Chapter 2

Multiple Testing Methodology

Alex Dmitrienko
Eli Lilly and Company

Frank Bretz
Novartis

Peter H. Westfall
Texas Tech University

James Troendle
National Institutes of Health

Brian L. Wiens
Alcon Laboratories

Ajit C. Tamhane
Northwestern University

Jason C. Hsu
Ohio State University

2.1 Introduction

Multiplicity issues are encountered in a variety of pharmaceutical applications with multiple objectives. In a pre-clinical setting, the objectives can correspond to multiple genetic markers. In clinical applications, the objectives can be defined in terms of multiple dose levels, endpoints or subgroup analyses. Most common sources of multiplicity in clinical trials are listed below:

- Multiple dose-control comparisons are commonly included in dose-finding studies to evaluate efficacy and safety properties of a treatment compared to a control.
- Multiplicity is often induced by multiple criteria for assessing the effi-

cacy profile of a treatment. These criteria are introduced to help the trial's sponsor better evaluate multiple dimensions of the treatment effect. Depending on the trial's objectives, the overall outcome may be declared positive if (i) one or more criteria are met or (ii) all criteria are met or (iii) some composite criterion is met.

- Another source of multiplicity in clinical trials is multiple secondary analyses, e.g., analysis of secondary endpoints or subgroup effects that are typically performed after the trial's primary objective is met.

This chapter provides an overview of key concepts and approaches in multiple testing methodology. For more information about the theory of multiple comparisons and a detailed review of popular multiple testing procedures and their properties, see Hochberg and Tamhane (1987), Westfall and Young (1993) and Hsu (1996).

The chapter is organized as follows. Sections 2.2 through 2.4 review fundamental concepts and principles that provide a foundation for the theory of multiple comparisons. Sections 2.5 through 2.8 describe commonly used multiple tests in pre-clinical and clinical studies. Lastly, Section 2.9 describes software packages that implement popular multiple testing procedures.

2.2 Error rate definitions

In order to choose an appropriate multiple testing method, it is critical to select the definition of correct decisions that reflect the objective of a clinical study or pre-clinical experiment. This section introduces an error rate definition commonly used in clinical trials (familywise error rate) as well as other definitions (generalized familywise error rate, false discovery rate and false discovery proportion) that have found applications in pre-clinical and clinical studies.

2.2.1 Comparisonwise and familywise error rates

The concept of a Type I error rate originates in the problem of testing a single hypothesis. It is defined as the probability of rejecting the hypothesis when it is true. As an example, consider a dose-finding study with m doses tested versus placebo. The primary endpoint is normally distributed with larger values indicating improvement. Let μ_0 be the mean improvement in the placebo arm and μ_i be the mean improvement in the ith dose group, $i = 1, \ldots, m$. The testing problem is formulated in terms of the difference in the mean responses: The hypothesis of treatment effect no greater than δ

$$H_i : \ \mu_i - \mu_0 \leq \delta$$

is tested versus a one-sided alternative

$$K_i: \ \mu_i - \mu_0 > \delta,$$

where δ is a non-negative constant defining the clinically important difference. The Type I error rate for H_i is the probability of concluding that a clinically relevant treatment effect is present when the treatment difference is actually no greater than δ.

If each of the m hypotheses is tested separately at a pre-specified significance level α, e.g., $\alpha = 0.05$, it can be shown that the proportion of incorrectly rejected hypotheses will not exceed α. This is known as the control of the *comparisonwise error rate*. However, preserving the comparisonwise error rate is not considered an adequate approach to controlling the probability of incorrect decisions in a clinical trial setting. The hypotheses of interest are considered together as a family. Even a single Type I error in this family is assumed to lead to an incorrect decision. Accordingly, the overall Type I error rate is defined as the probability of rejecting at least one true hypothesis. The probability can be computed under the assumption that all m hypotheses are simultaneously true. This is known as the *weak control of the familywise error rate* (FWER). In the context of clinical trials with multiple endpoints, the weak FWER control can be interpreted as the probability of concluding an effect on at least one endpoint when there is no effect on any endpoint, i.e., the probability of concluding an ineffective treatment has an effect.

In general, the assumption that all hypotheses are true at the same time may be restrictive in many pharmaceutical applications and is not appropriate in the cases when the clinical trial sponsor is interested in making claims about specific outcomes. For example, in dose-finding clinical trials, the treatment difference is likely to vary across the dose levels and the trial's sponsor is generally interested in testing the drug effect at each particular dose and being able to claim that this effect is significant. To achieve this goal, one needs to preserve the probability of an incorrect decision (that is, the probability of erroneously finding a significant result) for each dose regardless of the size of the treatment effect in other dose groups. Using mathematical terminology, this requirement can be reformulated as the control of the probability of incorrectly rejecting any true hypothesis regardless of which and how many other hypotheses are true. In other words, if T is the index set of true null hypotheses, we require that

$$\sup \text{FWER} = \max_{T} \sup_{\{\mu_i(T)\}} P \ (\text{Reject at least one } H_i, \ i \in T) \leq \alpha,$$

where the supremum is taken over all μ_i satisfying $\mu_i - \mu_0 \leq \delta$ for $i \in T$ and $\mu_i - \mu_0 > \delta$ for $i \notin T$, and the maximum is taken over all index sets T. This approach to protecting the overall error rate is known as *strong control of the familywise error rate*. Strong control of the FWER for the primary objectives is mandated by regulators in all confirmatory clinical trials (CPMP, 2002). A

detailed description of multiple tests that protect the FWER in the strong sense is given in Sections 2.6–2.8.

It is worth noting that some multiple tests such as the unprotected and protected least significance difference (LSD) tests do not protect the FWER (Hochberg and Tamhane, 1987, Chapter 1). The former does not control the FWER even in the weak sense while the latter does not control it in the strong sense. These and similar tests will not be discussed in this chapter.

2.2.2 Generalized familywise error rate, false discovery rate and false discovery proportion

The definition of the FWER in the previous section is based on preserving the chances of rejecting at least one true null hypothesis for any number of hypotheses. This approach is reasonable when one deals with a few hypotheses but can become impractical in studies involving a large number of hypotheses, for example, in microarray experiments (Chapter 7). Specifically, as the number of hypotheses, m, increases, FWER-controlling multiple tests become conservative and fail to detect significant results unless the treatment effect is overwhelmingly positive.

The standard FWER definition can be extended by relaxing the requirement to protect the probability of at least one incorrect conclusion. This approach improves the power of multiple tests by increasing the probability of (correctly) rejecting false hypotheses. Romano, Shaikh and Wolf (2005) pointed out that, although one will need to pay a price for this in terms of an increased Type I error rate, "the price to pay can be small compared to the benefits to reap."

Generalized familywise error rate

The *generalized familywise error rate* (gFWER) definition assumes that one can tolerate a certain fixed number k ($1 \leq k < m$) of incorrect conclusions regardless of how many hypotheses are considered (Victor, 1982; Hommel and Hoffmann, 1987; Lehmann and Romano, 2005). In mathematical terms, the control of the generalized FWER is achieved if

$$\sup \mathrm{gFWER}(k) = \max_{T} \sup_{\{\mu_i(T)\}} P\left(\text{Reject at least } k \text{ hypotheses } H_i,\ i \in T\right) \leq \alpha,$$

where T is the index set of at least k true null hypotheses. Note that the gFWER simplifies to the usual FWER when $k = 1$. Multiple testing procedures for controlling the gFWER are discussed in Chapter 7 in the context of problems arising in pharmacogenomic studies.

False discovery rate and false discovery proportion

Two closely connected approaches to extend the FWER are known as the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995) and the *false discovery proportion* (FDP) (Korn et al., 2004).

If the number of rejected hypotheses is positive, then the FDP is defined as

$$\text{FDP} = \left(\frac{\text{Number of rejected true null hypotheses}}{\text{Number of rejected hypotheses}} \right).$$

The FDP is defined as 0 if no hypotheses are rejected. The FDR is said to be controlled at the γ level if

$$\text{FDR} = E(\text{FDP}) \leq \gamma.$$

Note that control of the FDR at the γ level does not imply that the FDP is less than or equal to γ with high probability. To ensure this, one can choose an acceptable probability of exceedence, α, and require that

$$P(\text{FDP} > \beta) \leq \alpha.$$

The interpretation is that of those hypotheses that are rejected, the proportion of false discoveries may exceed a specified fraction β with probability no larger than α. Note that control of the FWER is equivalent to control of the FDP with $\beta = 0$. Control of the FDP makes sense in many nonconfirmatory settings like genetic or pre-clinical studies, where a certain proportion of errors is considered acceptable.

Control of the FDR at the α level does not imply control of the FWER at the α level, nor does any ($\beta > 0$) control of the FDP at the α level imply control of the FWER at the α level. In fact, it is often possible to manipulate the design of a clinical trial so that any desired conclusion can be almost surely inferred without inflating the FDR (Finner and Roter, 2001). Thus, FDR or FDP controlling procedures are not suitable for confirmatory clinical trials.

2.2.3 Role of Type II errors

Similarly as in the case of the Type I error rate, the Type II error rate is not extended uniquely from the univariate case to the multiple-hypotheses case. Different possibilities exist to measure the success of a clinical trial in terms of power. A standard approach is to consider the probability of rejecting *at least one* false null hypothesis (disjunctive power), that is, to calculate

$$P \text{ (Reject at least one } H_i,\ i \notin T),$$

where the probability is evaluated for a given set of parameter values: $\mu_i > \mu_0 + \delta$ if $i \notin T$ and $\mu_i = \mu_0 + \delta$ if $i \in T$ (Senn and Bretz, 2007). The use of disjunctive power is recommended in, for example, studies involving multiple comparisons with a control or in studies with multiple endpoints, where it

is sufficient to demonstrate the treatment's effect on at least one endpoint. Alternatively, one may be interested in calculating the probability of rejecting *all* false null hypotheses (conjunctive power)

$$P \text{ (Reject all } H_i,\ i \notin T),$$

where probability is again evaluated a given set of parameter values. One may argue that conjunctive power should be used in, for example, fixed drug combination studies or studies in which the treatment's effect must be established on two or more co-primary endpoints (see also Section 4.5).

Other power concepts exist and we refer to Maurer and Mellein (1987) and Westfall et al. (1999) for further details. General software implementations are not available, even for some of the simpler multiple comparison procedures and in most case extensive simulations need to be performed. The practically relevant question about the appropriate power concept needs to be addressed on a case-by-case basis tailored to the study objectives, but see Hommel and Bretz (2008) for a balance between power and other considerations in multiple testing. It should be noted that adequately powering a clinical study is typically in the interest of the trial sponsor: It is the sponsor's choice to control the risk of failing to detect a truly significant drug effect.

