusted p-values and simultaneous confidence intervals can be provided as illustrated for the problem phrased in terms of relative treatment effects earlier in this section.

Adjusted p-values $\tilde{p}^F_{ij}$'s can be computed as:

$$\tilde{p}_{ij}^F = 1 - \Phi_{rc}\left(-\left|t_{ij}^F\right|, \left|t_{ij}^F\right|; \mathbf{0}, \mathbf{R}^F\right) \tag{7.29}$$

where $\left|t_{ij}^F\right|$ is the observed value of the test statistic $\left|T_{ij}^-\right|$ (i = 1, …, r and j = 1, …, c).

Two-sided 100(1 - a)% simultaneous confidence intervals for p_{ij} can be computed as:

$$\hat{p}_{ij} \pm |x|_{rc,\mathbf{0},\mathbf{R}^F;1-a} \sqrt{\frac{N_{ij}^2}{n_{i0} n_{ij}} \hat{s}_{ij0}^{\,2}} \qquad (i = 1, \ldots, r \text{ and } j = 1, \ldots, c). \tag{7.30}$$

where $\hat{s}_{ij0}^{\,2}$ is given in formula (7.25).

<u>Computations of the example</u>

The test in terms of distribution functions is illustrated using the standard example introduced in Chapter 3.

The situation is as follows:

$$H_0 : F_{11} = F_{12} = F_{10} \text{ and } F_{21} = F_{22} = F_{20}$$

versus

$$H_1 : F_{11} \neq F_{10} \text{ or } F_{12} \neq F_{10} \text{ or } F_{21} \neq F_{20} \text{ or } F_{22} \neq F_{20}$$

where i = 1 and i = 2 represent again the males and females and j = 1 and j = 2 represent the low and high dose respectively.

The data are analyzed using the multivariate normal distribution and the ‘Dunnett’-factors.

The results of the analysis are presented in the following table.

Table 7.1 Analysis of example applying nonparametric procedure

Stratum	Contrast	Estimator	Adjusted p-value	Asymptotic two-sided
		$\hat{p}_{ij}$	$\tilde{p}_{ij}^{F}$	95% Confidence interval
M	Plac-Low	0.143	0.056	(-0.221, 0.506)
	Plac-High	0.020	0.013	(-0.365, 0.435)
F	Plac-Low	0.350	0.783	(-0.032, 0.732)
	Plac-High	0.060	0.027	(-0.345, 0.465)

The program code can be found in program Ch7.sas of Appendix 3, which also computes the two sided upper percentage point $d_{2,0.05}^{F} = d_{2}^{F}(0.05,2,2,\{ r_{i(j_1,j_2)}^{F} \}) = |x|_{4,\mathbf{R}^F;0.95} = 2.483$.

The analysis shows that all estimators of the relative effects p_{ij} are smaller than 0.5. Qualitatively the conclusions are similar to the analysis of the two-sided testing problem in terms of differences assuming normal distributed data and testing for $\Delta = 0$ as illustrated in Section 4.2: the low dose is not statistically significant (p < 0.05) but the high dose is statistically significant in both genders. The interpretation of the $\hat{p}_{ij}$'s is more difficult than the interpretation of the estimators of the relative differences of the active treatment versus control in the situation of Chapter 4.

Notice that the confidence intervals include values < 0 which are impossible.

8 Resampling methods

Resampling method are methods in which the observed data are used repeatedly, in a computer intensive simulation analysis, to provide inferences. The idea is to re-assigned the observed data randomly and to re-compute the test statistics many many times. The original test statistic is considered unusual if it is unusual compared to the resampling distribution of the test statistic. Adjusted p-values are the natural output of these resampling methods. The computation of critical values and resampling standard errors of these critical values are more complicated.

General advantages of resampling methods are that they can cope with many complicated testing situations and that they have the ability to incorporate distributional characteristics, which can make the tests more robust. The main disadvantage is the heavy computational effort although this is less of a problem now a day.

It is not the intention of this chapter to provide a complete overview of all possibilities of resampling methods in the context of many-to-one comparisons in a stratified design, but just to show some methods that are standard available within the SAS system. The following three resampling methods are considered: the parametric simulation method of Edwards and Berry (1987) in Section 8.1, the bootstrap method in Section 8.2 and the permutation method in Section 8.3. They are illustrated for the example introduced in Chapter 3 and compared by means of a small simulation study in Section 8.4.

An extensive overview of resampling-based multiple testing methods is given in the book by Westfall and Young (1993).

8.1 Stochastic approximation

For many multiple testing situations, the upper-α percentage point can't be easily determined. Therefore Edwards and Berry (1987) proposed a method that approximated the upper-α percentage point, say d_a , by parametric computer simulation. The basic idea is to substitute a random variable D_a obtained by computer simulation instead of d_a itself, in much the same way as s^2 is substituted for s^2.

The validity of this method depends on a result, which is already referred to be Dwass (1957).

Assume that D_1, ..., D_N and D are independent random variables, each with the same continuous probability distribution. Given a probability level a, let $r = (N + 1)(1 - a)$ and suppose that a and *N* are such that *r* is and integer. Then $P\left(D > D_{(r)}\right) = a$ if $D_{(1)} \leq \ \ldots \leq D_{(N)}$ are the order statistics of $D_1, \ldots, D_N$.

A disadvantage of using $D_a = D_{(r)}$ instead of d_a itself is that extraneous variability is introduced. However, the amount of added variability is under control by choosing an appropriate simulation size *N*. Under control means that the distance between the probability $P\left(D \leq D_{(r)}\right)$ and the probability of the true upper-α percentage point $P\left(D \leq d_a\right) = 1 - a$ is as small as requested with high probability.

Let F denote the cumulative distribution function of D. Then $F\left(D_{(r)}\right)$ has a beta distribution with shape parameters *r* and *N* – *r* + 1, i.e. $\text{Beta}\left(r, N - r + 1\right)$, since $F\left(D_{(r)}\right)$ can be seen as the *r*-th order statistic of a random sample of size *N* from a $\text{Uniform}\left(0,1\right)$ distribution. Hence $E\left(F\left(D_{(r)}\right)\right) = 1 - a$ and $Var\left(F\left(D_{(r)}\right)\right) = a\left(1 - a\right)/\left(N + 2\right)$.

For example, with $a = 0.05$ and *N* + 1 = 3200, r = 3040 and $\sqrt{Var\left(F\left(D_{(r)}\right)\right)} < 0.0039$, placing the tail area of $D_{(r)}$ within 0.01 of 1 - a with 99% confidence.

Thus, for any desired g (> 0) and e (> 0), the simulation size *N* can be set so that the tail area for the simulated percentage point $D_{(r)}$ is within g of 1 - a with 100(1 - e)% confidence, i.e. in equation form: $P\left(\left|F\left(D_{(r)}\right) - \left(1 - a\right)\right| \leq g\right) = 1 - e$.

Lets consider the testing situation

$$H_0 : m_{ij} = m_{i0} \quad (i = 1,\ldots,r \quad j = 1,\ldots,c)$$

versus (8.1)

$$H_1 : \exists ij : m_{ij} > m_{i0} \quad (i = 1,\ldots r \quad j = 1,\ldots,c)$$

with test statistic $D = \max_{1 \leq i \leq r; 1 \leq j \leq c}\{D_{ij}\}$ where $D_{ij} = \dfrac{\bar{X}_{ij} - \bar{X}_{i0}}{s\sqrt{n_{ij}^{-1} + n_{i0}^{-1}}} \quad (i = 1,\ldots,r \quad j = 1,\ldots,c)$ as extensively discussed in Chapter 3.

Then D is the maximum of the random vector $\left(D_{11},\ldots,D_{rc}\right)$ that has an *rc*-variate central t-distribution with n degrees of freedom and correlation matrix characterized by the set of $\left\{b_{ij}\right\}$ under the null hypothesis. Apply the results above by noticing that here F is the cumulative distribution function of the maximum of an *rc*-variate t-distributed random vector.

For this testing situation the stochastic approximation resampling algorithm can be outlined as follows to obtain an estimate of the upper-α percentage point:

(i) initialize a, *r*, *c*, n, $\left\{b_{ij}\right\}$ and *N* (or g and e)

(ii) Do *l* = 1 to *N*

- obtain D_l as the maximum of a generated *rc*-variate t-distributed random variable
- store D_l in an ordered way

(iii) write $D_a = D_{(r)}$

The SAS system offers the possibility to apply the stochastic approximation method for this situation directly. Adjusted p-values and confidence limits can be computed using the SIMULATE adjustment option in the LSMEANS statement of the procedure PROC MIXED. The simulation size *N*, g and e can be controlled with the simulation options NSAMP, ACC and EPS respectively. By default $g = 0.005$ and $e = 0.01$. (See also the SAS manual (1996))

Computations of the example

The example introduced in Section 3.1 has been analyzed using the stochastic approximation resampling technique with g = 0.001 and e = 0.01. The adjusted p-values and simultaneous 95% confidence intervals are shown in the following table.

Table 8.1 Analysis of example applying stochastic approximation method

Stratum	Contrast	Estimate	Adjusted p-value	95% Confidence interval
M	Plac-Low	0.864	0.072	(-0.068, ∞)
	Plac-High	2.163	<0.001	(1.128, ∞)
F	Plac-Low	0.582	0.287	(-0.394, ∞)
	Plac-High	1.265	0.015	(0.230, ∞)

The SAS program code can be found in Appendix 3.

The table shows almost identical adjusted p-values and simultaneous confidence intervals as obtained by using the extended Dunnett procedure as discussed in Chapter 3 assuming normal distributed data. But this is in line with the expectation, because this method makes also use of the multivariate t distribution.

8.2 Bootstrap

The bootstrap resampling method was introduced by Efron (1979) and is widely used for many different situations since then.

The idea of the bootstrap method is to approximate the true but unspecified distribution function by the empirical distribution function and use this empirical distribution function in the remainder as if it is the true distribution function. The following simple situation illustrates this.

Assume that $Y_1, \ldots, Y_n$ are a random sample from a larger population with mean m and further unknown underlying distribution, and let F denote the cumulative distribution function of Y_i. Assume that $T = T_n(Y_i, m)$ is the test statistic to test some hypothesis about the location of m and that large values of T are in favor of the alternative hypothesis. Then one is interested in the probability $P(T > t)$. Notice that the probability $P(T > t)$ depends on F.

Then the unknown distribution function of the test statistic can be approximated through simulation by substitution the empirical distribution function $\hat{F}(y) = (\# Y_i\text{'s} \le y)/n$ for F:

$$P(T > t) = P(T > t|F) \approx P\left(T > t\middle|\hat{F}\right) \tag{8.2}$$

The bootstrap method generates pseudo-data sets $\{Y_i^*\}$ having this distribution function $\hat{F}$ by sampling observations with replacement from the original set of observations.

Then the probability $P(T_n(Y_i, m) > t)$ may be estimated by $\dfrac{\#\left(T_n(Y_i^*, \hat{m}) > t\right)}{N}$, where $T_n(Y_i^*, \hat{m})$ is the test statistics computed on the pseudo-data set $\{Y_i^*\}$ and $\hat{m}$ is an estimate of m based on the observed data set $\{Y_i\}$ and N is the number of times a pseudo-data set has been generated.

Westfall and Young (1993) mentioned that there are two sources of error in this process: the simulation error and the error inherent in using $\hat{F}$ instead of F. The accuracy of the simulation error, which is a function of simulation size *N*, can be estimated using the binomial distribution and goes to zero for increasing *N*. The error induced by replacing F with $\hat{F}$ also goes to zero for increasing *N*, since $\hat{F}$ generally approaches F for increasing *N*.

It is not the intention of this section to go into too much detail, though it is worthwhile to briefly mention two important guidelines for the bootstrap resampling method as highlighted by Hall and Wilson (1991): centering and pivoting. The first guideline concerning centering means that resampling should be done to reflect the distribution under the null hypothesis even if the observations are drawn from a population that fails to satisfies the null hypothesis. The second guideline means that the test statistic is pivotal, i.e. under the null hypothesis the sampling distribution of the test statistic should not depend on the distribution function of the observed data within the assumed family of possible distributions. These guidelines can be extended to the multiple testing situation, were the concept of pivotality is then called subset pivotality. It can be shown that the bootstrap procedure, under the subset pivotality condition, has (approximately) control of the FWE in the strong sense. For example the subset pivotality condition is satisfied in the case of multiple comparisons using the t-statistics when the data come from a location shift model, which does not require normal distributions. These details as well other details are well described by Westfall and Young (1993), chapter 2.

Consider again the one-sided superiority testing problem for the many-to-one comparisons in a stratified design as stated by (8.1).

For this testing situation the bootstrap resampling algorithm to obtain adjusted p-values can be outlined as follows:

(i) Initialize counting variables $Count_{ij} = 0$ (*i* = 1, ..., *c* and *j* = 1, …, *r*)

(ii) Center the data $C_{ijk} = X_{ijk} - \overline{X}_{ij}$ (*i* = 1, ..., *c*, *j* = 0, 1, …, *r* and *k* = 1, …, n_{ij})

(iii) Generate resampled data X^*_{101}, …, $X^*_{1cn_{1c}}$ to X^*_{rc1}, …, $X^*_{rcn_{rc}}$, within each stratum a with replacement sample from the centered data C_{101}, …, $C_{1cn_{1c}}$ to C_{rc1}, …, $C_{rcn_{rc}}$

(iv) Compute the sample means $\overline{X^*_{ij}}$ as well as the residual mean square s^{*2} from the permutated dataset $\left\{X^*_{ijk}\right\}$

(v) Compute the test statistics $D_{ij}^{*}=\dfrac{\overline{X_{ij}^{*}}-\overline{X_{i0}^{*}}}{s^{*}\Big/\sqrt{n_{ij}^{-1}+n_{i0}^{-1}}}$ (i = 1, ..., c and j = 1, …, r)

(vi) If $\max\limits_{1\le i\le k;1\le j\le k}\left\{D_{ij}^{*}\right\}\ge d_{ij}$ where d_{ij} is the observed test statistic based on the original data, then increment the count variable $Count_{ij}=Count_{ij}+1$ (i = 1, ..., c and j = 1, …, r)

(vii) Repeat steps (i) to (vi) N times. The estimated adjusted p-value is $\tilde{p}_{ij}=Count_{ij}\big/N$ (i = 1, ..., c and j = 1, …, r)

The bootstrap procedure is available in the procedure PROC MULTTEST within the SAS system. The option BOOTSTRAP specifies that the p-values be adjusted using the bootstrap method and the NSAMPLE options specifies the number of resamples. Continuous variables are mean-centered by default prior to resampling. The t-test for the mean can be requested by specifying MEAN in the TEST statement. (See also the SAS manual (1996) or Westfall and Young (1993).)

The SAS program code for the example can be found in Appendix 3.

The bootstrap resampling technique can also be applied in a step-down manner directly available in procedure PROC MULTTEST by using the option STEPBOOT. This will not be discussed here any further. See e.g. Chapter 2 of Westfall and Young (1993)

The output of the standard example can be found at the end of this Chapter.

8.3 Permutation

The idea behind permutation or rerandomisation tests goes back to Fisher (1935). The concept is as follows. Suppose that treatments are randomly assigned to the experimental units, but these units themselves are not randomly selected from a larger population. Then the only legitimate form of inference seems to be based on the probability mechanism of the random assignments of the treatments, see Ludbrook and Dudley (1998). Permutation tests calculate how extreme the observer results are in comparison with those that would have occurred with other randomisations. The permutation approach is conditional with respect to the data and so it gives rise to conditional inferences, whereas bootstrap is not a strictly conditional procedure, in fact it is asymptotically unconditional.

The computation of the permutation adjusted p-values is almost identical as the computation of the bootstrap adjusted p-values. The exception is that permutation tests use resampling without replacement whereas bootstrap tests use resampling with replacement. In the spirit of rerandomisation analyses, the raw data are resampled in contrast to the bootstrap method where the variables are centered prior to resampling.

Adapting the bootstrap algorithm of Section 8.2 provides the following permutation algorithm:

(i) Initialize counting variables $Count_{ij} = 0$ (*i* = 1, ..., *c* and *j* = 1, …, *r*)

(ii) Generate resampled data X^*_{101}, …, $X^*_{1cn_{1c}}$ to X^*_{rc1}, …, $X^*_{rcn_{rc}}$, within each stratum a without replacement sample (or permutation) of the observed data X_{101}, …, $X_{1cn_{1c}}$ to X_{rc1}, …, $X_{rcn_{rc}}$

(iii) Compute the sample means $\overline{X^*_{ij}}$ as well as the residual mean square s^{*2} from the permutated dataset $\left\{X^*_{ijk}\right\}$

(iv) Compute the test statistics $D^*_{ij} = \dfrac{\overline{X^*_{ij}} - \overline{X^*_{i0}}}{s^* \Big/ \sqrt{n^{-1}_{ij} + n^{-1}_{i0}}}$ (*i* = 1, ..., *c* and *j* = 1, …, *r*)

(v) If $\max\limits_{1 \le i \le k; 1 \le j \le k} \left\{D^*_{ij}\right\} \ge d_{ij}$ where d_{ij} is the observed test statistic based on the original data, then increment the count variable $Count_{ij} = Count_{ij} + 1$ (*i* = 1, ..., *c* and *j* = 1, …, *r*)

(vi) Repeat steps (i) to (v) N times. The estimated adjusted p-value is $\tilde{p}_{ij} = Count_{ij} \big/ N$ (*i* = 1, ..., *c* and *j* = 1, …, *r*)

The permutation resampling method can also be applied in a step-down manner using the option STEPPERM in the procedure PROC MULTTEST. However, one should be cautious because this process doesn’t control the FWE in the strong sense. Westfall and Wolfinger (2000) illustrate that the permutation resampling technique within PROC MULTTEST does not provide closed tests in the situation of comparisons of means involving more than three groups. The reason is that PROC MULTTEST always uses the global hypothesis for calculating the adjusted p-values.

Like the bootstrap resampling technique, the permutation resampling technique can also be applied in a step-down manner by using the option STEPBOOT.

Computations of the example

The same example has been analyzed using the bootstrap and permutation resampling techniques. The simulation size for both the bootstrap and permutation techniques was set on *N* = 50000. The adjusted p-values are shown in the following table. The adjusted p-values obtained using the stochastic approximation technique of Section 8.1 are added for reason of comparisons.

Table 8.2 Single-step adjusted p-values of resampling methods

Stratum	Contrast	Estimate	Stoch. Approx.	Bootstrap (N=50000)	Permutation (N=50000)
M	Plac-Low	0.864	0.072	0.078	0.074
	Plac-High	2.163	<0.001	<0.001	<0.001
F	Plac-Low	0.582	0.287	0.293	0.296
	Plac-High	1.265	0.015	0.018	0.015

These figures show that both the bootstrap and permutation resampling techniques result in similar adjusted p-values and that those are also similar to the adjusted p-values obtained using the stochastic approximation technique.

Although the step-down adjusted p-values were only briefly mentioned in this chapter, these are presented in the following table. The adjusted p-values using the step-down method proposed by Cheung and Holland (1994) assuming normal distributed data as described in Section 3.5 are also included.

Table 8.3 Step-down adjusted p-values of bootstrap and permutation methods

Stratum	Contrast	Estimate	Cheung & Holland	Bootstrap (N=50000)	Permutation (N=50000)
M	Plac-Low	0.864	0.039	0.041	0.039
	Plac-High	2.163	<0.001	<0.001	<0.001
F	Plac-Low	0.582	0.089	0.104	0.075
	Plac-High	1.265	0.011	0.015	0.008

The p-values of the bootstrap method are in the same order of the step-down adjusted p-values assuming normal distributed data although slightly more conservative. The permutation method provides more liberal adjusted p-values, although one should not forget that this process doesn’t control the FWE in the strong sense.

The program code can be found in Appendix 3.

8.4 Simulation study

A small simulation study was conducted to compare the behaviour of the resampling procedures described above as well as the extended Dunnett procedure for the stratified situation and the Bonferroni corrected Dunnett-within-strata procedure. This latter procedure consists of applying the original Dunnett procedure within each of the strata at a significance level of α / r to control the FWE in the strong sense.

Some of the many configurations that were investigated are reported here.

The number of observations for each of the control treatments is chosen to be equal, say n_0, and the number of observations for each of the active treatments within each of the strata is also taken to be equal, say n_a. The following pairs of combinations of (n_0, n_a) were considered: (2, 5), (5,5), (5,10), (10,5) and (10,10). The number of strata is taken to be two (*r* = 2) and there are three active treatment arms within each of the strata (*c* = 3). The random error terms are taken as independently identically distributed random variables from the standard normal distribution or from the lognormal distribution, which are being generated as the exponential of a standard normal random variable. The performance of these methods are compared under the null hypothesis to check whether the FWE is correctly kept at an alpha level of α = 0.05 using 10.000 replications for each of the settings. The results are shown in Table 8.4.

All five methods approximate the alpha level in the situation of normally distributed data quit well. However, in case of lognormal distributed data, the stratified Dunnett procedure, the Bonferroni corrected Dunnett-within-strata method as well as the stochastic approximation technique become much too liberal. This is not a surprise because it is well known that the original Dunnett's procedure does not control the FWE in case of non-normal distributed data. The other two resampling methods (bootstrap and permutation) behave much better in this situation.

Table 8.4: Empirical Type I errors ($\alpha = 0.05$) for $r = 2$ groups and $c = 3$ active treatment arms within each group based on 10000 replications

		Method				
		Stoch.			Stratified	Bonferroni-
n_0	n_a	Approx.	Bootstrap	Permutation	Dunnett	Dunnett
Normal data						
2	5	0.0508	0.0487	0.0510	0.0507	0.0512
5	5	0.0525	0.0493	0.0527	0.0534	0.0483
5	10	0.0529	0.0495	0.0516	0.0519	0.0565
10	5	0.0563	0.0529	0.0564	0.0565	0.0508
10	10	0.0494	0.0487	0.0502	0.0497	0.0512
Lognormal data						
2	5	0.0173	0.0220	0.0337	0.0170	0.0121
5	5	0.0737	0.0630	0.0593	0.0730	0.0442
5	10	0.0280	0.0403	0.0403	0.0283	0.0835
10	5	0.1097	0.0637	0.0567	0.1087	0.0155
10	10	0.0630	0.0560	0.0540	0.0627	0.0468

In addition, the methods are compared under the alternative hypothesis using the any-pair power. The setting of the one-sided alternative hypothesis represents a linear shift of 0.5 for each of the active treatment means; i.e. the means of the three active treatments have a positive shift of 0.5, 1.0 and 1.5, respectively compared to placebo. The simulations are conducted using 3000 replications for each of the settings. Table 8.5 shows the results.

In the situation of normally distributed data, the stratified Dunnett and the resampling methods have very similar power results, although the bootstrap method seems to have the lowest power of these four methods. Even in the case of only two strata, the Bonferroni corrected Dunnett-within-strata procedure shows the lowest power of these five methods for almost all settings as shown above. This is in line with our expectation and one can image that the loss of power in comparison to the other methods will increase if the number of groups (*r*) increases. Before one compares the power of these five methods directly in the situation of lognormal distributed data, one should keep in mind that the stratified Dunnett, the stochastic approximation and the Bonferroni corrected Dunnett-within-strata methods are not maintaining the FWE. On the other hand, Table 8.5 does not show that the bootstrap and permutation methods have a much lower power in these settings.

Table 8.5: Any-Pair power a linear shift and for *r* = 2 groups, *c* = 3 active treatment arms within each group based on 3000 replications

		Method				
		Stoch.			Stratified	Bonferroni-
N_0	n_a	Approx.	Bootstrap	Permutation	Dunnett	Dunnett
Normal data						
2	5	0.5317	0.5117	0.5353	0.5327	0.5017
5	5	0.7707	0.7477	0.7717	0.7687	0.7343
5	10	0.8820	0.8793	0.8853	0.8833	0.8640
10	5	0.8840	0.8713	0.8867	0.8843	0.8800
10	10	0.9760	0.9743	0.9757	0.9760	0.9730
Lognormal data						
2	5	0.2280	0.2970	0.3190	0.2307	0.2670
5	5	0.4277	0.3943	0.3927	0.4280	0.4707
5	10	0.4367	0.5143	0.5070	0.4370	0.5700
10	5	0.5107	0.4017	0.4050	0.5127	0.4930
10	10	0.5913	0.5557	0.5603	0.5923	0.6487

These simulation results indicate that the stratified Dunnett procedure maintains the FWE in the situation of normal distributed data, as do the other proposed methods. The power of all five methods, except the Bonferroni style adjusted method, are similar for the situation of normal distributed data.

The simulation study also indicates that the FWE of the stratified Dunnett procedure, the Bonferroni corrected Dunnett-within-strata method and the stochastic approximation method are inflated in case of lognormal distributed data. Both the bootstrap and permutation resampling methods seem to behave better without substantial loss in power.