2.3 Multiple testing principles

This section reviews key principles that provide a foundation for multiple tests described in this chapter. It begins with two general principles, known as the principles of *union-intersection testing* and *intersection-union testing*, that define the underlying testing problem. The section also introduces two methods for constructing multiple tests (*closure principle* and *partitioning principle*).

2.3.1 Union-intersection testing

Multiple testing problems in pharmaceutical applications are commonly formulated as union-intersection problems (Roy, 1953). Within the union-intersection framework, one rejects the global hypothesis of no effect if there is evidence of a positive effect with respect to at least one individual objective. To provide a mathematical definition, let $H_1, \ldots, H_m$ denote the hypotheses corresponding to the multiple objectives. The hypotheses are tested against the alternative hypotheses $K_1, \ldots, K_m$. The global hypothesis H_I, defined as the intersection of the hypotheses, is tested versus the union of the alternative

hypotheses (K_U):

$$H_I : \bigcap_{i=1}^{m} H_i \text{ versus } K_U : \bigcup_{i=1}^{m} K_i.$$

In the context of union-intersection testing, carrying out the individual tests at an unadjusted α level leads to an inflated probability of rejecting H_I and can compromise the validity of statistical inferences. To address this problem, a multiplicity adjustment method needs to be utilized to control the appropriately defined probability of a Type I error.

2.3.2 Intersection-union testing

A different class of multiple testing problems requires a different approach called the intersection-union testing approach. Intersection-union testing arises naturally in studies when a significant outcome with respect to two or more objectives is required in order to declare the study successful. For example, new therapies for the treatment of Alzheimer's disease are required to demonstrate their effects on both cognition and global clinical scores.

In other words, the intersection-union method involves testing the union of the hypotheses (H_U) against the intersection of the alternative hypotheses (K_I):

$$H_U : \bigcup_{i=1}^{m} H_i \text{ versus } K_I : \bigcap_{i=1}^{m} K_i.$$

When the global hypothesis H_U is rejected, one concludes that all K_i's are true, i.e., there is evidence of a positive effect with respect to all of the m objectives.

An interesting feature of intersection-union tests is that no multiplicity adjustment is necessary to control the size of a test but the individual hypotheses cannot be tested at levels higher than the nominal significance level either (Berger, 1982). Note that intersection-union tests are sometimes biased in the sense that their power function can drop below their size (Type I error rate) in the alternative space. For a detailed discussion of intersection-union tests in the analysis of multiple endpoints, see Section 4.5.

2.3.3 Closure principle

The closure principle proposed by Marcus, Peritz and Gabriel (1976) plays a key role in the theory of multiple testing and provides a foundation for virtually all multiple testing methods arising in pharmaceutical applications. This principle has been used to construct a variety of stepwise testing procedures.

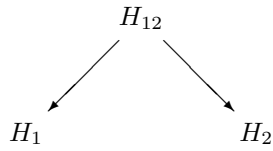

FIGURE 2.1: Closed testing procedure in the dose-finding clinical trial example with two null hypotheses.

Motivating example

To introduce the closure principle and demonstrate how it can be used to derive powerful multiple tests, consider the dose-finding trial example described in Section 2.2.1. Assume that two doses are compared to a placebo ($m = 2$) and the clinically important difference δ is equal to 0. The two associated hypotheses are given by $H_1 : \ \mu_1 \leq \mu_0$ and $H_2 : \ \mu_2 \leq \mu_0$. Let p_1 and p_2 denote the p-values for testing H_1 and H_2.

To construct a closed testing procedure for H_1 and H_2, we first need to define the *closed family of hypotheses* by considering all possible intersections of the two hypotheses. This family contains the following three intersection hypotheses:

$$H_1, \ H_2 \text{ and } H_{12} = H_1 \cap H_2.$$

The closure principle states that an FWER-controlling testing procedure can be constructed by testing each hypothesis in the closed family using a suitable *local* α-level test. This procedure rejects a hypothesis if all intersection hypotheses containing this hypothesis are rejected by the associated local tests. The decision rule is depicted in Figure 2.1. To conclude that $\mu_1 > \mu_0$, the two intersection hypotheses containing H_1, i.e., H_1 and H_{12}, need to be rejected. Likewise, to show that $\mu_2 > \mu_0$, H_2 and H_{12} need to be rejected.

As an illustration, we will construct a closed testing procedure based on the Bonferroni test. Let α denote the significance level, e.g., $\alpha = 0.05$. The local tests reject H_1 and H_2 if $p_1 \leq \alpha$ and $p_2 \leq \alpha$, respectively. Further, the Bonferroni test is carried out to test H_{12}, i.e., the hypothesis is rejected by its local test if $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$. When the decision rules are combined, it is easy to show that the resulting procedure has the following form:

- The procedure begins with the more significant p-value and rejects the corresponding hypothesis if the p-value is no greater than $\alpha/2$.
- If the first hypothesis is rejected, the other hypothesis is tested at an α level. Otherwise, the other hypothesis is retained without testing.

Since the second test is carried out at the α level, the closed testing procedure rejects as many (and sometimes more) hypotheses as the Bonferroni test from which it was derived. The power gain is due to the application of the

closure principle. As a side note, this closed testing procedure is actually the stepwise testing procedure proposed by Holm (1979). For more information about the Bonferroni and Holm procedures, see Section 2.6.

General case

In the general case of testing m hypotheses, the process of constructing a closed testing procedure goes through the following steps:

- Define the closed family of hypotheses. For each non-empty index set $I \subseteq \{1, \ldots, m\}$, consider an intersection hypothesis defined as

$$H_I = \bigcap_{i \in I} H_i.$$

- Establish implication relationships. An intersection hypothesis that contains another intersection hypothesis is said to imply it, i.e., H_I implies H_J if $J \subset I$.

- Define local α-level tests for individual intersection hypotheses. Let p_I denote the p-value produced by the associated local test and reject H_I iff $p_J \leq \alpha$ for all $J \subseteq I$. In particular, reject H_i if and only if (iff) all intersection hypotheses containing H_i are rejected by their local tests. In other words, reject H_i if and only if $p_I \leq \alpha$ for all index sets I that include i.

Marcus et al. (1976) showed that this closed testing procedure for the hypotheses $H_1, \ldots, H_m$ controls the FWER in the strong sense at the α level. To see how closed testing procedures achieve strong FWER control, it is instructive to revisit the FWER definition given in Section 2.2. By considering all possible combinations of hypotheses and defining an α-level test for each intersection, we ensure that the resulting procedure protects the Type I error rate for any configuration of true hypotheses. This immediately implies that the FWER is controlled in the strong sense.

The closed testing algorithm is generally computationally intensive since approximately 2^m individual tests need to be carried out to test m hypotheses. Because of this, shortcut versions of closed testing procedures have attracted attention in the multiple testing literature. Shortcut procedures have a stepwise form and reduce the number of computational steps from order-2^m to order-m or order-m^2. In addition, as will be explained in Section 2.6, stepwise procedures provide useful insights into the process of performing multiplicity adjustments and are easy to communicate to non-statisticians. For more information on stepwise closed testing procedures, see Grechanovsky and Hochberg (1999), Westfall et al. (2001), Dmitrienko et al. (2006b), Hommel, Bretz and Maurer (2007) and Bretz et al. (2009b).

2.3.4 Properties of closed testing procedures

This section briefly describes important properties of closed testing procedures that will be referenced later in this chapter and other chapters.

Monotone procedures

A monotone procedure rejects a hypothesis whenever it rejects another hypothesis with a larger p-value. For example, if $p_i < p_j$ then the rejection of H_j automatically implies the rejection of H_i. Monotonicity helps to avoid logical inconsistencies; as such it is an essential requirement for multiple testing procedures. When a procedure does not have this property, monotonicity needs to be enforced by updating adjusted p-values. The Shaffer procedure introduced in Section 2.6.2 serves as an example of a procedure that requires monotonicity to be enforced. For a review of other monotonicity considerations, see Hommel and Bretz (2008).

Consonant procedures

A closed testing procedure is termed consonant (Gabriel, 1969) if the rejection of an intersection hypothesis H_I with $I \subseteq \{1, \ldots, m\}$ and $|I| > 1$ always leads to the rejection of at least one H_J implied by H_I, i.e., H_J with $J \subset I$. While consonance is generally desirable, nonconsonant procedures can be of practical importance. The Hommel procedure defined in Section 2.6.8 is an example of a nonconsonant closed testing procedure. It is possible for this procedure to reject the global null hypothesis H_I, $I = \{1, \ldots, m\}$, without rejecting any other intersection hypotheses.

α-exhaustive procedures

An α-exhaustive procedure is a closed testing procedure based on intersection hypothesis tests the size of which is exactly α (Grechanovsky and Hochberg, 1999). In other words, $P(\text{Reject } H_I) = \alpha$ for any intersection hypothesis H_I, $I \subseteq \{1, \ldots, m\}$. If a procedure is not α-exhaustive, one can construct a uniformly more powerful procedure by setting the size of all intersection hypothesis tests at α. It is worth noting that some popular multiple testing procedures, for example, the fallback and Hochberg procedures described in Sections 2.6.4 and 2.6.9, respectively, are not α-exhaustive. These procedures are used in pharmaceutical applications due to other desirable properties such as computational simplicity.

2.3.5 Partitioning principle

The partitioning principle was introduced by Stefansson, Kim and Hsu (1988) and Finner and Strassburger (2002). The advantage of using this principle is two-fold:

- It can be used to construct procedures that are more powerful than procedures derived using the closed testing principle.
- Partitioning procedures are easy to invert in order to set up simultaneous confidence sets for parameters of interest (these sets are constructed by inverting partitioning tests as explained in Section 2.4.2).