Similar results are found for other settings. Evaluation of the all-pairs power instead of the any-pair power shows similar results as presented in Table 8.5 and evaluation of the situation where the error terms are exponential distributed shows similar results as presented for the lognormal distributed random error terms.

This suggests that the resampling techniques, and in particular the bootstrap method, that are standard available within SAS seem to be worthwhile to calculate p-values in case of non-normal distributed data. Keep in mind that PROC MULTTEST doesn’t provide simultaneous confidence intervals.

9 Summary and Outlook

The topic of this thesis was to investigate multiple comparisons procedures for many-to-one comparisons in a stratified design while controlling the familywise error rate strongly at level α .

The situation of testing several active treatments versus a control treatment in each of several strata simultaneously does occur in several practical settings as illustrated by the examples shown in the introduction of this thesis. A naïve way would be to perform Dunnett's procedure within each stratum without any other multiplicity correction. This would lead to an inflation of the FWE in the overall experiment. And the use of an additional Bonferroni correction to correct for the number of strata would result in a conservative approach under the assumption of an unknown common variance as assumed in this thesis.

Cheung and Holland (1992) extended the Dunnett procedure to the stratified situation. However, they only derived upper percentage points for a common correlation coefficient and suggested interpolation of these percentage points for all other testing situations. This thesis shows that these approximations are not needed any more and that correct percentage points can be computed quite easily with current available software (SAS).

In addition, this thesis described how power calculations and sample size determination could be performed, which was not considered by Cheung and Holland.

Although the interest for most of the many-to-one comparisons in practical testing situations is in showing that an active treatment is superiority to the control treatment or different from the control treatment, there are testing situations where this type of trials are not appropriate. This thesis showed that it is also feasible to perform many-to-one comparisons in a stratified design in case of a non-inferiority testing problem or in case of a global equivalence testing problem, while still controlling the FWE.

Also if the testing problem is better expressed in terms of proportions rather than in terms of differences, it has been showed in this thesis how to perform many-to-one comparisons in a stratified layout.

All these procedures assume that the data are normally distributed. If this assumption is suspect, the use of a nonparametric approach might be more appropriate. Munzel and Hothorn (2001) discussed an asymptotic approach to perform many-to-one comparisons for the one-way layout based on a pairwise ranking procedure. In this thesis it has been illustrated how this procedure could be extended to handle the testing problem in case of a stratified two-way layout.

At last this thesis discussed the stochastic approximation method, the bootstrap method and the permutation method as alternative methods. It was illustrated for a situation were these three computer intensive resampling methods are standard available within the SAS system.

To summarize, this thesis showed that it is possible to perform many-to-one comparisons in a stratified two-way layout for several different practical testing situations and provided program code to analyze these situations.

The work of this thesis also generated new areas of interest.

In this thesis it supposed that the experimenter is interested to analyze the data within each of the strata, while controlling the overall FWE. However in practice this is not always known upfront, but may become clear during the course of the experiment or even when the experiment is in the analysis phase. Suppose for example, that the experiment was designed to perform many-to-one comparisons averaged over all levels of a second factor. Then assume that it becomes clear, during the course of the experiment due to external information, that one cannot speak about *the* treatment effect, but that there are different treatment effects for each of the levels of the second factor. The same situation can occur if a first statistical analysis shows significant treatment by stratum interactions. A similar phenomenon arises in the analyses of subgroups in clinical trials. It would be interesting to develop strategies to deal with these practical situations.

A second interesting topic would be to describe the procedures and to provide corresponding computer programs to perform stratified many-to-one comparisons in a two-way analysis of covariance model which allows adjustment for covariates like a baseline value or a continuous covariate like age (see also Wong and Cheung (2000)).

Another inviting topic that could need more attention is the construction of simultaneous confidence intervals in case the testing problem is expressed in terms of proportions.

References

Akritas, M.G. and Brunner, E. A unified approach to rank tests in mixed models. Journal of Statistical Planning and Inference, 61, 249-277 (1997)

Altman, D. and Bland, M. Absence of evidence is not evidence of absence. British Medical Journal, 311, 485 (1995)

Bauer, P. Multiple testing in clinical trials. Statistics in Medicine, 10, 871-890 (1991)

Bauer, P and Kieser, M. A unifying approach for confidence intervals and testing of equivalence and difference. Biometrika, 83, 4, 934-937 (1996)

Bechhofer, R.E. and Dunnett, C.W. Percentage points of multivariate t distributions, in: Odeh, R.E. and Davenporth, J.M. (Eds.), Selected Tables in Mathematical Statistics, Vol. 11, American Mathematical Society, Providence, 1-371 (1988)

Bechhofer, R.E. and Tamhane, A.C. An iterated integral representation for a multivariate normal integral having a block covariance structure. Biometrika, 62, 3, 615-619 (1974)

Berger, R.L. Multiparameter hypothesis testing and acceptance sampling. Technometrics, 24, 295-300 (1982)

Bretz, F., Genz, A., Hothorn, L.A. On the numerical availability of multiple comparison procedures. Biometrical Journal, 43, 5, 645-656 (2001)

Brunner, E. and Munzel, U. The non-parametric Behrens-Fisher problem: asymptotic theory and small-sample approximation. Biometrical Journal, 42, 17-25 (2000)

Brunner, E., Puri, M.L., Sun S. Nonparametric methods for stratified two-sample designs with application to multiclinic trials. Journal of the American Statistical Association, 90, 1004-1014, (1995)

Brunner, E. and Puri, M.L. Nonparametric methods in design and analysis of experiments, in: Ghosh, S. and Rao, C.R. (Eds.), Handbook of Statistics 13: Design and Analysis of Experiments, Chapter 19, Elsevier Science, Holland (1996)

Cheung, S.H. and Holland, B. Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics, 47, 21-32 (1991)

Cheung, S.H. and Holland, B. Extension of Dunnett's multiple comparison procedure with differing sample sizes to the case of several groups. Computational Statistics & Data Analysis, 14, 165-182 (1992)

Cheung, S.H. and Holland, B. A step-down procedure for multiple tests of treatment versus control in each of several groups. Statistics in Medicine, 13, 2261-2267 (1994)

Cornish, E.A. The multivariate t-distribution associated with a set of normal sample deviates. Australian Journal of Physics, 7, 531-542 (1954)

Curnow, R.N. and Dunnett, C.W. The numerical evaluation of certain multivariate normal integrals. Annals of Mathematical Statistics, 33, 571-579 (1962)

Dubey, S.D. Some thoughts on the one-sided and two-sided tests. Journal of Biopharmaceutical Statistics, 1, 139-150 (1991)

Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1096-1121 (1955)

Dunnett, C.W. Multivariate normal probability integrals with product correlation structure; Algorithm AS251. Applied Statistics, 38, 564-579 (1988)

Dunnett, C.W. and Gent M. An alternative to the use of two-sided tests in clinical trials. Statistics in Medicine, 15, 1729-1738 (1996)

Dunnett, C.W., Horn, M., Vollandt, R. Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. Journal of Statistical Planning and Inference, 97, 367-384 (2001)

Dunnett, C.W. and Sobel, M. A bivariate generalisation of Student's t-distribution with tables for certain special cases. Biometrika, 41, 153-169 (1954)

Dunnett, C.W. and Sobel, M. Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution. Biometrika, 42, 258-260 (1955)

Dunnett, C.W. and Tamhane, A.C. Step-down multiple tests for comparing treatments with a control. Statistics in Medicine, 10, 939-947 (1991)

Dunnett, C.W. and Tamhane, A.C. A step-up multiple test procedure. Journal of the American Statistical Association, 87, 162-170 (1992)

Dunnett, C.W. and Tamhane, A.C. Step-up multiple testing of parameters with unequally correlated estimates. Biometrics, 51, 217-227 (1995)

Dwass, M. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics, 28, 181-187 (1957)

E9 Statistical principles for clinical trials. International conference of harmonization of technical requirements for registration of pharmaceutical for human use, September 8 (1998)

Edwards, D. and Berry, J.J. The efficiency of simulation-based multiple comparisons. Biometrics, 43, 913-928 (1987)

Efron, B. Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26 (1979)

Fisher, R.A. The design of experiments. Oliver Boyd, Edinburgh (1935)

Fieller, E.C. Some problems in interval estimation. Journal of the Royal Statistical Society B, 16, 175-185 (1954)

Finner, H. Stepwise multiple test procedures and control of directional errors. Annals of Statistics, 27, 274-289 (1999)

Finner, H. and Roters, M. Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. The Annals of Statistics, 26, 505-524 (1998)

Fisher, L.D. The use of one-sided tests in drug trials: an FDA advisory committee member's perspective. Journal of Biopharmaceutical Statistics, 1, 151-156 (1991)

Fligner, M.A. A note on two-sided distribution-free treatment versus control multiple comparisons. Journal of the American Statistical Association, 79, 208-211 (1984)

Gabriel, K.R. Simultaneous test procedures – some theory of multiple comparisons. Annals of Mathematical Statistics, 40, 224-250 (1969)

Genz, A. Numerical computations of multivariate normal probabilities. Journal of Computational and Graphical Statistics, 1, 141-150 (1992)

Genz, A. and Bretz, F. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation, 63, 361-378 (1999)

Grechanovsky, E. and Hochberg, Y. Closed procedures are better and often admit a shortcut. Journal of Statistical Planning and Inference, 76, 79-91 (1999)

Hall, P and Wilson, S.R. Two guidelines for bootstrap hypothesis testing. Biometrics, 47, 757-762 (1991)

Hauschke, D. Biometrische Methoden zur Planung und Auswertung von Sicherheitsstudien. Habilitationsschrift in German, University of Dordmunt (1999)

Hauschke, D., Kieser, M, Diletti, E, Burke, M. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. Statistics in Medicine, 18, 93-105 (1999)

Hayter, A.J. and Hsu, J.C. On the relationship between stepwise decision procedures and confidence sets. Journal of the American Statistical Association, 89, 128-136 (1994)

Hayter, A.J. and Liu, W. A method of power assessment for tests comparing several treatments with a control. Cummunications in Statistics, Part A – Theory and Methods, 21, 1871-1889 (1992)

Hayter, A.J., Miwa, T. and Liu, W. ‘Combining the advantages of one-sided and two-sided procedures for comparing several treatments with a control. Journal of Statistical Planning and Inference, 86, 81 – 99 (2000)

Hayter, A.J. and Tamhane, A.C. Sample size determination for step-down multiple test procedures: orthogonal contrasts and comparisons with a control. Journal of Statistical Planning and Inferences, 27, 271-190 (1991)

Hochberg, Y. and Tamhane, A.C. Multiple comparison procedures. John Wiley & Sons, New York (1987)

Holm, S.A. Simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70 (1979)

Horn, M. and Vollandt, R. Sample sizes for comparisons of k treatments with a control based on different definitions of the power. Biometrical Journal, 40, 5, 589-612 (1998)

Horn, M. and Vollandt, R. A survey of sample size formulas for pairwise and many-one multiple comparisons in the parametric, nonparametric and binomial case. Biometrical Journal, 42, 1, 27-44 (2000)

Hsu, J.C. Multiple Comparisons – Theory and methods. Chapman & Hall (1996)

Johnson, N.L. and Kotz, S. Distributions in statistics: continuous multivariate distributions. Wiley, New York. (1972)

Jugdutt, B.I. Effects of nitroglycerin and ibuprofen on left ventricular topography and rupture threshold during healing after myocardial infarction in the dog. Canadian Journal of Physiology and Pharmacology, 66, 385-395 (1988)

Kieser, M. A confirmatory strategy for therapeutic equivalence trials. International Journal of Clinical Pharmacology and Therapeutics, 33, 7, 388-390 (1995)

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

Liu, W. Some results on step-up tests for comparing treatments with a control in unbalanced one-way layouts. Biometrics, 53, 1508-1512 (1997)

Liu, W. Control of directional errors with step-up multiple tests. Statistics & Probability Letters, 31, 239-242 (1997)

Ludbrook, J. and Dudley, H. Why permutation tests are superior to t and F tests in biomedical research. The American Statistician, 52, 127-132 (1998)

Mann, H.B. and Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50-60 (1947)

Marcus, R., Peritz, E., Gabriel, K.R. On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63, 655-660 (1976)

Morales-Ramírez, P. and García-Rodríguez, M.C. In vivo effect of chlorophyllin on γ-ray-induced sister chromatid exchange in murine bone marrow cells. Mutation Research, 320, 329-334 (1994)

Munzel, U. and Hothorn, L.A. A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. Biometrical Journal, 43, 5, 553-569 (2001)

Nichols, M.B., Maickel, R.P. and Yim, G.K.W. Brain catecholamine alterations accompanying development of anorexia in rats bearing the Walker 256 carcinoma. Life Sciences, 36, 2223-2231 (1985)

Naik, U.D. Some selection rules for comparing p processes with a standard. Communications in Statistics, Part A – Theory and Methods, 4, 519-535 (1975)

Oude Voshaar, J.H. (k-1) mean significance levels of nonparametric multiple comparisons procedures. Annals of Statistics, 8, 75-86 (1980)

Overall J.E. A comment concerning one-sided tests of significance in new drug applications. Journal of Biopharmaceutical Statistics, 1, 157-160 (1991)

Pan, Z. and Kupper, L.L. Sample size determination for multiple comparison studies treating confidence interval width as random. Statistics in Medicine, 18, 1475-1488 (1999)

Peace, K.E. One-sided or two-sided p values: which most appropriate address the question of drug efficacy? Journal of Biopharmaceutical Statistics, 1, 133-138 (1991)

Points to consider on switching between superiority and non-inferiority. Committee for proprietary medicinal products, London, July 27 (2000)

Ramsey, P.H. Power differences between pairwise multiple comparisons. Journal of the American Statistical Association, 73, 479-485 (1978)

Roy, S.N. On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics, 24, 220-238 (1953)

Roy, S.N. and Bose, R.C. Simultaneous confidence interval estimation. Annals of Mathematical Statistics, 24, 513-536 (1953)

Ruymgaart, F.H. A unified approach to the asymptotic distribution theory of certain midrank statistics. In J.P. Raoult (ed.) Statistique non Parametrique Asymptotique. Lecture notes on Mathematics, 821. Springer, Berlin, 1-18 (1980)

SAS/STAT® Software, 1996: Changes and Enhancements through Release 6.11. SAS Institute Inc., Cary, NC, USA.