Motivating example

To illustrate the process of carrying out partitioning tests, consider the clinical trial example with two doses and a placebo from Section 2.3.3. The first step involves partitioning the union of the hypotheses

$$H_1 : \ \mu_1 \leq \mu_0, \quad H_2 : \ \mu_2 \leq \mu_0.$$

into three mutually exclusive hypotheses:

$$\begin{aligned} H_1^* &: \ \mu_1 \leq \mu_0 \quad \text{and} \quad \mu_2 \leq \mu_0, \\ H_2^* &: \ \mu_1 \leq \mu_0 \quad \text{and} \quad \mu_2 > \mu_0, \\ H_3^* &: \ \mu_1 > \mu_0 \quad \text{and} \quad \mu_2 \leq \mu_0. \end{aligned}$$

Since the three hypotheses are disjoint, each one of them can be tested at level α without compromising the FWER control. The final decision rule is constructed by considering all possible outcomes for the three mutually exclusive hypotheses. For example,

- If H_1^* is rejected, we conclude that $\mu_1 > \mu_0$ or $\mu_2 > \mu_0$.
- If H_1^* and H_2^* are rejected, we conclude that $\mu_1 > \mu_0$ and, similarly, rejecting H_1^* and H_3^* implies that $\mu_2 > \mu_0$.
- If H_1^*, H_2^* and H_3^* are all rejected, the conclusion is that $\mu_1 > \mu_0$ and $\mu_2 > \mu_0$.

This test appears conceptually similar to the closed test described in Section 2.3.3. However, unlike the closure, the partitioning principle does not deal with the hypotheses in the closed family (i.e., H_1, H_2 and $H_1 \cap H_2$) but rather with mutually exclusive hypotheses that partition the union of H_1 and H_2.

General case

To briefly describe a general version of the partitioning principle, let θ be the k-dimensional parameter of interest in a pre-clinical experiment or a clinical trial, $k \geq 1$. Suppose m hypotheses are considered and assume that H_i states that $\theta \in \Theta_i$, where Θ_i is a subset of the k-dimensional space, $i = 1, \ldots, m$. For example, in the dose-finding example discussed above, θ is a three-dimensional vector of true treatment means, $\theta = (\mu_0, \mu_1, \mu_2)$, and H_1 and H_2 are formulated in terms of

$$\Theta_1 = \{(\mu_0, \mu_1, \mu_2) : \ \mu_1 \leq \mu_0\}, \quad \Theta_2 = \{(\mu_0, \mu_1, \mu_2) : \ \mu_2 \leq \mu_0\}.$$

Given $\Theta_1, \ldots, \Theta_m$, partition the union of the m subsets into disjoint subsets Θ_I^*, $I \subseteq \{1, \ldots, m\}$. Each subset can be interpreted as a part of the k-dimensional space in which the hypotheses H_i, $i \in I$, are true and the remaining hypotheses are false. The next step is to define hypotheses corresponding to the constructed subsets,

$$H_I^* : \ \theta \in \Theta_I^*,$$

and test them at the α level. Since these hypotheses are mutually exclusive, at most one of them is true. Thus, even though no multiplicity adjustment is made, the resulting multiple test controls the FWER at the α level.

For more information about the partitioning principle and its applications to multiplicity problems in pre-clinical and clinical studies, see Hsu and Berger (1999), Xu and Hsu (2007), Strassburger, Bretz and Finner (2007) and Strassburger and Bretz (2008).

2.4 Adjusted significance levels, p-values and confidence intervals

Multiple inferences are performed by adjusting decision rules for individual hypotheses. This can be accomplished by computing multiplicity-adjusted significance levels, multiplicity-adjusted p-values or simultaneous confidence intervals. To avoid inflation of the overall Type I error rate in multiple testing problems, significance levels for individual hypotheses are adjusted downward or p-values are adjusted upward. Similarly, wider confidence intervals for parameters of interest need to be chosen to keep the overall coverage probability at a pre-determined level.

2.4.1 Adjusted significance levels and p-values

In most simple cases, a multiplicity adjustment can be performed by computing a reduced significance level for each individual hypothesis. For example,

in the problem of testing the hypotheses $H_1, \ldots, H_m$, a multiple test can be carried out by comparing the p-value associated with H_i to a significance level, α_i, which is lower than the nominal α level. The α_i's are selected to maintain the FWER at the α level.

In general, adjusted significance levels are used less frequently than adjusted p-values, mainly because adjusted significance levels depend on the α level. However, there are cases when the use of adjusted significance levels simplifies multiplicity adjustments. Consider, for example, a meta analysis that combines several multinational studies. If different multiplicity adjustment strategies are required by different regulatory agencies, the meta analysis may be easier to implement using raw p-values with appropriately adjusted significance levels.

Unlike adjusted significance levels, adjusted p-values capture the degree of multiplicity adjustment without reference to the pre-specified error rate and thus one can choose different α levels for different sets of hypotheses. For example, a clinical trial sponsor can pre-specify the 0.05 significance level for hypotheses corresponding to the primary objectives and a higher level (e.g., $\alpha = 0.1$) for secondary hypotheses. Another advantage of adjusted p-values is that they incorporate the structure of the underlying decision rule which can be quite complex. Considerations of this type become important, for example, in the context of gatekeeping procedures described in Chapter 5.

A general definition of an adjusted p-value is given in Westfall and Young (1993): The adjusted p-value for a hypothesis is the smallest significance level at which one would reject the hypothesis using the given multiple testing procedure. This definition can be illustrated by applying it to closed testing procedures. As was explained in Section 2.3.3, a closed testing procedure rejects a hypothesis, for example, H_i, if all intersection hypotheses containing H_i are rejected. If p_I, $I \subseteq \{1, \ldots, m\}$, denotes the p-value for testing the intersection hypothesis H_I, the adjusted p-value for H_i is the largest p-value associated with the index sets including i:

$$\widetilde{p}_i = \max_{I:\ i \in I} p_I.$$

The hypothesis H_i is rejected if the adjusted p-value does not exceed the pre-specified α level, i.e., $\widetilde{p}_i \leq \alpha$. This general approach will be utilized in Sections 2.6–2.8 to derive adjusted p-values for multiple testing procedures commonly used in pharmaceutical applications (all of which can be formulated as closed testing procedures).

2.4.2 Simultaneous confidence intervals

Lehmann (1986, page 90) described the following general method for constructing a confidence set from a significance test. Let θ denote the parameter of interest. For each parameter value θ_0, test the hypothesis $H: \theta = \theta_0$ using an α-level test and then consider the set of all parameter values θ_0 for which

$H: \theta = \theta_0$ is retained. The set is, in fact, a $100(1-\alpha)\%$ confidence set for the true value of θ. This method is essentially based on partitioning the parameter space into subsets consisting of a single parameter point each.

In the context of multiple hypothesis testing the partitioning principle described in Section 2.3.5 provides a natural extension of this general method to derive simultaneous confidence intervals that are compatible with a given multiple testing procedure (Hayter and Hsu, 1994; Finner and Strassburger, 2002).

Applying the partitioning principle, the parameter space is partitioned into small disjoint subhypotheses, each of which is tested with an appropriate test. The union of all non-rejected hypotheses then yields a confidence set C for the parameter vector of interest (see Finner and Strassburger, 2002, for a formal description). Note that the finest possible partition is given by a pointwise partition such that each point of the parameter space represents an element of the partition. Most of the classical (simultaneous) confidence intervals can be derived by using the finest partition and an appropriate family of one- or two-sided tests. However, this is not true in general. Note that a confidence set C can always be used to construct simultaneous confidence intervals by simply projecting C on the coordinate axes. Compatibility can be ensured by enforcing mild conditions on the partition and the test family (Strassburger, Bretz and Hochberg, 2004). In the following sections we will define simultaneous confidence intervals for popular multiple testing procedures. We will see that simultaneous confidence intervals are easily obtained for single-step procedures, but are often difficult to derive for stepwise procedures.

2.5 Common multiple testing procedures

This section provides background information and sets the stage for the next three sections (Sections 2.6–2.8) which review popular multiple testing procedures in pharmaceutical applications. We will begin by introducing several possible classification schemes based on the testing sequence, distributional assumptions and control of the Type I error rate.

2.5.1 Classification of multiple testing procedures

Single-step and stepwise procedures

Two important types of multiple testing procedures considered in Sections 2.6–2.8 are *single-step* and *stepwise* procedures described below.

Single-step procedures are multiple testing procedures for which the decision to reject any hypothesis does not depend on the decision to reject any other hypothesis. In other words, the order in which the hypotheses are

tested is not important and one can think of the multiple inferences as being performed simultaneously in a single step. The Bonferroni procedure (Section 2.6.1) and Dunnett procedure (Section 2.7.1) are examples of single-step procedures.

Unlike single-step procedures, stepwise procedures are carried out in a sequential manner. Some hypotheses are not tested explicitly and may be retained or rejected by implication. Stepwise procedures provide an attractive alternative to single-step procedures because they can reject more hypotheses without inflating the overall error rate.

The stepwise testing approach can be implemented via *step-down* or *step-up* procedures:

- A step-down procedure starts with the most significant p-value and continues in a sequentially rejective fashion until a certain hypothesis is retained or all hypotheses are rejected. If a hypothesis is retained, testing stops and the remaining hypotheses are retained by implication. The Holm procedure is an example of a step-down testing procedure.

- Step-up procedures approach the hypothesis testing problem from the opposite direction and carry out individual tests from the least significant one to the most significant one. The final decision rule is reversed compared to step-down procedures; i.e., once a step-up procedure rejects a hypothesis, it rejects the rest of the hypotheses by implication. The Hochberg procedure is an example of a step-up testing procedure.

The Holm and Hochberg procedures mentioned above are defined in Sections 2.6.2 and 2.6.9, respectively.

Distributional assumptions

Another useful approach to the classification of multiple testing procedures is based on the assumptions they make about the joint distribution of the test statistics. This approach leads to the following classification scheme:

- Procedures that don't make any assumptions about the joint distribution of the test statistics. These procedures rely on univariate p-values and thus tend to have a rather straightforward form. They are referred to as *p-value based procedures* or nonparametric procedures. Examples include many popular procedures such as the Bonferroni and Holm procedures. These and similar procedures are discussed in Section 2.6.

- Procedures that make specific distributional assumptions, for example, that the test statistics follow a multivariate normal or t-distribution. To contrast this approach with nonparametric procedures based on univariate p-values, they are termed *parametric procedures*. Examples include the Dunnett and related procedures introduced in Section 2.7.

- Procedures that do not make specific assumptions and attempt to approximate the true joint distribution of the test statistics. The approximation relies on resampling-based methods (bootstrap or permutation methods) and thus procedures in this class are often referred to as *resampling-based procedures*. The resampling-based approach is described in Section 2.8.

It is important to point out that p-value-based procedures tend to perform poorly, compared to parametric and resampling-based procedures, when the testing problem involves a large number of hypotheses or the test statistics are strongly correlated. This is due to the fact that procedures that do not account for the correlation among test statistics become conservative in these cases.

Control of the Type I error rate

The multiple testing procedures described in Sections 2.6–2.8 focus on the strong control of the FWER. Procedures that control alternative error rate definitions, e.g., the generalized FWER, are discussed in Chapter 7.

2.5.2 Notation

The following notation will be used in this section. As before, $H_1, \ldots, H_m$ denote the hypotheses of interest. We will assume throughout Sections 2.6–2.8 that the m hypotheses are tested under the free combination condition; i.e., no logical dependencies exist among the hypotheses. The only two exceptions are the problems considered in Section 2.6.2 (Shaffer procedure) and Section 2.7.3 (extended Shaffer-Royen procedure).

When the hypotheses are not equally important, the weights, $w_1, \ldots, w_m$, are introduced to quantify their importance (each weight is between 0 and 1 and the weights add up to 1). Weighted hypotheses are encountered, for example, in dose-finding trials. The trial's sponsor can assign weights to the dose-placebo comparisons according to the expected effect size at each dose level to improve the overall power of the multiple test.

The test statistics associated with the hypotheses are denoted by $t_1, \ldots, t_m$. Let p_i be the p-value computed from the null distribution of t_i, $i = 1, \ldots, m$. These p-values are frequently called *raw p-values* to distinguish them from multiplicity adjusted p-values. The ordered p-values are denoted by $p_{(1)} < \ldots < p_{(m)}$ and the associated hypotheses are denoted by $H_{(1)}, \ldots, H_{(m)}$.

2.5.3 Dose-finding trial example

To illustrate the use of multiple testing procedures in pharmaceutical applications, we will use the following example. Consider a dose-finding trial in

TABLE 2.1: Summary of the mean increase in HDL cholesterol (mg/dl) in the dose-finding trial under three scenarios (mean difference, standard error, lower limit of the one-sided 97.5% confidence limit, two-sample t statistic and raw one-sided p-value). The asterisk identifies the p-values that are significant at the 0.025 level.

Test	Mean	Standard error	Lower confidence limit	t statistic	P-value
			Scenario 1		
D1-Placebo	2.90	1.44	0.07	2.01	0.0228*
D2-Placebo	3.14	1.44	0.31	2.17	0.0152*
D3-Placebo	3.56	1.44	0.73	2.46	0.0071*
D4-Placebo	3.81	1.44	0.98	2.64	0.0043*
			Scenario 2		
D1-Placebo	2.60	1.45	−0.23	1.80	0.0364
D2-Placebo	2.73	1.45	−0.10	1.89	0.0297
D3-Placebo	3.45	1.45	0.61	2.38	0.0088*
D4-Placebo	3.57	1.45	0.74	2.47	0.0070*
			Scenario 3		
D1-Placebo	3.10	1.45	0.27	2.15	0.0162*
D2-Placebo	3.35	1.45	0.52	2.32	0.0105*
D3-Placebo	3.69	1.45	0.86	2.55	0.0055*
D4-Placebo	2.67	1.45	−0.17	1.85	0.0329

patients with dyslipidemia. The trial will be conducted to compare the effect of four doses of the drug, labeled D1 (lowest dose) through D4 (highest dose), to that of a placebo. The primary efficacy endpoint is based on the mean increase in HDL cholesterol at 12 weeks. The sample size in each treatment group is 77 patients.

Table 2.1 displays the mean treatment effects of the four doses compared to placebo, associated standard errors, lower limits of one-sided 97.5% confidence intervals, t statistics based on the two-sample t test with a pooled variance computed from all treatment groups and raw one-sided p-values. The table includes three scenarios that represent three different dose-response relationships in this trial. These scenarios will be used to evaluate the performance of the multiple testing procedures described in Sections 2.6–2.8.

The mean treatment differences with one-sided 97.5% confidence intervals in the three scenarios are plotted in Figure 2.2. Key features of the three dose-response functions are summarized below:

- Scenario 1. The dose-response function increases over the dose range and the drug effect is present in all dose groups (all doses are superior to placebo at 0.025).

- Scenario 2. The dose-response function increases over the dose range

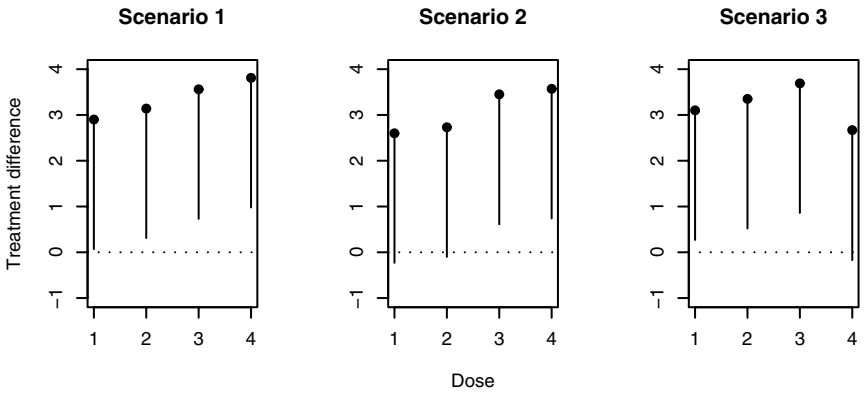

FIGURE 2.2: Mean treatment differences in HDL cholesterol change from baseline to 12 weeks with one-sided 97.5% confidence intervals in the dose-finding trial example.

but the drug effect is present only at the higher doses (D3 and D4 are superior to placebo but D1 and D2 are not).

- Scenario 3. The dose-response function is generally flat at the lower three doses with a drop at the highest dose due to tolerability problems (D1, D2 and D3 are superior to placebo but D4 does not separate from placebo).

It should be emphasized that we use the numerical example from Table 2.1 mainly to illustrate the multiple testing procedures described in Sections 2.6–2.8. Chapter 3 describes alternative analysis strategies based on multiple testing, modeling dose-response functions or a combination of multiple testing and modeling, which are usually more appropriate in the context of dose-finding studies.

2.6 Multiple testing procedures based on univariate p-values

2.6.1 Bonferroni procedure

The Bonferroni procedure is a widely used single-step procedure commonly attributed to Sir Ronald Fisher. In the problem of testing m equally weighted

hypotheses, $H_1, \ldots, H_m$, the Bonferroni procedure rejects H_i if $p_i \leq \alpha/m$. Due to the (first-order) Bonferroni inequality, this procedure controls the FWER for any joint distribution of the raw p-values.[1]

As an illustration, consider Scenario 1 of the dose finding trial example given in Section 2.5.3. The Bonferroni-adjusted significance level is $\alpha/4 = 0.00625$ and thus the D4-Placebo test is significant at this level whereas the other three tests are not.

The Bonferroni procedure tends to be rather conservative if the number of hypotheses is large or the test statistics are strongly positively correlated. Figure 2.3 displays the actual Type I error rate of the Bonferroni procedure in multiple testing problems with $m = 2$ and 5 comparisons when the error rate is controlled at the one-sided 0.025 level. The test statistics are assumed to be equally correlated and follow a multivariate normal distribution. The common correlation coefficient is denoted by ρ ($-1 < \rho \leq 1$ in the two-dimensional case and $-1/4 < \rho \leq 1$ in the five-dimensional case). The probability of a Type I error is evaluated under the global null hypothesis (all hypotheses are true) based on 1,000,000 simulation runs. With $m = 2$ comparisons, the error rate is very close to the nominal level when $\rho \leq 0.3$ and becomes severely deflated when the test statistics are strongly positively correlated ($\rho \geq 0.8$). In the case of $m = 5$ comparisons, the actual error rate is below 0.02 even when the test statistics are moderately positively correlated (ρ is around 0.6).

2.6.2 Holm procedure and its extensions

The Holm (1979) procedure is a popular multiple testing procedure that demonstrates the advantages of a stepwise testing method.

Assume first that the hypotheses are equally weighted. The Holm procedure is a step-down procedure that starts with the hypothesis associated with the most significant p-value and rejects it if the p-value is no greater than α/m. If the first ordered hypothesis is rejected, the Holm procedure examines the next hypothesis in the sequence and so on. In general, this procedure is based on the following algorithm:

- Step 1. If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$ and go to the next step. Otherwise retain all hypotheses and stop.
- Steps $i = 2, \ldots, m-1$. If $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(i)}$ and go to the next step. Otherwise retain $H_{(i)}, \ldots, H_{(m)}$ and stop.
- Step m. If $p_{(m)} \leq \alpha$, reject $H_{(m)}$. Otherwise retain $H_{(m)}$.

This stepwise procedure is more powerful than the single-step Bonferroni procedure because it begins at the same significance level as the Bonferroni procedure (α/m) and tests the other hypotheses at successively higher levels.

[1] Although the Bonferroni inequality is named after the Italian mathematician Carlo Emilio Bonferroni, it is worth noting that Bonferroni's research focused on refining this inequality that actually goes back to the work of the British mathematician George Boole.

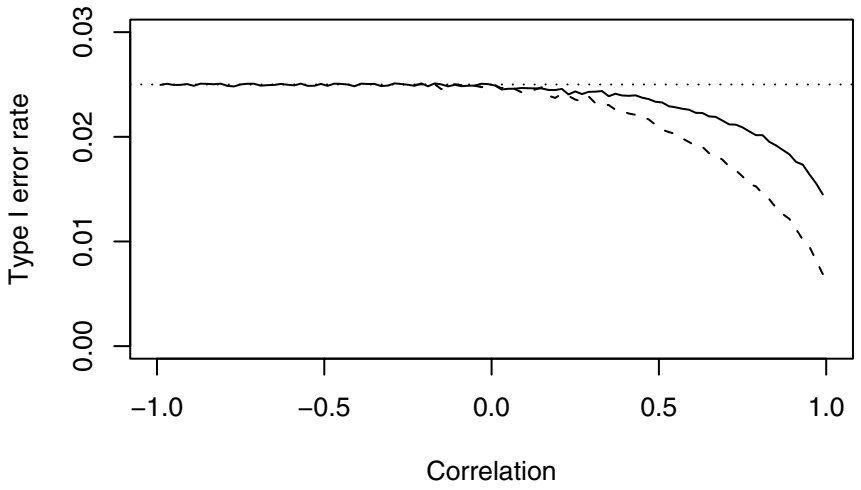

FIGURE 2.3: Type I error rate of the Bonferroni procedure under the global null hypothesis as a function of the number of comparisons and correlation (solid curve, $m = 2$ comparisons, correlation > -1; dashed curve, $m = 5$ comparisons, correlation > -0.25). The Bonferroni procedure is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025.