Sasabuchi, S. 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 (1988)

Scheffé, H.A. method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104 (1953)

Senn, S. Statistical issues in drug development. John Wiley & Sons, Chichester (1997)

Shaffer, J.P. Control of directional errors with stagewise multiple test procedures. Annals of Statistics, 8, 1342-1347 (1980)

Shaffer, J.P. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831 (1986)

Somerville, P.N. and Bretz, F. Obtaining critical values for simultaneous confidence intervals and multiple testing. Biometrical Journal, 43, 5, 657-663 (2001)

Spurrier, J.D. and Nizam, A. Sample size allocation for simultaneous inference in comparison with control experiments. Journal of the American Statistical Association, 85, 181-186 (1990)

Steel, R.D.G. A multiple comparison rank test: treatments versus control. Biometrics, 15, 560-572 (1959)

Steel, R.D.G. and Torrie, J.H. Principles and procedures of statistics – a biometrical approach, 2nd edition. New York, McGraw-Hill (1980)

Stefansson, G., Kim, W.C., Hsu J.C. On confidence sets in multiple comparisons. Statistical decision theory and related topics, IV, 2, 89-104 (1988)

Tong, Y.L. The multivariate normal distribution. Springer, New York. (1990)

Trevisan M, De Santo N, Laurenzi M, Di Muro M, De Chiara F, Latte M, Franzese A, Iacone R, Capodicasa G, Giordano C. Intracellular ion metabolism in erythrocytes and uraemia: the effect of different dialysis treatments. Clinical Science, 71, 545-552 (1986)

Tukey, J.W. The problem of multiple comparisons. Mimegraphed monograph (1953)

Tukey, J.W. Some thoughts on clinical trials, especially problems of multiplicity. Science, 198, 679-684 (1977)

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

Westfall, P.H., Tobias R.D., Rom, D., et.al. Multiple comparisons and multiple tests using the SAS System. Cary, NC: SAS Institute Inc. (1999)

Westfall, P.H. and Wolfinger, R.D. Closed multiple testing procedures and PROC MULTTEST, Observations, 23 (2000) (see: www.sas.com/service/library/periodicals/obs/obswww23.)

Wong, Y.W. and Cheung, S.H. Simultaneous pairwise multiple comparisons in a two-way analysis of covariance model. Journal of Applied Statistics, 27, 3, 281-291 (2000)

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

Appendix 1: Multivariate normal and multivariate t-distribution

The multivariate normal distribution and the multivariate t-distribution play an important role in many statistical applications and in many multiple comparisons procedures. Also the test statistics considered in this thesis are based on these multivariate distributions.
The multivariate normal distribution has been given a lot of attention in the literature. In contrast to the multivariate normal distribution, the multivariate t-disitribution has been given much less attention in the literature. See Johnson and Kotz (1972) and also Tong (1990) for a detailed discussion of the multivariate normal distribution and for details of the multivariate t-distribution.

This appendix describes the definition and some basic properties of both these multivariate distributions. In addition it shows the relationship between the multivariate normal and multivariate t-distribution.

Multivariate Normal distribution

Definition:
Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)'$ denotes a random vector of dimension n with independent and identically distributed (i.i.d.) components Z_i with $Z_i \sim N(0,1)$ i = 1, …, n.
If a random vector $\mathbf{X}$ of dimension k can be expressed as $\mathbf{X} = \boldsymbol{\mu} + \mathbf{CZ}$ where $\boldsymbol{\mu}$ is a k-vector, $\mathbf{C}$ a (k x n) matrix with rank $n \leq k$ and $\mathbf{CC}' = \mathbf{S}$, then $\mathbf{X}$ is said to follow a k-variate normal distribution which will be denoted as $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \mathbf{S})$.

By this definition, $\mathbf{Z}$ is said to follow a standard normal distribution of dimension n, which is denoted as $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, where $\mathbf{I}_n$ is the n-dimensional identity matrix.

The k-dimensional random variable $\mathbf{X}$ is said to have a non-singular multivariate normal distribution if k = n and $|\mathbf{S}| > 0$. Otherwise, $|\mathbf{S}| = 0$ and $\mathbf{X}$ is said to follow a singular k-variate normal distribution.

Some basic properties of the multivariate normal distribution are:

i) $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$ if and only if its characteristic function is given by

$$y_{\mathbf{x}}(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{x}}) = e^{i\mathbf{t}'\boldsymbol{\mu}-\frac{1}{2}\mathbf{tSt}}, \ \mathbf{t} \in \mathrm{R}^k \text{ where } i = \sqrt{-1}$$

Thus the distribution of $N_k(\boldsymbol{\mu},\mathbf{S})$ is uniquely determined by $\boldsymbol{\mu}$ and $\mathbf{S}$.

ii) Assume $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$ and $|\mathbf{S}| \neq 0$, then the density function of $\mathbf{X}$ is given by

$$f_k(\mathbf{x}) = \frac{1}{(2p)^{k/2}\sqrt{|\mathbf{S}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

iii) Assume $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$, then as expected

$E(\mathbf{X}) = \boldsymbol{\mu}$ and $Cov(\mathbf{X}) = \mathbf{S}$

iv) Assume $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$ and $\mathbf{Y} = \mathbf{?} + \mathbf{BX}$, where $\mathbf{?}$ is a *l*-vector and $\mathbf{B}$ is a (*l* x *k*) matrix, then

$\mathbf{Y} \sim N_l(\mathbf{?} + \mathbf{B}\boldsymbol{\mu}, \mathbf{BSB}')$

Thus a linear transformation of a multivariate normal distributed random variable remains multivariate normal distributed.

v) Assume $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$ and $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}$ and $\mathbf{S} = \begin{pmatrix}\mathbf{S}_{11} & \mathbf{S}_{12}\\ \mathbf{S}_{21} & \mathbf{S}_{22}\end{pmatrix}$, where $\mathbf{X}_1$ and $\boldsymbol{\mu}_1$ are *l*-vectors and $\mathbf{S}_{11}$ is a (*l* x *l*) matrix with *l* < *k*, then

$\mathbf{X}_1 \sim N_l(\boldsymbol{\mu}_1, \mathbf{S}_{11})$ and $\mathbf{X}_2 \sim N_{n-l}(\boldsymbol{\mu}_2, \mathbf{S}_{22})$

Thus the marginal distribution of a multivariate normal distribution is again a multivariate normal distribution.

Notation:

Assume $\mathbf{X} \sim N_k(\boldsymbol{\mu},\mathbf{S})$ with $|\mathbf{S}| > 0$ and density function $f_k(\mathbf{x})$, then the cumulative distribution function of this multivariate normal distribution is denoted as

$$\Phi_k(\mathbf{a},\mathbf{b};\boldsymbol{\mu},\mathbf{S}) = P_{\boldsymbol{\mu},\mathbf{S}}(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b}) = P_{\boldsymbol{\mu},\mathbf{S}}(a_1 \leq X_1 \leq b_1, \ldots, a_k \leq X_k \leq b_k) =$$

$$= \int_{a_1}^{b_1} \ldots \int_{a_k}^{b_k} f_k(\mathbf{x})dX_1 \ldots dX_k$$

The a (equi-)percentage point of this distribution is denoted as $x_{k,\boldsymbol{\mu},\mathbf{S};a}$, i.e. $\Phi_k\left(-\infty, ?_{k,\boldsymbol{\mu},\mathbf{S};a}; \boldsymbol{\mu}, \mathbf{S}\right) = a$.

Multivariate t-distribution

Definition:

Let $\mathbf{X} = (X_1, X_2, ..., X_k)' \sim N(\boldsymbol{\mu}, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix of $\mathbf{X}$, and let U be a univariate random variable that is c_n^2 distributed independently of the X_i's. Then $\mathbf{T} = (T_1, T_2, ..., T_k)' = \left(\frac{X_1}{\sqrt{U/n}}, \frac{X_2}{\sqrt{U/n}}, ..., \frac{X_k}{\sqrt{U/n}}\right)'$ has a *k*-variate t-distribution with n degrees of freedom and associated correlation matrix $\mathbf{R}$.

In case $\boldsymbol{\mu} = \mathbf{0}$ the distribution is called a central *k*-variate t-distribution and the notation is $\mathbf{T} \sim t_k(n, \mathbf{R})$. Otherwise the distribution is called a non-central *k*-variate t-distribution with non-centrality parameter $\boldsymbol{\mu}$, which notation is $\mathbf{T} \sim t_{k,\boldsymbol{\mu}}(n, \mathbf{R})$.

Two basic properties of the multivariate t-distribution are:

i) Assume $\mathbf{T} \sim t_k(n, \mathbf{R})$ and $|\mathbf{R}| > 0$, then the density function of $\mathbf{T}$ is given by

$$g_k(\mathbf{x}) = \frac{\Gamma\left(\frac{k+n}{2}\right)}{(np)^{k/2}\Gamma\left(\frac{n}{2}\right)\sqrt{|\mathbf{R}|}}\left(1 + \frac{1}{n}\mathbf{t}'\mathbf{R}^{-1}\mathbf{t}\right)^{-\frac{k+n}{2}}, \ \mathbf{t} \in \mathrm{R}^{\mathrm{k}}$$

where $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ denotes the gamma function.

This was independently derived by Dunnett and Sobel (1954) and Cornish (1954).

ii) Assume $\mathbf{T} \sim t_{k,\boldsymbol{\mu}}(n, \mathbf{R})$, then

$$E(\mathbf{T}) = \boldsymbol{\mu} \text{ for } n > 1 \text{ and } Corr(\mathbf{T}) = \mathbf{R} \text{ for } n > 2.$$

Notation:

Assume $\mathbf{T} \sim t_k(n,\mathbf{R})$ and $|\mathbf{R}| > 0$ and density function $g_k(\mathbf{x})$, then the cumulative distribution function of this multivariate t-distribution is denoted as

$$T_k(\mathbf{a},\mathbf{b};n,\mathbf{R}) = P_{n,\mathbf{R}}(\mathbf{a} \le \mathbf{T} \le \mathbf{b}) = P_{n,\mathbf{R}}(a_1 \le T_1 \le b_1,\ldots,a_k \le T_k \le b_k) =$$

$$= \int_{a_1}^{b_1}\ldots\int_{a_k}^{b_k} g_k(\mathbf{t})d\mathbf{t} = \frac{\Gamma\left(\frac{k+n}{2}\right)}{(np)^{k/2}\Gamma\left(\frac{n}{2}\right)\sqrt{|\mathbf{R}|}}\int_{a_1}^{b_1}\ldots\int_{a_k}^{b_k}\left(1+\frac{1}{n}\mathbf{t}'\mathbf{R}^{-1}\mathbf{t}\right)^{-\frac{k+n}{2}} d\mathbf{t}$$

The a (equi-)percentage point of the distribution of $\mathbf{T}$ is denoted as $t_{k,n,\mathbf{R};a}$, i.e. $T_k\left(-\infty,\mathbf{t}_{k,n,\mathbf{R};a};n,\mathbf{R}\right) = a$.

The two-sided a (equi-)percentage point of the distribution of $\mathbf{T}$ is denoted as $|t|_{k,n,\mathbf{R};a}$, i.e. $T_k\left(-|\mathbf{t}|_{k,n,\mathbf{R};a},|\mathbf{t}|_{k,n,\mathbf{R};a};n,\mathbf{R}\right) = a$.

Relationships between multivariate normal and multivariate t-distribution

Dunnett (1955) showed that a general k-variate t-distribution with n degrees of freedom and associated correlation matrix $\mathbf{R}$ can be transformed into a single integral over a k-variate standard normal distribution with the same matrix $\mathbf{R}$ as covariance matrix.

Relationship 1:

Let $T_k(\mathbf{a},\mathbf{b};n,\mathbf{R})$ and $\Phi_k(\mathbf{a},\mathbf{b};\boldsymbol{\mu},\mathbf{S})$ be the cumulative density functions of the k-variate t-distribution and k-variate normal distribution, respectively. In addition, let $h_n(x) = \frac{n^{n/2}e^{-xn/2}x^{n/2-1}}{\Gamma(n/2)2^{n/2}}$ be the density function of a c_n^2/n distributed random variable.

Then $T_k(\mathbf{a},\mathbf{b};n,\mathbf{R}) = \int_0^\infty \Phi_k(\mathbf{a}\sqrt{x},\mathbf{b}\sqrt{x};\mathbf{0},\mathbf{R})h_n(x)dx =$

$$= \frac{n^{n/2}}{\Gamma(n/2)2^{n/2}}\int_0^\infty e^{-xn/2}x^{n/2-1}\Phi_k(\mathbf{a}\sqrt{x},\mathbf{b}\sqrt{x};\mathbf{0},\mathbf{R})dx$$

In particular, assume that $\mathbf{T} \sim t_{k,\boldsymbol{\mu}}(n,\mathbf{R})$ and $\mathbf{Z}$ is a standardized k-variate normal distribution random variable with correlation matrix $\mathbf{R}$ and independently U is a c_n^2 /n distributed random variable, then $P\left(\mathbf{T} < \mathbf{b}\right) = P\left(\frac{\mathbf{Z}+\boldsymbol{\mu}}{\sqrt{U}} < \mathbf{b}\right) = P\left(\mathbf{Z} < \mathbf{b}\sqrt{U} - \boldsymbol{\mu}\right)$.

Relationship 2:

Let $g_k(\mathbf{t};n\,,\mathbf{R})$ and $f_k(\mathbf{t};\mathbf{0},\mathbf{R})$ be the density functions of a multivariate t-distribution and multivariate normal distribution, respectively. Then

$$\lim_{n\to\infty} g_k(\mathbf{t};n,\mathbf{R}) = f_k(\mathbf{t};\mathbf{0},\mathbf{R}) \quad \forall \mathbf{t} \in \mathrm{R}^{\mathrm{k}}$$

Thus this relationship shows that the multivariate t-distribution converges to the multivariate normal distribution for increasing degrees of freedom, like it holds true for the univariate situation.

Product correlation structure

A correlation matrix $\mathbf{R} = \left\{ r_{ij} \right\}$ is said to satisfy the product structure condition if $r_{ij} = l_i l_j \;\; \forall i \neq j$ with $l_i \in (-1,1)$.

Let $\mathbf{X} = (X_1, X_2,..., X_k)'$ have a k-variate normal distribution with zero mean vector, unit variances and correlation matrix $\mathbf{R} = \left\{ r_{ij} \right\}$ which satisfies the product structure condition. Then the X_i's can be represented by $X_i = \sqrt{1-l_i^2}Y_i - l_i Y_0 \quad i = 1,...,k$ where $Y_0, Y_1,..., Y_k$ are i.i.d. $N(0,1)$ random variables.

Under this condition, the calculation of the probability of the cumulative density function of the *k*-variate normal distribution does not involve a *k*-variate integral but can be expressed by univariate integrals as follows:

$$\Phi_k(\mathbf{a},\mathbf{b};\mathbf{0},\mathbf{R}) = \int_{-\infty}^{\infty} \prod_{i=1}^{k} \left\{ \Phi\left(\frac{l_i y + b_i}{\sqrt{1-l_i^2}} \right) - \Phi\left(\frac{l_i y + a_i}{\sqrt{1-l_i^2}} \right) \right\} d\Phi(y)$$

where $\Phi(y)$ is the cumulative density function of the univariate standard normal distribution. See also Dunnett and Sobel (1955) or Curnow and Dunnett (1962).

Proof:

$$\Phi_k(\mathbf{a},\mathbf{b};\mathbf{0},\mathbf{R}) = P(\mathbf{a} \le \mathbf{X} \le \mathbf{b}) = P(a_i \le X_i \le b_i;\ i = 1,\dots,k) =$$

$$= P(a_i \le \sqrt{1-l_i^2}Y_i - l_iY_0 \le b_i;\ i = 1,\dots,k) =$$

$$= \int_{-\infty}^{\infty} P\left(a_i \le \sqrt{1-l_i^2}Y_i - l_iY_0 \le b_i;\ i = 1,\dots,k \text{ and } Y_0 = y\right) d\Phi(y) =$$

$$= \int_{-\infty}^{\infty} P\left(l_iy + a_i \le \sqrt{1-l_i^2}Y_i \le l_iy + b_i;\ i = 1,\dots,k\right) d\Phi(y) =$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{k} \left\{ \Phi\left(\frac{l_iy + b_i}{\sqrt{1-l_i^2}} \right) - \Phi\left(\frac{l_iy + a_i}{\sqrt{1-l_i^2}} \right) \right\} d\Phi(y) .$$

Similarly the cumulative density function of the k-variate t-distribution can be expressed by univariate integrals:

$$T_k(\mathbf{a},\mathbf{b};n,\mathbf{R}) = \int_0^{\infty} \Phi_k(\mathbf{a}\sqrt{x},\mathbf{b}\sqrt{x};\mathbf{0},\mathbf{R})h_n(x)dx =$$

$$= \int_0^{\infty} \left[\int_{-\infty}^{\infty} \prod_{i=1}^{k} \left\{ \Phi\left(\frac{l_iy + b_i\sqrt{x}}{\sqrt{1-l_i^2}} \right) - \Phi\left(\frac{l_iy + a_i\sqrt{x}}{\sqrt{1-l_i^2}} \right) \right\} d\Phi(y) \right] h_n(x)dx$$

where $h_n(x) = \dfrac{n^{n/2}e^{-xn/2}x^{n/2-1}}{\Gamma(n/2)2^{n/2}}$ is the density function of a c_n^2/n distributed random variable.

Appendix 2: SAS/STAT function PROBMC

The PROBMC function provided with the SAS/STAT software enables to compute probabilities and quantiles from the one-sided and two-sided Dunnett distributions with finite and infinite degreed of freedom for the variance estimate.

This section describes the PROBMC function restricted to the many-to-one test situation. A full description of the PROBMC function can be found in the SAS manual SAS/STAT Software (1996).

Syntax

value = PROBMC(*string*, *q*, *prob*, *df*, *nparms*, <*parameters*>);

Return value:

value = either the probability or the quantile from the distribution

Input arguments:

string = a character string identifying the distribution, which is either ‘DUNNETT1’ or ‘DUNNETT2’.

q = the quantile from the distribution. Only one of the parameters *q* or *prob* should be specified; the other should be set to missing. (q > 0 in case of ‘DUNNETT2’)

prob = the left probability of the distribution. Only one of the parameters *q* or *prob* should be specified; the other should be set to missing.

df = the degrees of freedom. A missing value is interpreted as an infinite value.

nparms = the number of active treatment groups.

parameters = the set of *nparms* parameters that must be specified to handle the unequal case. If *parameters* is not specified, equal parameters are assumed.

The precision of the computed probability *prob* will be $O(10^{-8})$ (absolute error), and the precision of the computed quantile *q* will be $O(10^{-5})$.

Formulas and parameters

Let $h_n(x) = \dfrac{n^{n/2}e^{-xn/2}x^{n/2-1}}{\Gamma(n/2)2^{n/2}}$ be the density function of a c_n^2 /n distributed random variable.

Then the following expressions relate the probability, *prob*, and the quantile, q, for different situations.

- Unequal case with finite degrees of freedom:

$$prob = \int_0^\infty \int_{-\infty}^\infty f(y) \prod_{i=1}^k \Phi\left(\frac{l_i y + q\sqrt{x}}{\sqrt{1-l_i^2}} \right) h_n(x)dydx \qquad \text{one-sided case}$$

$$prob = \int_0^\infty \int_{-\infty}^\infty f(y) \prod_{i=1}^k \left[\Phi\left(\frac{l_i y - q\sqrt{x}}{\sqrt{1-l_i^2}} \right) - \Phi\left(\frac{l_i y + q\sqrt{x}}{\sqrt{1-l_i^2}} \right) \right] h_n(x)dydx \qquad \text{two-sided case}$$

In this case, the *parameters* are $l_1,..., l_k$, the value of *nparms* is set to k, and the value of *df* is set to n .

- Equal case with finite degrees of freedom:

$$prob = \int_0^\infty \int_{-\infty}^\infty f(y) \left[\Phi\left(y + \sqrt{2}q\sqrt{x} \right) \right]^k h_n(x)dydx \qquad \text{one-sided case}$$

$$prob = \int_0^\infty \int_{-\infty}^\infty f(y) \left[\Phi\left(y + \sqrt{2}q\sqrt{x} \right) - \Phi\left(y - \sqrt{2}q\sqrt{x} \right) \right]^k h_n(x)dydx \qquad \text{two-sided case}$$

In this case, no *parameters* are passed, the value of *nparms* is set to k, and the value of *df* is set to n .

- Unequal case with infinite degrees of freedom:

$$prob = \int_{-\infty}^\infty f(y) \prod_{i=1}^k \Phi\left(\frac{l_i y + q}{\sqrt{1-l_i^2}} \right) dy \qquad \text{one-sided case}$$

$$prob = \int_{-\infty}^\infty f(y) \prod_{i=1}^k \left[\Phi\left(\frac{l_i y + q}{\sqrt{1-l_i^2}} \right) - \Phi\left(\frac{l_i y - q}{\sqrt{1-l_i^2}} \right) \right] dy \qquad \text{two-sided case}$$

In this case, the *parameters* are $l_1,\ldots,l_k$, the value of *nparms* is set to k, and the value of *df* is set to missing.

- Equal case with infinite degrees of freedom:

$$prob = \int_{-\infty}^{\infty} f(y)\left[\Phi\left(y+\sqrt{2}q\right)\right]^k dy \qquad \text{one-sided case}$$

$$prob = \int_{-\infty}^{\infty} f(y)\left[\Phi\left(y+\sqrt{2}q\right)-\Phi\left(y-\sqrt{2}q\right)\right]^k dy \qquad \text{two-sided case}$$

In this case, no *parameters* are passed, the value of *nparms* is set to k, and the value of *df* is set to missing.

Appendix 3: SAS Program code

This appendix contains the SAS code of all procedures used throughout this thesis to analyze the example dataset introduced in Chapter 3.

The example dataset consist of the following data:

OBS	STRATA	TRT	Y
1	1	0	10.5212
2	1	0	10.8392
3	1	0	9.6872
4	1	0	10.6900
5	1	0	9.2314
6	1	0	9.3274
7	1	0	10.8205
8	1	0	11.8538
9	1	0	10.0951
10	1	0	9.8664
11	1	1	10.2495
12	1	1	10.9874
13	1	1	11.8561
14	1	1	10.9736
15	1	1	10.8699
16	1	1	11.6841
17	1	1	11.4768
18	1	2	13.8552
19	1	2	12.6919
20	1	2	12.3069
21	1	2	11.9455
22	1	2	11.4824
23	2	0	15.2332
24	2	0	13.8679
25	2	0	15.0877
26	2	0	14.7369
27	2	0	13.8194
28	2	0	13.4193
29	2	0	14.8510
30	2	0	14.4201
31	2	0	15.5445
32	2	0	15.3915
33	2	1	13.4587
34	2	1	15.4549
35	2	1	17.2838
36	2	1	15.4497
37	2	1	14.5990
38	2	1	15.0679
39	2	2	16.1797
40	2	2	16.6505
41	2	2	15.3983
42	2	2	15.9460
43	2	2	15.3376

The value *Strata* = 1 represents the males and *Strata* = 2 represents the females.