In the case of unequally weighted hypotheses, the Holm procedure can be defined based on the algorithms proposed by Holm (1979) or Benjamini and Hochberg (1997).

In order to prove that the Holm procedure controls the FWER, one can use a direct approach, as was done by Holm (1979), or utilize the closure principle introduced in Section 2.3.3. It is easy to show that, under the free combination condition, the Holm procedure is, in fact, a closed testing procedure in which each intersection hypothesis is tested using a global test based on the Bonferroni procedure. Since the Bonferroni procedure cannot reject more hypotheses than the Holm procedure, the latter serves as an example of how the power of a multiple testing procedure can be enhanced in a uniform manner by applying the closure principle. Note, however, that the Holm procedure rejects the global hypothesis if and only if the Bonferroni procedure does and therefore the conclusions regarding the conservative nature of the Bonferroni procedure (see Figure 2.3) also apply to the Holm procedure.

To compare the performance of the Holm procedure to that of the Bonferroni procedure, consider Scenario 1 in the dose-finding trial example (Sec-

tion 2.5.3). The ordered p-values are given by

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

At the first step of the Holm procedure, the smallest p-value, $p_{(1)}$, is compared with $\alpha/4 = 0.00625$. Since $p_{(1)} < 0.00625$, the corresponding hypothesis of no treatment effect is rejected and thus Dose D4 is superior to Placebo. Next, $p_{(2)}$ is compared to $\alpha/3 = 0.00833$ and the associated hypothesis is rejected (Dose D3 is superior to Placebo). Next, $p_{(3)}$ is greater than the critical value, $\alpha/2 = 0.0125$, and therefore the Holm procedure retains the remaining two hypotheses (Doses D1 and D2 do not separate from Placebo).

Shaffer procedure

Shaffer (1986) introduced a version of the Holm procedure for multiple testing problems with logical dependencies. Logical dependencies exist when truth of a subset of hypotheses necessarily implies truth of some other hypotheses. The most common example of logical dependencies occurs with all pairwise comparisons. For example, in the context of the dose-finding trial example introduced in Section 2.5.3, let μ_0 denote the mean improvement in the placebo group and μ_i denote the mean improvement in the ith dose group, $i = 1, 2, 3, 4$. The hypotheses of interest are defined as follows:

$$H_{ij} : \ \mu_i = \mu_j.$$

Suppose that H_{12} and H_{13} are true, then it logically follows that H_{23} is also true. On the other hand, there are no logical dependencies when dose-placebo comparisons are considered. If H_{01} and H_{02} are true, H_{03} is not necessarily true.

When there are logical dependencies among the hypotheses, the divisors $(m-i+1)$ in the Holm procedure may be replaced by divisors k_i, where k_i is the maximum number of the hypotheses $H_{(i)}, \ldots, H_{(m)}$ that can be simultaneously true given that $H_{(1)}, \ldots, H_{(i-1)}$ are false. Specifically, the Shaffer procedure rejects the hypothesis $H_{(i)}$, $i = 1, \ldots, m$, at the ith step if

$$p_{(j)} \leq \frac{\alpha}{k_j}, \quad j = 1, \ldots, i.$$

In the dose-finding trial example introduced in Section 2.5.3, there are 10 pairwise comparisons of interest and thus $k_1 = 10$. This means that the critical value for the smallest p-value, $p_{(1)}$, is equal to that used by the Holm procedure, i.e., $\alpha/10$. However, at the second step $k_2 = 6$. This represents a substantial improvement over the remaining number of comparisons, i.e., $(10 - 2 + 1) = 9$, needed for the Holm procedure at the second step. Note that, when there are no logical dependencies among hypotheses, the Shaffer procedure reduces to the regular Holm procedure.

Shaffer developed two methods, Method 1 (described above) and Method

2, which uses the sequence of hypotheses $H_{(1)}, \ldots, H_{(m)}$ corresponding to the specific ordered p-values $p_{(1)}, \ldots, p_{(m)}$ observed in the study. The divisors l_i for Shaffer's Method 2 satisfy $l_i \leq k_i$; hence Method 2 is uniformly more powerful than Method 1. Like the Holm procedure, the Method 2 procedure is a type of closed testing procedure based on the Bonferroni procedures for each intersection hypothesis. The procedure is set up as follows: closed testing is performed, in sequence, for $H_{(1)}, \ldots, H_{(m)}$ and testing stops at the first non-significant outcome. Thus Shaffer's Method 2 is called a *truncated closed testing procedure*. Truncation ensures that the procedure is monotone; i.e., $H_{(j)}$ cannot be rejected if $H_{(i)}$ is not rejected and $i < j$ (see Westfall and Tobias (2007) for details).

Shaffer's method has been modified recently to account for dependence structures as noted by Westfall and Tobias (2007); see Section 2.7.3. This class of methods is especially useful for pairwise comparisons, which are uncommon in Phase III clinical trials but used more frequently in early-phase studies, and in general for comparisons that are logically intertwined, such as for non-pairwise comparisons using trend tests applied to dose-response analyses (see Chapter 3).

2.6.3 Fixed-sequence procedure

The fixed-sequence testing approach (Maurer et al., 1995; Westfall and Krishen, 2001) has found a variety of applications in clinical trials due to its straightforward stepwise form. The fixed-sequence procedure assumes that the order in which the hypotheses are tested, $H_1, \ldots, H_m$, is pre-specified (this order normally reflects the clinical importance of the multiple analyses). Testing begins with the first hypothesis, H_1, and each test is carried out without a multiplicity adjustment as long as significant results are observed in all preceding tests. In other words, the hypothesis H_i, $i = 1, \ldots, m$, is rejected at the ith step if

$$p_j \leq \alpha, \quad j = 1, \ldots, i.$$

The fixed-sequence procedure controls the FWER because, for each hypothesis, testing is conditional upon rejecting all hypotheses earlier in the sequence.

To demonstrate how a fixed-sequence strategy can be used in a clinical study, we will use the dose-finding trial example described in Section 2.5.3. It may be reasonable to order the doses from D4 (highest dose) to D1 (lowest dose) since the higher doses are generally more likely to produce a significant treatment effect than the lower doses. In all three scenarios defined in Section 2.5.3, the fixed-sequence procedure starts with the D4-Placebo comparison, proceeds to the next comparison in the sequence if the D4-Placebo statistic is significant at the one-sided 0.025 level and so on. Consider, for example, Scenario 1. Since all p-values are significant at 0.025 in this scenario, the fixed-sequence procedure rejects all hypotheses of no treatment effect.

2.6.4 Fallback procedure

The Holm and fixed-sequence procedures described in Sections 2.6.2 and 2.6.3 represent two different approaches to carrying out multiple testing procedures. In the case of the Holm procedure, testing is performed in a data-driven order. By contrast, the fixed-sequence procedure uses an *a priori* specified testing sequence. A compromise between the two testing approaches can be achieved by utilizing the fallback procedure introduced by Wiens (2003) and further studied by Wiens and Dmitrienko (2005), Dmitrienko, Wiens and Westfall (2006), Hommel, Bretz and Maurer (2007), Hommel and Bretz (2008) and Bretz et al. (2009b).

To introduce the fallback procedure, suppose the hypotheses $H_1, \ldots, H_m$ are ordered and allocate the overall error rate α among the hypotheses according to their weights $w_1, \ldots, w_m$ (we will consider the general version of this procedure because it was designed specifically for the case of unequal weights). Specifically, the amount of the overall error rate assigned to H_i is equal to αw_i, $i = 1, \ldots, m$. This process is similar to allocating the overall α among the hypotheses in the weighted Bonferroni procedure. The fallback procedure is carried out as follows:

- Step 1. Test H_1 at $\alpha_1 = \alpha w_1$. If $p_1 \leq \alpha_1$, reject this hypothesis; otherwise retain it. Go to the next step.

- Steps $i = 2, \ldots, m-1$. Test H_i at $\alpha_i = \alpha_{i-1} + \alpha w_i$ if H_{i-1} is rejected and at $\alpha_i = \alpha w_i$ if H_{i-1} is retained. If $p_i \leq \alpha_i$, reject H_i; otherwise retain it. Go to the next step.

- Step m. Test H_m at $\alpha_m = \alpha_{m-1} + \alpha w_m$ if H_{m-1} is rejected and at $\alpha_m = \alpha w_m$ if H_{m-1} is retained. If $p_m \leq \alpha_m$, reject H_m; otherwise retain it.

It is instructive to compare the fallback and Holm procedures. Unlike the Holm procedure, the fallback procedure can continue testing even if a non-significant outcome is encountered by utilizing the fallback strategy (this explains why it is called the *fallback* procedure). If a hypothesis is retained, the next hypothesis in the sequence is tested at the level that would have been used by the weighted Bonferroni procedure. It was shown by Wiens and Dmitrienko (2005) that the fallback procedure is a closed testing procedure and thus it controls the FWER in the strong sense.

The fallback procedure is uniformly more powerful than the weighted Bonferroni procedure based on the same set of weights. In addition, the fallback procedure simplifies to the fixed-sequence procedure when $w_1 = 1$ and $w_2 = \ldots = w_m = 0$. Wiens and Dmitrienko (2005) suggested that the fallback procedure can be thought of as a compromise between the fixed-sequence and Hommel procedures (the Hommel procedure will be introduced in Section 2.6.8).

It was pointed out in Section 2.3.4 that the fallback procedure is not α-exhaustive; i.e., when it is cast as a closed testing procedure, not all intersection hypotheses are tested at the full α level. This means that one can construct a procedure that is uniformly more powerful than the fallback procedure and maintains the FWER at the same level. Wiens and Dmitrienko (2005) discussed several approaches to defining extended fallback procedures of this kind. In Section 2.6.5 we describe further properties of the regular and extended fallback procedures and discuss graphical tools for their visualization.

Using Scenario 1 in the dose-finding trial example (Section 2.5.3), we will demonstrate how to apply the fallback procedure to a multiple testing problem with equal weights (the four dose-placebo tests are equally weighted, i.e., $w_1 = \ldots = w_4 = 1/4$). The fallback procedure begins with the comparison of Dose D4 to placebo and tests the associated hypothesis H_4 at $\alpha/4 = 0.0063$. This hypothesis is rejected since $p_4 < 0.0063$. The α level at which H_4 was tested is carried over to the next hypothesis in the sequence, H_3 (D3-Placebo comparison). This hypothesis is tested at $2\alpha/4 = 0.0125$. Note that $p_3 < 0.0125$ and thus the fallback procedure rejects H_3, which means that $\alpha/2$ is carried over to H_2 (D2-Placebo comparison). This hypothesis is rejected at $3\alpha/4 = 0.0188$ and the last hypothesis in the sequence, H_1, is tested at $4\alpha/4 = 0.025$. This hypothesis is also rejected by the fallback test.