The value Trt = 0 represents the control treatment and Trt = 1, Trt = 2 represent the low dose and high dose respectively.

The data are assumed to be stored in the location `'c:\My SAS Files\...';`

Program ***CH3_12.SAS*** performs the analyses of Chapter 3, Sections 1 and 2

```
/*******************************************************/
/* Program calculating one-sided 'Dunnett' corrected     */
/* p-values and simultanous CI's                         */
/*                                                       */
/* Input:  dataset Example                               */
/* Output: dij  = estimate of treatment effect of        */
/*                treatment j in stratum i               */
/*         pval = one-sided adjusted p-value             */
/*         cval = critical values at alpha 5% level      */
/*         cij  = lower limit of one-sided 95% CI of     */
/*                treatment j in stratum i               */
/*******************************************************/

%GLOBAL _PRINT_;
%LET _PRINT_ = OFF;

OPTIONS NOBYLINE;

/****************************/
/* Input of example dataset */
/****************************/
LIBNAME DAT 'c:\My SAS Files\...';

DATA WORK.example;
SET DAT.example;
trts = 10 * strata + trt; /* Unique treatment code per stratum */
RUN;

PROC MEANS DATA = WORK.example NOPRINT;
CLASS trts;
VAR y;
OUTPUT OUT = WORK.means N = n MEAN = mean STD = std;
RUN;

TITLE 'Summary statistics';
PROC PRINT DATA = WORK.means;
RUN;

/*****************************/
/* Compute t-values and sigma */
/*****************************/
PROC MIXED DATA = WORK.example;
CLASS strata trt;
MODEL y = strata trt strata*trt;
ESTIMATE 'C1 plac - 1' strata 0 0 trt -1 1 0 strata*trt -1 1 0  0 0 0;
ESTIMATE 'C1 plac - 2' strata 0 0 trt -1 0 1 strata*trt -1 0 1  0 0 0;
ESTIMATE 'C2 plac - 1' strata 0 0 trt -1 1 0 strata*trt  0 0 0 -1 1 0;
ESTIMATE 'C2 plac - 2' strata 0 0 trt -1 0 1 strata*trt  0 0 0 -1 0 1;
LSMEANS strata*trt/ DIFFS;
MAKE 'DIFFS' OUT = WORK.diffs;
MAKE 'COVPARMS' OUT = WORK.cov;
RUN;

DATA WORK.diffs;
SET WORK.diffs;
IF strata = _strata;
IF trt = 0 and _trt > 0;
t = - _t_;
p = 1 - PROBT(t,_df_); /* one-sided unadjusted p-value */
```

```
RUN;

TITLE 'Statistics and unadjusted p-values';
PROC PRINT DATA = WORK.diffs;
RUN;

/******************************************/
/* Create variables needed within SAS/IML */
/******************************************/
DATA _NULL_;
SET WORK.cov;
CALL SYMPUT('sigma',SQRT(est));
RUN;

DATA _NULL_;
SET WORK.means;
IF trts = 10 THEN CALL SYMPUT('n10',n);
IF trts = 11 THEN CALL SYMPUT('n11',n);
IF trts = 12 THEN CALL SYMPUT('n12',n);
IF trts = 20 THEN CALL SYMPUT('n20',n);
IF trts = 21 THEN CALL SYMPUT('n21',n);
IF trts = 22 THEN CALL SYMPUT('n22',n);
IF trts = . THEN CALL SYMPUT('df',n-6);
RUN;

DATA _NULL_;
SET WORK.diffs;
IF strata = 1 AND _trt = 1 THEN DO;
  CALL SYMPUT('t11',t); CALL SYMPUT('d11',-_diff_); END;
IF strata = 1 AND _trt = 2 THEN DO;
  CALL SYMPUT('t12',t); CALL SYMPUT('d12',-_diff_); END;
IF strata = 2 AND _trt = 1 THEN DO;
  CALL SYMPUT('t21',t); CALL SYMPUT('d21',-_diff_); END;
IF strata = 2 AND _trt = 2 THEN DO;
  CALL SYMPUT('t22',t); CALL SYMPUT('d22',-_diff_); END;
RUN;

/*******************************************/
/* Compute adjusted p-values and sim. CI's */
/* using algorithm 3: PROBMC               */
/*******************************************/
PROC IML;
n10 = &n10; n11 = &n11; n12 = &n12;
n20 = &n20; n21 = &n21; n22 = &n22;

d11 = &d11; d12 = &d12; d21 = &d21; d22 = &d22;

df= &df;             /* sum overij (nij-1) */
lambda11 = SQRT(n11/(n10 + n11));
lambda12 = SQRT(n12/(n10 + n12));
lambda21 = SQRT(n21/(n20 + n21));
lambda22 = SQRT(n22/(n20 + n22));

d = df/2;
m = d**d / GAMMA(d);

/* Define integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda12,lambda21,lambda22);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT1",q,.,.,2,lambda11,lambda12);
  p2 = PROBMC("DUNNETT1",q,.,.,2,lambda21,lambda22);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = &t11;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value stratum 1 treatment 1' d11 t pval[FORMAT=7.5];

t = &t12;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value stratum 1 treatment 2' d12 t pval[FORMAT=7.5];

t = &t21;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value stratum 2 treatment 1' d21 t pval[FORMAT=7.5];

t = &t22;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value stratum 2 treatment 2' d22 t pval[FORMAT=7.5];


/* find critical values using bisection method */
alpha = 0.05;
c = 2;                          /* number of active treatments    */
r = 2;                          /* number of groups               */
q1 = TINV(1 - alpha,df);        /* start value: uncorr. t value   */
q2 = TINV(1 - alpha/(c*r),df);  /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Critical value' cval[FORMAT=7.5];

c11 = &d11 - cval * &sigma * SQRT(n11**-1 + n10**-1);
c12 = &d12 - cval * &sigma * SQRT(n12**-1 + n10**-1);
c21 = &d21 - cval * &sigma * SQRT(n21**-1 + n20**-1);
c22 = &d22 - cval * &sigma * SQRT(n22**-1 + n20**-1);

PRINT 'Lower limit of one-sided 95% CIs' c11 c12 c21 c22;

QUIT;
```

Program ***CH3_3.SAS*** performs the analyses of Chapter 3, Section 3

```
/*********************************************************/
/* Program calculating power                             */
/*                                                       */
/* Input:  parameters of dataset Example including       */
/*         critical value                                */
/*         program PROBMVT.SAS of Genz and Bretz         */
/* Output: powAll = All-pairs power                      */
/*         powAny = Any-pair power                       */
/*********************************************************/

/*****************************************************/
/* Include the SAS/IML program PROBMVT that computes */
/* probabilities of the multivariate t distribution  */
/* available on the homepage of Bretz:               */
/* http://www.bioinf.uni-hannover.de/~bretz          */
/*****************************************************/

/****************************/
/* Input parameters example */
/****************************/
s = SQRT(0.7);
b11 = (7 / (10+7))##(0.5);
b12 = (5 / (10+5))##(0.5);
b21 = (7 / (10+7))##(0.5);
b22 = (5 / (10+5))##(0.5);
rho1_12 = b11*b12;
rho2_12 = b21*b22;

/*************************/
/* S is the complete set */
/*************************/
n = 4;
nu = 38;
covar = (   1   ||rho1_12||   0   ||   0   )//
        (rho1_12||   1   ||   0   ||   0   )//
        (   0   ||   0   ||   1   ||rho2_12)//
        (   0   ||   0   ||rho2_12||   1   );

maxpts = 2000*n*n;
abseps = .0001;
releps = 0;

delta11 = 1.5 / (s*SQRT(1/10 + 1/7));
delta12 = 1.5 / (s*SQRT(1/10 + 1/5));
delta21 = 1.5 / (s*SQRT(1/10 + 1/7));
delta22 = 1.5 / (s*SQRT(1/10 + 1/5));
delta = delta11||delta12||delta21||delta22;

/***************************/
/* Compute All-pairs power */
/***************************/
lower = J(1,n,2.30336);   /* fill in the crit.value */
upper = J(1,n,15);
infin = J(1,n,1);         /* interval [lower,inf)   */

RUN MVN_DIST(n,nu,delta,lower,upper,infin,covar,maxpts,abseps,releps,error,
                                                        value,nevals,inform);
powAll = value;
PRINT 'All-pairs power';
PRINT powAll, n error nevals inform;

/***************************/
/* Compute Any-pair power */
/***************************/
lower = J(1,n,0);
upper = J(1,n,2.30336);   /* fill in the crit.value */
infin = J(1,n,0);         /* interval (-inf,upper]  */

RUN MVN_DIST(n,nu,delta,lower,upper,infin,covar,maxpts,abseps,releps,error,
                                                        value,nevals,inform);
powAny = 1 - value;
PRINT 'Any-pair power';
PRINT powAny, n error nevals inform;

/*************************************/
/* S is only the two highest dosages */
/*************************************/
n = 2;
nu = 38;

covar = (   1   ||   0   )//
        (   0   ||   1   );

maxpts = 2000*n*n*n;
abseps = .0001;
releps = 0;

delta12 = 2 / (s*SQRT(1/10 + 1/5));
delta22 = 2 / (s*SQRT(1/10 + 1/5));
delta = delta12||delta22;

/***************************/
/* Compute All-pairs power */
/***************************/
lower = J(1,n,2.30336);   /* fill in the crit.value */
upper = J(1,n,15);
infin = J(1,n,1);         /* interval [lower,inf)   */

RUN MVN_DIST(n,nu,delta,lower,upper,infin,covar,maxpts,abseps,releps,error,
                                                        value,nevals,inform);
powAll = value;
PRINT 'All-pairs power';
PRINT powAll, n error nevals inform;

/***************************/
/* Compute Any-pair power */
/***************************/
lower = J(1,n,0);
upper = J(1,n,2.30336);   /* fill in the crit.value */
infin = J(1,n,0);         /* interval (-inf,upper]  */

RUN MVN_DIST(n,nu,delta,lower,upper,infin,covar,maxpts,abseps,releps,error,
                                                        value,nevals,inform);
powAny = 1 - value;
PRINT 'Any-pair power';
PRINT powAny, n error nevals inform;

QUIT;
```

Program ***CH3_4.SAS*** performs the analyses of Chapter 3, Section 4

```
/**********************************************************/
/* Program calculating sample sizes                       */
/*                                                        */
/*                                                        */
/* Input:  parameters of dataset Example including        */
/*         critical value and relevant difference         */
/*         program PROBMVN.SAS of Genz and Bretz          */
/* Output: powAll = All-pairs power                       */
/*         (powAny = Any-pair power)                      */
/**********************************************************/

/**********************************************************/
/* Include the SAS/IML program PROBMVN that computes      */
/* probabilities of the multivariate normal distribution */
/* available on the homepage of Bretz:                    */
/* http://www.bioinf.uni-hannover.de/~bretz               */
/**********************************************************/

/****************************/
/* Input parameters example */
/****************************/
l = 1 / SQRT(2);
rho = l / (1+l);

/************************/
/* S is the complete set */
/************************/
n = 4;
covar = ( 1 ||rho|| 0 || 0 )//
        (rho|| 1 || 0 || 0 )//
        ( 0 || 0 || 1 ||rho)//
        ( 0 || 0 ||rho|| 1 );

maxpts = 2000*N*N*N;
abseps = .0001;
releps = 0;

dalpha = 2.215;                  /* crit.value */
delta = 1.5;         /* relevant difference    */

/*********************************************/
/* fill in (n1,n2) and compute All-pairs power */
/* change until power >= required value        */
/*********************************************/
n1 = 8;
n2 = 8;

s = SQRT(0.7);
b1 = (delta * SQRT(n1) / (s*SQRT(1.707107))) - dalpha;
b2 = (delta * SQRT(n2) / (s*SQRT(1.707107))) - dalpha;

lower = J(1,N,0);
upper = b1||b1||b2||b2;
infin = J(1,N,0);     /* interval (-inf,upper] */

RUN MVN_DIST(n,lower,upper,infin,covar,maxpts,abseps,releps,error,value,
                                                            nevals,inform);
powAll = value;
PRINT 'All-pairs power';
PRINT powAll, n error nevals inform;

QUIT;
```

Program ***CH3_5.SAS*** performs the analyses of Chapter 3, Section 5

```
/*********************************************************/
/* Program calculating one-sided adjusted p-values       */
/* applying step-down procedure                          */
/*                                                       */
/* Input:  parameters of dataset Example                 */
/* Output: pval = one-sided adjsuted pvalue              */
/*         cval = critical value at alpha 5% level       */
/*********************************************************/

/*********************************************/
/* Compute adjusted p-values and crit.values */
/* using algorithm 3: PROBMC                 */
/*********************************************/
PROC IML;
n10 = 10; n11 = 7; n12 = 5;
n20 = 10; n21 = 6; n22 = 5;

df = SUM(n10,n11,n12,n20,n21,n22) - 6;  /* sum overij (nij-1) */
lambda11 = SQRT(n11/(n10 + n11));
lambda12 = SQRT(n12/(n10 + n12));
lambda21 = SQRT(n21/(n20 + n21));
lambda22 = SQRT(n22/(n20 + n22));

d = df/2;
m = d**d / GAMMA(d);

/**********/
/* Step 1 */
/**********/
/*Define integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda12,lambda21,lambda22);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT1",q,.,.,2,lambda11,lambda12);
  p2 = PROBMC("DUNNETT1",q,.,.,2,lambda21,lambda22);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = 4.82028;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value step 1' t pval[FORMAT=7.5];

/* find critical values using bisection method */
alpha = 0.05;
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = TINV(1 - alpha,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(c*r),df);   /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Critical value step 1' cval[FORMAT=7.5];

/**********/
/* Step 2 */
/**********/
/*Redefine integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda21,lambda22);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT1",q,.,.,1,lambda11);
  p2 = PROBMC("DUNNETT1",q,.,.,2,lambda21,lambda22);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = 2.81947;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value step 2' t pval[FORMAT=7.5];

/* find critical values using bisection method */
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = TINV(1 - alpha,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(c*r),df);   /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Critical value step 2' cval[FORMAT=7.5];

/**********/
/* Step 3 */
/**********/
/*Redefine integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda21);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT1",q,.,.,1,lambda11);
  p2 = PROBMC("DUNNETT1",q,.,.,1,lambda21);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = 2.13874;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value step 3' t pval[FORMAT=7.5];

/* find critical values using bisection method */
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = TINV(1 - alpha,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(c*r),df);   /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Critical value step 3' cval[FORMAT=7.5];

/**********/
/* Step 4 */
/**********/
/*Redefine integrand */
START dunnett(u) GLOBAL(d,t,lambda21);
  q = t * SQRT(u);
  p2 = PROBMC("DUNNETT1",q,.,.,1,lambda21);
  g = u **(d-1) * EXP(-u * d);
  v = p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = 1.37519;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval = 1 - m * z;
PRINT 'P-value step 4' t pval[FORMAT=7.5];

/* find critical values using bisection method */
alpha = 0.05;
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = TINV(1 - alpha,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(c*r),df);   /* start value: Bonf. corr. value */
```

```
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Critical value step 4' cval[FORMAT=7.5];

QUIT;
```

Program ***CH4.SAS*** performs the analyses of Chapter 4

```
/*********************************************************/
/* Program calculating two-sided 'Dunnett' corrected     */
/* p-values and simultaneous CI's                        */
/*                                                       */
/* Input:  dataset Example                               */
/* Output: dij   = estimate of treatment effect of       */
/*                 treatment j in stratum i              */
/*         pval  = two-sided adjusted p-value            */
/*         cval  = critical values at alpha 5% level     */
/*         cij_l = lower limit of two-sided 95% CI of    */
/*                 treatment j in stratum i              */
/*         cij_u = upper limit of two-sided 95% CI of    */
/*                 treatment j in stratum i              */
/*********************************************************/

%GLOBAL _PRINT_;
%LET _PRINT_ = OFF;

OPTIONS NOBYLINE;

/****************************/
/* Input of example dataset */
/****************************/
LIBNAME DAT 'c:\My SAS Files\...';

DATA WORK.example;
SET DAT.example;
trts = 10 * strata + trt; /* Unique treatment code per stratum */
RUN;

/******************************/
/* Compute t-values and sigma */
/******************************/
PROC MIXED DATA = WORK.example;
CLASS strata trt;
MODEL y = strata trt strata*trt;
ESTIMATE 'C1 plac - 1' strata 0 0 trt -1 1 0 strata*trt -1 1 0  0 0 0;
ESTIMATE 'C1 plac - 2' strata 0 0 trt -1 0 1 strata*trt -1 0 1  0 0 0;
ESTIMATE 'C2 plac - 1' strata 0 0 trt -1 1 0 strata*trt  0 0 0 -1 1 0;
ESTIMATE 'C2 plac - 2' strata 0 0 trt -1 0 1 strata*trt  0 0 0 -1 0 1;
LSMEANS strata*trt/ DIFFS;
MAKE 'DIFFS' OUT = WORK.diffs;
MAKE 'COVPARMS' OUT = WORK.cov;
RUN;

DATA WORK.diffs;
SET WORK.diffs;
IF strata = _strata;
IF trt = 0 and _trt > 0;
t = - _t_;
RUN;

TITLE 'Statistics and Unadjusted p-values';
PROC PRINT DATA = WORK.diffs;
RUN;

/******************************************/
/* Create variables needed within SAS/IML */
/******************************************/
```

```
DATA _NULL_;
SET WORK.cov;
CALL SYMPUT('sigma',SQRT(est));
RUN;

PROC MEANS DATA = WORK.example NOPRINT;
CLASS trts;
VAR y;
OUTPUT OUT = WORK.means N = n;
RUN;

DATA _NULL_;
SET WORK.means;
IF trts = 10 THEN CALL SYMPUT('n10',n);
IF trts = 11 THEN CALL SYMPUT('n11',n);
IF trts = 12 THEN CALL SYMPUT('n12',n);
IF trts = 20 THEN CALL SYMPUT('n20',n);
IF trts = 21 THEN CALL SYMPUT('n21',n);
IF trts = 22 THEN CALL SYMPUT('n22',n);
IF trts = . THEN CALL SYMPUT('df',n-6);
RUN;

DATA _NULL_;
SET WORK.diffs;
IF strata = 1 AND _trt = 1 THEN DO;
  CALL SYMPUT('t11',t); CALL SYMPUT('d11',-_diff_); END;
IF strata = 1 AND _trt = 2 THEN DO;
  CALL SYMPUT('t12',t); CALL SYMPUT('d12',-_diff_); END;
IF strata = 2 AND _trt = 1 THEN DO;
  CALL SYMPUT('t21',t); CALL SYMPUT('d21',-_diff_); END;
IF strata = 2 AND _trt = 2 THEN DO;
  CALL SYMPUT('t22',t); CALL SYMPUT('d22',-_diff_); END;
RUN;

/*******************************************/
/* Compute adjusted p-values and sim. CI's */
/* using algorithm 3: PROBMC               */
/*******************************************/
PROC IML;
n10 = &n10; n11 = &n11; n12 = &n12;
n20 = &n20; n21 = &n21; n22 = &n22;

d11 = &d11; d12 = &d12; d21 = &d21; d22 = &d22;

df= &df;             /* sum overij (nij-1) */
lambda11 = SQRT(n11/(n10 + n11));
lambda12 = SQRT(n12/(n10 + n12));
lambda21 = SQRT(n21/(n20 + n21));
lambda22 = SQRT(n22/(n20 + n22));

d = df/2;
m = d**d / GAMMA(d);

/*Define integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda12,lambda21,lambda22);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT2",q,.,.,2,lambda11,lambda12);
  p2 = PROBMC("DUNNETT2",q,.,.,2,lambda21,lambda22);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
```

```
FINISH;

int = {0 .P};

t = &t11;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
PRINT 'P-value stratum 1 treatment 1' d11 t pval2[FORMAT=7.5];

t = &t12;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
PRINT 'P-value stratum 1 treatment 2' d12 t pval2[FORMAT=7.5];

t = &t21;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
PRINT 'P-value stratum 2 treatment 1' d21 t pval2[FORMAT=7.5];

t = &t22;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
PRINT 'P-value stratum 2 treatment 2' d22 t pval2[FORMAT=7.5];

/* find critical values using bisection method */
alpha = 0.05;
c = 2;                             /* number of active treatments */
r = 2;                             /* number of groups */
q1 = TINV(1 - alpha/2,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(2*c*r),df);   /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Two-sided critical value' cval[FORMAT=7.5];

c11_l = &d11 - cval * &sigma * SQRT(n11**-1 + n10**-1);
c11_u = &d11 + cval * &sigma * SQRT(n11**-1 + n10**-1);
c12_l = &d12 - cval * &sigma * SQRT(n12**-1 + n10**-1);
c12_u = &d12 + cval * &sigma * SQRT(n12**-1 + n10**-1);
c21_l = &d21 - cval * &sigma * SQRT(n21**-1 + n20**-1);
c21_u = &d21 + cval * &sigma * SQRT(n21**-1 + n20**-1);
c22_l = &d22 - cval * &sigma * SQRT(n22**-1 + n20**-1);
c22_u = &d22 + cval * &sigma * SQRT(n22**-1 + n20**-1);

PRINT 'Lower and upper limits of two-sided 95%CIs',
    c11_l c11_u, c12_l c12_u, c21_l c21_u, c22_l c22_u;

QUIT;
```

Program ***CH6.SAS*** performs the analyses of Chapter 6

```
/*********************************************************/
/* Program calculating two-sided adjusted p-values and   */
/* critical values for ratio's                           */
/*                                                       */
/* Input:  dataset Example                               */
/* Output: dij = estimate of treatment effect of         */
/*               treatment j in group i                  */
/*         cal = two-sided critical values at            */
/*               alpha 5% level                          */
/*********************************************************/

%GLOBAL _PRINT_;
%LET _PRINT_ = OFF;

OPTIONS NOBYLINE;

/****************************/
/* Input of example dataset */
/****************************/
LIBNAME DAT 'c:\My SAS Files\...';

DATA WORK.example;
SET DAT.example;
trts = 10 * strata + trt; /* Unique treatment code per stratum */
RUN;

/******************************************/
/* Create variables needed within SAS/IML */
/******************************************/
PROC MIXED DATA = WORK.example;
CLASS strata trt;
MODEL y = strata trt strata*trt;
MAKE 'COVPARMS' OUT = WORK.cov;
RUN;

DATA _NULL_;
SET WORK.cov;
CALL SYMPUT('sigma',SQRT(est));
RUN;

PROC MEANS DATA = WORK.example NOPRINT;
CLASS trts;
VAR y;
OUTPUT OUT = WORK.means MEAN = MEAN N = n STD =std;
RUN;

DATA _NULL_;
SET WORK.means;
IF trts = 10 THEN DO;
  CALL SYMPUT('x10',mean); CALL SYMPUT('n10',n); END;
IF trts = 11 THEN DO;
  CALL SYMPUT('x11',mean); CALL SYMPUT('n11',n); END;
IF trts = 12 THEN DO;
  CALL SYMPUT('x12',mean); CALL SYMPUT('n12',n); END;
IF trts = 20 THEN DO;
  CALL SYMPUT('x20',mean); CALL SYMPUT('n20',n); END;
IF trts = 21 THEN DO;
  CALL SYMPUT('x21',mean); CALL SYMPUT('n21',n); END;
IF trts = 22 THEN DO;
```

```
  CALL SYMPUT('x22',mean); CALL SYMPUT('n22',n); END;
IF trts = . THEN CALL SYMPUT('df',n-6);
RUN;

/*********************************************/
/* Compute adjusted p-values and crit. value */
/* using algorithm 3: PROBMC                 */
/*********************************************/
PROC IML;
x10 = &x10; x11 = &x11; x12 = &x12;
x20 = &x20; x21 = &x21; x22 = &x22;

n10 = &n10; n11 = &n11; n12 = &n12;
n20 = &n20; n21 = &n21; n22 = &n22;

df = &df;             /* sum overij (nij-1) */
s  = &sigma;

t11 = (x11 - x10) / (s * SQRT(1/n11 + 1/n10));
t12 = (x12 - x10) / (s * SQRT(1/n12 + 1/n10));
t21 = (x21 - x20) / (s * SQRT(1/n21 + 1/n20));
t22 = (x22 - x20) / (s * SQRT(1/n22 + 1/n20));


lambda11 = 1 / SQRT((n10/n11) + 1);
lambda12 = 1 / SQRT((n10/n12) + 1);
lambda21 = 1 / SQRT((n20/n21) + 1);
lambda22 = 1 / SQRT((n20/n22) + 1);

d = df/2;
m = d**d / GAMMA(d);

/*Define integrand */
START dunnett(u) GLOBAL(d,t,lambda11,lambda12,lambda21,lambda22);
  q = t * SQRT(u);
  p1 = PROBMC("DUNNETT2",q,.,.,2,lambda11,lambda12);
  p2 = PROBMC("DUNNETT2",q,.,.,2,lambda21,lambda22);
  g = u **(d-1) * EXP(-u * d);
  v = p1 * p2 * g;
RETURN (v);
FINISH;

int = {0 .P};

t = t11;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
r11 = x11 / x10;
PRINT 'P-value stratum 1 treatment 1' r11 t11 pval2[FORMAT=7.5];

t = t12;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
r12 = x12 / x10;
PRINT 'P-value stratum 1 treatment 2' r12 t12 pval2[FORMAT=7.5];

t = t21;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
r21 = x21 / x20;
PRINT 'P-value stratum 2 treatment 1' r21 t21 pval2[FORMAT=7.5];
```

```
t = t22;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
pval2 = 1 - m * z;
r22 = x22 / x20;
PRINT 'P-value stratum 2 treatment 2' r22 t22 pval2[FORMAT=7.5];


/* find critical values using bisection method */
alpha = 0.05;
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = TINV(1 - alpha/2,df);         /* start value: uncorr. t value */
q2 = TINV(1 - alpha/(2*c*r),df);   /* start value: Bonf. corr. value */
t = q2;
CALL QUAD(z,"dunnett",int) EPS = 1E-10;
crit = 1 - m * z;

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  t = qm;
  CALL QUAD(z,"dunnett",int) EPS = 1E-10;
  crit = 1 - m * z;
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;

cval = t;
PRINT 'Two-sided critical value' cval[FORMAT=7.5];

QUIT;
```