2.6.5 Bonferroni-based closed testing procedures

In this section we show that the multiple testing procedures described in Sections 2.6.1–2.6.4 are all closed testing procedures based on the (weighted) Bonferroni test and thus follow the same construction principle. Understanding the closure principle (Section 2.3.3) enables one to take full advantage of its flexibility and to tailor the multiple testing procedure to the study objectives. In the following we will

- describe the class of Bonferroni-based closed testing procedures;
- give a sufficient characterization to derive sequentially rejective multiple testing procedures and demonstrate that many common procedures are in fact special cases thereof;
- construct simultaneous confidence intervals for procedures in the class;
- provide graphical tools that facilitate the derivation and communication of Bonferroni-based closed testing procedures based on sequentially rejective rules that are tailored to study objectives.

Because this section provides a general perspective of the methods described previously, the description is slightly more technical. In order to keep this section at a reasonable size, we omit the technical details and refer to the original publications instead.

Class of Bonferroni-based closed testing procedures

As before, consider the problem of testing m hypotheses $H_1, \ldots, H_m$ and let $I = \{1, \ldots, m\}$ denote the associated index set. Recall from Section 2.3.3 that applying the closure principle leads to consideration of the intersection hypotheses $H_J = \bigcap_{j \in J} H_j$. For each intersection hypothesis H_J we assume a collection of non-negative weights $w_j(J)$ such that they sum to 1, that is, $0 \leq w_j(J) \leq 1$ and $\sum_{j \in J} w_j(J) = 1$. These weights quantify the relative importance of the hypotheses H_j included in the intersection H_J. As before, let p_j denote the raw p-value for H_j, $j \in I$.

In this section we assume that each intersection hypothesis is tested with a weighted Bonferroni test. Consequently, we obtain the multiplicity adjusted p-values

$$p_J = \min\{q_j(J) : \ j \in J\}$$

for the weighted Bonferroni test for H_J, where

$$q_j(J) = \begin{cases} p_j / w_j(J) & \text{if } w_j(J) > 0, \\ 1 & \text{if } w_j(J) = 0. \end{cases}$$

This defines Class $\mathcal{B}$ of all closed testing procedures that use weighted Bonferroni tests for each intersection hypothesis. Any collection of weights subject to the constraints given above can be used and thus one can choose the weights and tailor the closed testing procedure to the given study objectives.

To illustrate this, consider the simple two-hypothesis problem from Section 2.3.3. Consider the intersection hypothesis H_J with $J = \{1, 2\}$ and associated weights $w_1(J) = w_2(J) = 1/2$. This results in the regular Bonferroni test and the adjusted p-value $p_J = 2\min(p_1, p_2)$. If $H_{\{1,2\}} = H_1 \cap H_2$ is rejected, so is either H_1 or H_2, since they are tested subsequently at level α. In other words, if $H_{\{1,2\}}$ is rejected (the smaller of the two p-values is less than $\alpha/2$), the remaining elementary hypothesis is tested at level α, which is exactly the Holm procedure described in Section 2.6.2. Similarly, one can show that the Shaffer procedure (Section 2.6.2), fixed-sequence procedure (Section 2.6.3), fallback procedure (Section 2.6.4) and all Bonferroni-based gatekeeping procedures (Chapter 5) are examples of multiple testing procedures from Class $\mathcal{B}$ (Hommel, Bretz and Maurer, 2007).

Sequentially rejective Bonferroni-based closed testing procedures

It can further be shown that under a mild monotonicity condition on the weights $w_j(J)$ the closure principle leads to powerful consonant multiple testing procedures (see Section 2.3.4 for the definition of consonance). Short-cut versions can thus be derived, which substantially simplify the implementation and interpretation of the related procedures. Hommel, Bretz and Maurer (2007) showed that all the procedures mentioned previously (with the notable

exception for the Shaffer procedure) belong to a subclass $\mathcal{S} \subset \mathcal{B}$ of shortcut procedures characterized by the property

$$w_j(J) \leq w_j(J') \text{ for all } J' \subseteq J \subseteq I \text{ and } j \in J'.$$

This condition ensures that if an intersection hypothesis H_J is rejected, there is an index $j \in J$, such that $p_j/w_j(J) \leq \alpha$ and the corresponding elementary hypothesis H_j can be rejected immediately by the closed testing procedure. Therefore, short-cut procedures of order m can be constructed; i.e., instead of testing $2^m - 1$ hypotheses (as usually required by the closure principle), it is sufficient to test the elementary hypotheses $H_1, \ldots, H_m$ in m steps. This simplification is a key characterization of the Holm procedure and the results from Hommel, Bretz and Maurer (2007) ensure that this remains true for *any* procedure in $\mathcal{S}$. As a consequence, shortcut procedures from $\mathcal{S}$ can be carried out with the following m-step procedure. Start testing the global intersection hypothesis $H_I, I = \{1, \ldots, m\}$. If it is rejected, there is an index $i \in I$ as described above such that H_i is rejected by the closed testing procedure. At the next step, one continues testing the global intersection $H_{I \setminus i}$ of the remaining, not yet rejected hypotheses, and so on, until the first non-rejection.

Simultaneous confidence intervals

The previous characterization for Class $\mathcal{S}$ can also be used to construct compatible simultaneous confidence intervals introduced in Section 2.4.2 (Strassburger and Bretz, 2008; Guilbaud, 2008). Consider the one-sided null hypotheses $H_i : \ \theta_i \leq \delta_i$, $i \in I = \{1, \ldots, m\}$, where $\theta_1, \ldots, \theta_m$ are the parameters of interest and $\delta_1, \ldots, \delta_m$ are pre-specified constants (e.g., noninferiority margins). Let $\alpha_j(J) = \alpha w_j(J)$ denote the local significance levels with $j \in J \subseteq I$. Further, let $L_i(\bar{\alpha})$ denote the marginal lower confidence limit for θ_i at level $1 - \bar{\alpha}$, $1, \ldots, m$. Finally, let R denote the index set of hypotheses rejected by a multiple testing procedure from $\mathcal{S}$. Then, lower one-sided confidence limits for $\theta_1, \ldots, \theta_m$ with coverage probability of at least $1 - \alpha$ are given by

$$\widetilde{L}_i = \begin{cases} \delta_i & \text{if } i \in R \text{ and } R \neq I, \\ L_i(\bar{\alpha}_i) & \text{if } i \notin R, \\ \max(\delta_i, L_i(\bar{\alpha}_i)) & \text{if } R = I, \end{cases}$$

where $\bar{\alpha}_i = \alpha_i(I \setminus R)$ if $i \notin R \neq I$. In the case $R = I$, where all hypotheses are rejected, the choice of the local levels $\bar{\alpha}_i = \alpha_i(\emptyset)$ is arbitrary (Strassburger and Bretz, 2008). Thus, in order to compute the simultaneous confidence limits, one needs to know only the set R of rejected hypotheses and the corresponding local levels $\bar{\alpha}_i$ for all indices i of retained hypotheses. Note that if not all hypotheses are rejected ($R \neq I$), the confidence limits associated with the rejected hypotheses ($i \in R$) essentially reflect the test decision $\theta_i > \delta_i$ and the confidence limits associated with the retained hypotheses are the marginal

confidence limits at level $\alpha_i(I \setminus R)$. This method will be used to derive simultaneous confidence intervals for the Bonferroni, Holm, fixed-sequence and fallback procedures in Section 2.6.11.

Graphical visualization

It was shown above that Class $\mathcal{S}$ includes a variety of Bonferroni-based testing procedures, such as fixed-sequence, fallback and gatekeeping procedures. Using procedures in this class, one can map the difference in importance as well as the relationship between various study objectives onto a suitable multiple test procedure. However, since the procedures are based on the closure principle, one needs to specify the weights $w_j(J)$ for each of the $2^m - 1$ intersection hypotheses $H_J, J \subseteq I$. Unless these weights follow some simple and well-known specification rules (such as, for example, in the Holm procedure), the underlying test strategy may be difficult to communicate to clinical trial teams.

Graphical tools have been proposed instead, which help visualizing different sequentially rejective test strategies and thus to best tailor a multiple testing procedure to given study objectives (Bretz et al., 2009b). Using a graphical approach, the hypotheses $H_1, \ldots, H_m$ are represented by vertices with associated weights denoting the local significance levels $\alpha_1, \ldots, \alpha_m$. The weight associated with a directed edge between any two vertices indicates the fraction of the (local) significance level that is shifted if the hypothesis at the tail of the edge is rejected.

To help illustrate this concept, consider a problem with three hypotheses H_1, H_2 and H_3. The top left panel in Figure 2.4 displays a graphical representation of the fallback procedure introduced in Section 2.6.4. Each of the hypotheses is assigned an associated local significance level α_i, such that $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$. If H_1 is rejected, then the level α_1 is carried over to H_2, as indicated by the edge pointing from H_1 to H_2. If H_2 is rejected at its local significance level (either α_2 or $\alpha_1 + \alpha_2$), then that level is carried over to H_3, as indicated by the edge pointing from H_2 to H_3.

It is important to note that graphical tools of this kind also help derive other, potentially more powerful testing strategies. Returning to the top left panel in Figure 2.4, one can see that, if H_3 is rejected, its local significance level is not carried over to any other hypothesis. As shown in the top right panel, this significance level can be re-used by adding two further edges pointing back to H_1 and H_2, where $r = \alpha_2/(\alpha_1 + \alpha_2)$. The resulting testing procedure is equivalent to the α-exhaustive extension of the fallback procedure considered by Wiens and Dmitrienko (2005). Further, the bottom panel in Figure 2.4 displays yet another extension of the fallback procedure by shifting the significance level to the first hypothesis in the hierarchy that was not rejected so far (Hommel and Bretz, 2008). Here, ϵ denotes an infinitesimally small weight indicating that the significance level is carried over from H_2 to H_3 only if both H_1 and H_2 are rejected. The motivation for this extension is that H_1

is deemed more important than H_3. Thus, once H_2 is rejected, its associated significance level should be carried over first to H_1 before continuing to testing H_3.