Program ***CH7.SAS*** performs the analyses of Chapter 7

```
/*********************************************************/
/* Program calculating asymptotic two-sided adjusted     */
/* p-values and simulteneous CI's using nonparametrical  */
/* procedure based on distribution functions             */
/*                                                       */
/* Input:  dataset Example                               */
/* Output: dij  = estimate of treatment effect of        */
/*                treatment j in stratum i               */
/*         pval = two-sided adjusted p-value             */
/*         cval = critical values at alpha 5% level      */
/*         c_l  = lower limit of two-sided 95% CI of     */
/*                treatment j in stratum i               */
/*         c_u  = upper limit of two-sided 95% CI of     */
/*                treatment j in stratum i               */
/*********************************************************/

%GLOBAL _PRINT_;
%LET _PRINT_ = OFF;

OPTIONS NOBYLINE;

/****************************/
/* Input of example dataset */
/****************************/
LIBNAME DAT 'c:\My SAS Files\...';

DATA WORK.example;
SET DAT.example;
trts = 10 * strata + trt; /* Unique treatment code per stratum */
RUN;

/********************/
/* Pairwise ranking */
/********************/
DATA WORK.ex2;
LENGTH j 8;
SET WORK.ex (IN = in11 WHERE = (trts IN (10 11)))
    WORK.ex (IN = in12 WHERE = (trts IN (10 12)))
    WORK.ex (IN = in21 WHERE = (trts IN (20 21)))
    WORK.ex (IN = in22 WHERE = (trts IN (20 22)));
IF in11 THEN j = 11;
IF in12 THEN j = 12;
IF in21 THEN j = 21;
IF in22 THEN j = 22;
RUN;

PROC RANK DATA = WORK.ex2 OUT = WORK.rank TIES = MEAN;
BY j;
RANKS rank;
VAR y;
RUN;

/**************/
/* Mean ranks */
/**************/
PROC MEANS DATA = WORK.rank NOPRINT NWAY;
CLASS j trt;
VAR rank;
OUTPUT OUT = WORK.rmean (DROP = _TYPE_ _FREQ_) N = n MEAN = rmean;
```

```
RUN;

/* n_act: obs. on active treatment
   rmean1: mean of midranks Xi0k in sample Xi0k and Xijk */
DATA WORK.n_act(KEEP = j trt n)
     WORK.rmean1 (DROP = trt RENAME = (n = n0));
SET WORK.rmean;
IF trt = 0 THEN OUTPUT WORK.rmean1;
ELSE OUTPUT WORK.n_act;
RUN;

/*********************/
/* Compute sigma_ij0 */
/*********************/
PROC RANK DATA = WORK.ex2 OUT = WORK.rank2 FRACTION TIES = MEAN;
BY j;
RANKS rank;
VAR y;
RUN;

PROC MIXED DATA = WORK.rank2;
BY j;
CLASS trt;
MODEL rank =  / S;
MAKE 'COVPARMS' OUT = WORK.sigma;
RUN;

/***************************/
/* Compute test statistics */
/***************************/
DATA WORK.test;
MERGE WORK.rmean1 WORK.n_act WORK.sigma;
BY j;
lambda = SQRT(n/(n + n0));
p = (rmean - (n0+1)/2) / n;
temp = SQRT(n0) * lambda / SQRT(est);
t = temp * (p-0.5);
abst = ABS(t);
RUN;

PROC TRANSPOSE DATA = WORK.test OUT = WORK.lambda PREFIX = lambda;
VAR lambda;
ID j;
RUN;

/***********************************/
/* Compute adjusted p-values and CI'*/
/***********************************/
DATA WORK.pval;
IF _N_ = 1 THEN SET WORK.lambda;
SET WORK.test;
pval = 1 - (PROBMC("DUNNETT2",abst,.,.,2,lambda11,lambda12) *
            PROBMC("DUNNETT2",abst,.,.,2,lambda21,lambda22));
RUN;

DATA _NULL_;
SET WORK.lambda;
CALL SYMPUT('lambda11',lambda11);
CALL SYMPUT('lambda12',lambda12);
CALL SYMPUT('lambda21',lambda21);
CALL SYMPUT('lambda22',lambda22);
```

```
RUN;

PROC IML;
lambda11 = &lambda11;
lambda12 = &lambda12;
lambda21 = &lambda21;
lambda22 = &lambda22;

alpha = 0.05;
c = 2;                           /* number of active treatments */
r = 2;                           /* number of groups */
q1 = PROBIT(1 - alpha/2);        /* start value: uncorr. t value */
q2 = PROBIT(1 - alpha/(2*c*r));  /* start value: Bonf. corr. value */
crit = 1 - (PROBMC("DUNNETT2",q2,.,.,2,lambda11,lambda12) *
            PROBMC("DUNNETT2",q2,.,.,2,lambda21,lambda22));

n = 1;
DO WHILE ((ABS(crit - alpha) > 0.0001) & (n < 20));/* max 20 steps */
  n = n + 1;
  qm = (q1 + q2)/2;
  crit = 1 - (PROBMC("DUNNETT2",qm,.,.,2,lambda11,lambda12) *
              PROBMC("DUNNETT2",qm,.,.,2,lambda21,lambda22));
  IF crit > alpha THEN q1 = qm;
  ELSE q2 = qm;
END;
cval = qm;
CALL SYMPUT('cval', CHAR(cval));

QUIT;

DATA WORK.cis;
SET WORK.pval;
c_l = p - &cval / temp;
c_u = p + &cval / temp;
RUN;

TITLE 'Adjusted p-values, critical value and lower and upper bounds';
PROC PRINT DATA = WORK.cis;
RUN;
```

Program ***CH8.SAS*** performs the analyses of Chapter 8

```
/*********************************************************/
/* Program applying standard resampling methods          */
/* for ones-sided testing problem                        */
/*                                                       */
/* Input:  Example dataset                               */
/* Output: p-values and CI's for the stochastic approx.  */
/*       and bootstrap and permutation method            */
/*********************************************************/

%GLOBAL _PRINT_;
%LET _PRINT_ = ON;

OPTIONS NOBYLINE;

/****************************/
/* Input of example dataset */
/****************************/
LIBNAME DAT 'c:\My SAS Files\...';

DATA WORK.example;
SET DAT.example;
trt1 = trt;
IF trt NE 0 THEN trt1 = 10 * strata + trt;  /* Active treatments are unique
                                                               per stratum */
trts = 10 * strata + trt;                   /* Unique treatment code per
                                                                   stratum */
RUN;

/****************************/
/* Stochastic approximation */
/****************************/
TITLE1 'Resampling techniques';
TITLE2 'Stochastic Approximation: Acc = 0.001 Eps = 0.01';
PROC MIXED DATA = WORK.example;
CLASS strata trt1;
MODEL y = strata trt1;
LSMEANS trt1 / ADJUST = SIMULATE(ACC=0.001 EPS= 0.01 SEED =99) DIFF =
CONTROLU ('0') CL;
RUN;

/*************/
/* Bootstrap */
/*************/
TITLE2 'Bootstrap including Step-down procedure: N = 50000';
PROC MULTTEST BOOT N = 50000 DATA = WORK.example SEED = 99 OUT = WORK.pvals
STEPBOOT;
CLASS trts;
STRATA strata;
TEST MEAN (y / UPPER);
CONTRAST 'Stratum 1 Plac - 1' -1 1 0  0 0 0;
CONTRAST 'Stratum 1 Plac - 2' -1 0 1  0 0 0;
CONTRAST 'Stratum 2 Plac - 1'  0 0 0 -1 1 0;
CONTRAST 'Stratum 2 Plac - 2'  0 0 0 -1 0 1;
RUN;
```

```
/***************/
/* Permutation */
/***************/
TITLE2 'Permutation including Step-down: N = 50000';
PROC MULTTEST PERM N = 50000 DATA = WORK.example SEED = 99 OUT = WORK.pvals
STEPPERM;
CLASS trts;
STRATA strata;
TEST MEAN (y / UPPER);
CONTRAST 'Stratum 1 Plac - 1' -1 1 0  0 0 0;
CONTRAST 'Stratum 1 Plac - 2' -1 0 1  0 0 0;
CONTRAST 'Stratum 2 Plac - 1'  0 0 0 -1 1 0;
CONTRAST 'Stratum 2 Plac - 2'  0 0 0 -1 0 1;
RUN;
```

Acknowledgements

Only my name is mentioned on the cover of this work. However it is obvious that no one can do research by himself alone. Therefore I would like to thank anybody who has contributed to this work. In particular I would like to mention a few people.

First I would like to thank Hans Koch. Working as a statistician at Solvay Pharmaceuticals we had a conversation about my professional training in the end of 1996. Hans, because of your enthusiasm of multiple comparisons you made the proposal that initiated this work. You introduced me to prof. L.A. Hothorn, stressed the importance of doing a PhD and you arranged the necessary conditions to perform this research next to my regular job at the company. Thanks. In this context, I would also like to thank Solvay Pharmaceuticals for providing the support for doing my thesis.

Prof. Ludwig Hothorn, I’m very grateful to you. You introduced me to the exciting world of multiple comparisons. But more important you were willing to be my supervisor. I thank you for the patience you had during these years and your understanding of my situation of being a Dutch guy not living nearby the institute and having a regular job that made the progress sometimes very slow. Without your enthusiasm and drive this thesis would never have finalized. In addition I would like to thank the hospitality of your *LG Bioinformatik* during my stays in Hannover. (Hannelore Visser and Clemens Buczilowski thank you as well.)

Frank Bretz and Jeroen Tolboom, I would like to thank both of you for so many discussions we have had in the last couple of years about multiple comparisons and the more general statistical issues. You contributed a lot to the progress of my work.
Jan van der Stam, thanks for supporting me with programming the code in SAS.

Without the confidence of friends and family I wouldn’t had the motivation to continue this research. You made me aware never to forget that there is more than work and multiple comparisons. Thanks for putting this in the right perspective.
Last but not least I would like to thank my parents and Kees and Nelleke. Thanks for your never ending love and support.

Curriculum Vitae

Name:	Biesheuvel
Christian name:	Egbertus Hendrikus Evert
Birthday:	January 6th, 1967
Place of Birth:	Kampen, The Netherlands

Education

8/1979 – 7/1985 Atheneum at comprehensive school ‘De Brug’ in Lelystad, The Netherlands

9/1985 – 7/1991 Mathematics at the faculty of Mathematics, Free University of Amsterdam with specialization statistics (MSc. degree)

Including: student assistant for general statistics (9/1989 – 4/1990), practical work at the Mathematics, Informatics and Statistics department of Duphar BV, Weesp, The Netherlands (4/1990 – 6/1991)

1/1997 – PhD student of Prof. L. Hothorn at the department Bioinformatik, Gartenbau faculty, University of Hanover

Profession

7/1991 – 11/1992 Biostatistician at the Epidemiology and Biostatistics department of the medical faculty, Free University of Amsterdam

12/1992 – 7/2001 Project Statistician at the Biometrics department of the pharmaceutical company Solvay Pharmaceuticals BV, Weesp, The Netherlands

11/2001 – Group Head at the Biometrics department of the pharmaceutical company Organon NV, Oss, The Netherlands