We refer to Bretz et al. (2009b) for a detailed description of these ideas along with further extensions and examples, including a description of algorithms to derive the (updated) weights, simultaneous confidence intervals and adjusted p-values.

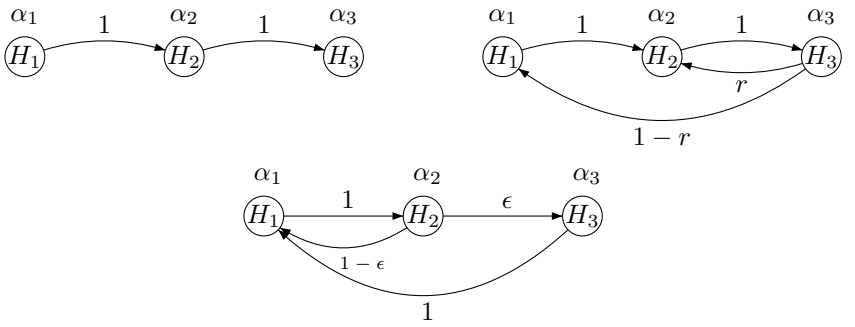

FIGURE 2.4: Graphical illustration of the fallback procedure (top left) and two extensions.

2.6.6 Reverse fixed-sequence procedure

The fixed-sequence procedure introduced in Section 2.6.3 has a sequentially rejective structure in the sense that each hypothesis can be rejected only if all previously examined hypotheses are rejected. The reverse fixed-sequence procedure is a mirror image of the regular fixed-sequence procedure.

Consider m ordered hypotheses $H_1, \ldots, H_m$. The reverse fixed-sequence procedure begins with the first hypothesis in the sequence, H_1, and tests it at the full α level. If the hypothesis is rejected, the other hypotheses are automatically rejected and the testing algorithm terminates. Otherwise, H_1 is retained and the next hypothesis is tested. At the ith step of the algorithm, the procedure retains $H_1, \ldots, H_{i-1}$ and rejects $H_i, \ldots, H_m$ if $p_j > \alpha$ for $j = 1, \ldots, i-1$ and $p_i \leq \alpha$, $i = 2, \ldots, m$. This procedure controls the FWER in the strong sense if $H_1, \ldots, H_m$ form a sequence of nested hypotheses; i.e., H_i is a subset of H_j if $i > j$. Therefore, rejection of H_j implies rejection of all H_i's for $i > j$.

To compute adjusted p-values for the reverse fixed-sequence procedure, note that $p_i \geq p_j$ if $i > j$ when the hypotheses are nested (in other words, it is easier to reject H_j compared to H_i when $i > j$). Therefore, the adjusted p-value for H_i is simply equal to p_i; i.e., $\widetilde{p}_i = p_i$, $i = 1, \ldots, m$.

Nested hypotheses are encountered in clinical trials with noninferiority and

superiority objectives. Specifically, consider a trial for testing the efficacy of a treatment versus a control with respect to a single endpoint. Let δ denote an appropriate measure of the treatment difference (for example, the mean difference if the endpoint is continuous) and assume that a positive treatment difference indicates improvement. The trial's sponsor is interested in testing the hypotheses of noninferiority and superiority. The noninferiority hypothesis is defined as

$$H_1 : \ \delta \leq -\gamma,$$

where γ is a pre-specified positive noninferiority margin. The superiority hypothesis is given by:

$$H_2 : \ \delta \leq 0.$$

The two hypotheses can be tested sequentially using the fixed-sequence procedure:

- Begin with the noninferiority test and test H_1 at the α level.
- If noninferiority is established (H_1 is rejected), switch to the superiority test, which is also carried out at the α level.

This sequentially rejective procedure was described, among others, by Morikawa and Yoshida (1995). Morikawa and Yoshida pointed out that the reverse fixed-sequence procedure can also be applied in this problem due to the fact that H_1 is a subset of H_2. The reverse fixed-sequence procedure is carried out as follows:

- Begin with the superiority test at the α level.
- If superiority cannot be established (H_2 is not rejected), carry out the noninferiority test at the α level.

2.6.7 Simes global test

In this and the next two sections we will introduce the Simes global test and multiple testing procedures derived from the Simes test, including the Hochberg, Rom and Hommel procedures that will be defined later in this section.

The Simes test (Simes, 1986) focuses on testing the global hypothesis of no treatment effect; i.e.,

$$H_I = \bigcap_{i=1}^{m} H_i.$$

It rejects H_I if

$$p_{(i)} \leq i\alpha/m \text{ for at least one } i = 1, \ldots, m,$$

where $p_{(1)} < \ldots < p_{(m)}$ are the ordered p-values.

Note that the Simes test makes use of all ordered p-values (not just the smallest p-value) to test the global hypothesis and thus it is more powerful than a global test based on the Bonferroni procedure. It is also important to note that, unlike the Bonferroni procedure, the Simes test cannot be directly used to test the individual hypotheses $H_1, \ldots, H_m$. In particular, one cannot reject $H_{(i)}$ if $p_{(i)} \leq i\alpha/m$, $i = 1, \ldots, m$, since the FWER is not controlled in this case (Hommel, 1988).

Simes proved that this global test is exact in the sense that its size equals α if $p_1, \ldots, p_m$ are independent. Since the assumption of independence is unlikely to be met in practice, several authors examined operating characteristics of this test under dependence. Hommel (1988) showed that the use of the Simes test can lead to an inflated Type I error probability. However, the worst-case scenario considered by Hommel corresponds to an extreme case that is very unlikely to be encountered in pharmaceutical applications. Hochberg and Rom (1995) examined the Type I error rate of the Simes test in the case of negatively-correlated normal variables and employed a simulation study to demonstrate that the Type I error rate is slightly inflated (about 10% inflation in the worst case). Samuel-Cahn (1996) showed via simulations that the Simes test preserves the Type I error rate in a one-sided setting with positively-correlated test statistics and in a two-sided setting regardless of the correlation. However, the test becomes anticonservative in the one-sided case if the test statistics are negatively correlated. Sarkar and Chang (1997) and Sarkar (1998) proved that the test protects the Type I error rate when the joint distribution of the test statistics exhibit a certain type of positive dependence; i.e., when the the joint distribution is multivariate totally positive of order two (Karlin and Rinott, 1980)[2]. They showed that this condition is met in studies with multiple treatment-control comparisons under normal assumptions.

Figure 2.5 depicts the relationship between the Type I error rate of the Simes test, number of comparisons $m = 2$, 5 and common correlation coefficient ρ in the case of normally distributed test statistics. The calculations are performed under the same assumptions as in Figure 2.3. Comparing the Type I error rate to that of the Bonferroni procedure (see Figure 2.3), it is easy to see that the Simes test performs better than the Bonferroni procedure under both weak and strong positive correlations. For example, the error rate in Figure 2.5 is only slightly below the nominal 0.025 level for both $m = 2$ and $m = 5$ comparisons when $\rho \leq 0.4$ and is close to 0.025 when ρ approaches 1.

[2]It is worth pointing out that positive dependence is a more restrictive condition than positive correlation. However, in the case of a multivariate normal distribution the latter generally implies the former. For example, the positive dependence condition is satisfied when the test statistics follow a multivariate normal distribution with equal correlations and the common correlation is non-negative. In addition, this condition is satisfied in the multivariate normal case if all partial correlations are non-negative (Bolviken, 1982; Karlin and Rinott, 1983).

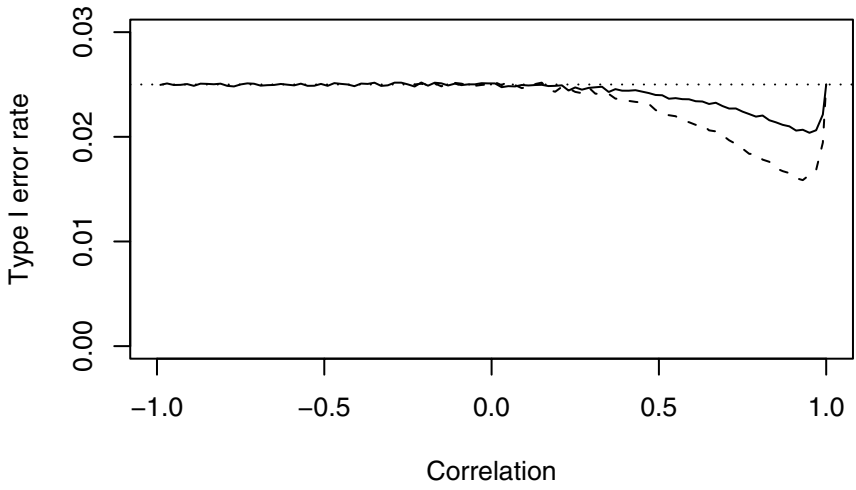

FIGURE 2.5: Type I error rate of the Simes test under the global null hypothesis as a function of the number of comparisons and correlation (solid curve, $m = 2$ comparisons, correlation > -1; dashed curve, $m = 5$ comparisons, correlation > -0.25). The Simes test is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025.

2.6.8 Hommel procedure

It was explained in Section 2.6.2 that the Holm procedure results from using a global test based on the Bonferroni for testing intersection hypotheses in a closed procedure. Similarly, the Hommel procedure (Hommel, 1988) results from using the Simes global test for testing individual intersection hypotheses. In the case of equally weighted hypotheses, the Hommel procedure can be applied using the following algorithm:

- Step 1. If $p_{(m)} > \alpha$, retain $H_{(m)}$ and go to the next step. Otherwise reject all hypotheses and stop.
- Steps $i = 2, \ldots, m-1$. If $p_{(m-j+1)} > (i-j+1)\alpha/i$ for $j = 1, \ldots, i$, retain $H_{(m-i+1)}$ and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step m. If $p_{(m-j+1)} > (i-j+1)\alpha/i$ for $j = 1, \ldots, m$, retain $H_{(1)}$; otherwise reject it.

The Hommel procedure is easily extended to problems with unequally weighted hypotheses (Hommel, 1988). It protects the FWER under conditions that guarantee Type I error rate control for the Simes global test. It is uniformly more powerful than the Holm procedure because the Simes test is uniformly more powerful than the global test based on the Bonferroni procedure. For example, the Holm procedure rejects $H_{(1)}$ if and only if $p_{(1)} \leq \alpha/m$ whereas the Hommel procedure can reject this hypothesis when $p_{(1)} > \alpha/m$, e.g., $H_{(1)}$ is rejected if $p_{(m)} \leq \alpha$.

Using the dose-finding trial example from Section 2.5.3, we will illustrate the application of the stepwise algorithm defined above. Beginning with Scenario 1, note that all unadjusted p-values are less than 0.025, which immediately implies that the Hommel procedure rejects all hypotheses of no treatment effect. Now consider Scenario 2. The ordered p-values are given by

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

Since the largest p-value, $p_{(4)}$, is greater than 0.025, the Hommel procedure retains the hypothesis $H_{(4)}$. At the second step of the procedure, $p_{(3)} > 0.025/2$ and $p_{(4)} > 2(0.025/2)$, which means that $H_{(3)}$ is also retained. Further, the Hommel procedure retains $H_{(2)}$ since

$$p_{(2)} > 0.025/3, \quad p_{(3)} > 2(0.025/3), \quad p_{(4)} > 3(0.025/3).$$

Finally, note that $p_{(1)} \leq 0.025/4$ and thus $H_{(1)}$ is rejected by the procedure.

2.6.9 Hochberg procedure

The Hochberg procedure (Hochberg, 1988) is another popular procedure based on the Simes global test. As was mentioned in Section 2.5.1, the Hochberg procedure is an example of a step-up procedure based on univariate p-values. Unlike step-down procedures (e.g., the Holm procedure), this procedure begins with the least significant p-value and examines the other p-values in a sequential manner until it reaches the most significant one.

Beginning with the case of equally weighted hypotheses, the decision rule for the Hochberg procedure is defined as follows:

- Step 1. If $p_{(m)} > \alpha$, retain $H_{(m)}$ and go to the next step. Otherwise reject all hypotheses and stop.
- Steps $i = 2, \ldots, m-1$. If $p_{(m-i+1)} > \alpha/i$, retain $H_{(m-i+1)}$ and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step m. If $p_{(1)} > \alpha/m$, retain $H_{(1)}$; otherwise reject it.

Extensions of the Hochberg procedure to the case of unequally weighted hypotheses were discussed in Tamhane and Liu (2008).

The Hochberg procedure controls the FWER under the same conditions

for which the Simes global test controls the Type I error rate. Further, this procedure is uniformly more powerful than the Holm procedure (Hochberg, 1988) but, on the other hand, it is uniformly less powerful than the Hommel procedure (Hommel, 1989).

Scenarios 1 and 2 in the dose-finding trial example (Section 2.5.3) will be used to illustrate the Hocberg procedure. Consider Scenario 1 and note that the Hochberg procedure is similar to the Hommel procedure in that it rejects all hypotheses if the largest unadjusted p-value is less than or equal to α. Since the largest p-value is significant at the 0.025 level, all doses are significantly superior to Placebo after the Hochberg multiplicity adjustment. Further, it is easy to see that the Hochberg procedure finds only one significant dose-placebo comparison in Scenario 2. Note that

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

Therefore,

$$p_{(2)} > 0.025/3, \quad p_{(3)} > 0.025/2, \quad p_{(4)} > 0.025$$

and $H_{(2)}$, $H_{(3)}$ and $H_{(4)}$ are retained. However, $p_{(1)} \leq 0.025/4$ and thus the Hochberg procedure rejects $H_{(1)}$.

It is worth mentioning that the Hochberg procedure serves as a good example of the importance of a simple transparent structure in the choice of a multiple testing procedure. This procedure is very popular in clinical trial applications despite the fact that it is not α-exhaustive and thus it can be improved in a uniform manner. In fact, the Hommel procedure is based on the full closure and consequently uniformly more powerful than the Hochberg procedure; however, the Hommel procedure is based on a more complicated algorithm. In addition, one can uniformly improve the power of the Hochberg procedure in the class of step-up procedures. For example, Rom (1990) derived a step-up procedure that is uniformly more powerful than the Hochberg procedure. The Rom procedure requires tabulation of critical values for ordered p-values whereas Hochberg's critical values do not require tabulation. Thanks to its computational simplicity, the Hochberg procedure remains the most popular Simes-based procedure used in practice.

2.6.10 Adjusted p-values

This section discusses the computation of multiplicity-adjusted p-values for multiple testing procedures introduced in Sections 2.6.1–2.6.9 in the case of equally weighted hypotheses. Adjusted p-values are denoted by $\widetilde{p}_1, \ldots, \widetilde{p}_m$. The hypothesis H_i is rejected if $\widetilde{p}_i \leq \alpha$.

Bonferroni procedure

The adjusted p-value for the hypothesis H_i is $\widetilde{p}_i = \min(1, mp_i)$, $i = 1, \ldots, m$.

Holm procedure

The adjusted p-values for the hypotheses $H_{(1)}, \ldots, H_{(m)}$ are defined sequentially in the following way:

$$\widetilde{p}_{(i)} = \begin{cases} \min(1, mp_{(i)}) & \text{if } i = 1, \\ \max(\widetilde{p}_{(i-1)}, (m-i+1)p_{(i)}) & \text{if } i = 2, \ldots, m. \end{cases}$$

Fixed-sequence procedure

The adjusted p-value for H_i is given by $\widetilde{p}_i = \max(p_1, \ldots, p_i)$, $i = 1, \ldots, m$.

Fallback and Hommel procedures

The adjusted p-values for the two procedures can be computed using the general method for closed testing procedures given in Section 2.4.1.

Hochberg procedure

The adjusted p-values are defined recursively beginning with the largest p-value:

$$\widetilde{p}_{(i)} = \begin{cases} p_{(i)} & \text{if } i = m, \\ \min(\widetilde{p}_{(i+1)}, (m-i+1)p_{(i+1)}) & \text{if } i = m-1, \ldots, 1. \end{cases}$$

Dose-finding trial example

Table 2.2 displays adjusted p-values for the Bonferroni, Holm, fixed-sequence, fallback (assuming equally weighted hypotheses), Hommel and Hochberg procedures in the dose-finding trial example introduced in Section 2.5.3 (the unadjusted p-values are shown in Table 2.1).

The Bonferroni procedure rejects one hypothesis in Scenario 1 (Dose D4 is superior to Placebo) and also one hypothesis in Scenario 3 (Dose D3 is superior to placebo).

Since the Holm procedure is uniformly more powerful than the Bonferroni procedure, the Holm-adjusted p-values are no greater than the Bonferroni-adjusted p-values. This results in an additional significant test in Scenario 1 (D3-Placebo comparison) compared to the Bonferroni procedure. However, the numbers of hypotheses rejected by the Holm procedure in Scenarios 2 and 3 are the same as for the Bonferroni procedure.

The fixed-sequence procedure finds the following significant results: all doses are superior to Placebo in Scenario 1 and Doses D3 and D4 are superior to Placebo in Scenario 2. It is important to note that the fixed-sequence procedure outperforms the Bonferroni and Holm procedures in Scenarios 1 and 2

TABLE 2.2: Adjusted p-values for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted p-values that are significant at the 0.025 level.

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
	Scenario 1			
Bonferroni	0.0912	0.0608	0.0284	0.0172*
Holm	0.0304	0.0304	0.0213*	0.0172*
Fixed-sequence	0.0228*	0.0152*	0.0071*	0.0043*
Fallback	0.0228*	0.0203*	0.0172*	0.0172*
Hommel	0.0228*	0.0228*	0.0213*	0.0142*
Hochberg	0.0228*	0.0228*	0.0213*	0.0172*
	Scenario 2			
Bonferroni	0.1456	0.1188	0.0352	0.0280
Holm	0.0594	0.0594	0.0280	0.0280
Fixed-sequence	0.0364	0.0297	0.0088*	0.0070*
Fallback	0.0396	0.0396	0.0280	0.0280
Hommel	0.0364	0.0364	0.0264	0.0210*
Hochberg	0.0364	0.0364	0.0264	0.0264
	Scenario 3			
Bonferroni	0.0648	0.0420	0.0220*	0.1316
Holm	0.0324	0.0315	0.0220*	0.0329
Fixed-sequence	0.0329	0.0329	0.0329	0.0329
Fallback	0.0220*	0.0220*	0.0220*	0.1316
Hommel	0.0324	0.0243*	0.0210*	0.0329
Hochberg	0.0324	0.0315	0.0220*	0.0329

when the true dose-response relationship is monotone. This illustrates an important property of the fixed-sequence procedure. The procedure maximizes the power when the (unknown) ordering of the hypotheses in terms of the true treatment effect is close to the selected ordering. However, if the monotonicity assumption is not met, the fixed-sequence procedure tends to perform poorly. For a discussion of the robustness of the fixed-sequence procedure with respect to the monotonicity assumption, see Section 4.3.

The fallback procedure rejects more hypotheses than the Holm procedure in Scenarios 1 and 3. Further, the fallback procedure is more robust to departures from the monotonicity assumption than the fixed-sequence procedure and leads to more rejections when the first dose in the pre-determined sequence does not separate from placebo. For instance, the fallback procedure detects three significant dose-placebo comparisons in Scenario 3 whereas the fixed-sequence procedure does not find any significant results.

The Hommel procedure rejects more hypotheses than the Holm procedure in Scenarios 1, 2 and 3 (note that the Hommel procedure always rejects all hypotheses if the largest unadjusted p-value is $\leq \alpha$). However, it performs only as well as the Holm procedure in Scenario 3 when the dose-response curve

has an umbrella shape and finds fewer significant results than the fallback procedure.

The Hochberg procedure finds all dose-placebo comparisons significant in Scenario 1 and thus rejects more hypotheses of no treatment effect than the Holm procedure. However, unlike the Hommel procedure, the Hochberg procedure fails to detect the drug effect at Dose D4 in Scenario 2. Scenario 3 shows that the Hochberg procedure can sometimes be less powerful compared to the fallback procedure.

2.6.11 Simultaneous confidence intervals

In this section we will define simultaneous confidence intervals for Bonferroni-based procedures defined in Section 2.6.5; i.e., for the Bonferroni, Holm, fixed-sequence and fallback procedures, in problems with equally weighted hypotheses. Simultaneous confidence intervals for the Hommel and Hochberg procedures have not been explicitly defined in the multiple comparison literature.

To define simultaneous confidence intervals, consider a one-sided parametric multiple testing problem defined as follows. The hypothesis of no treatment effect

$$H_i: \ \theta_i \leq 0$$

is tested versus a one-sided alternative

$$K_i: \ \theta_i > 0,$$

where $i = 1, \ldots, m$ and $\theta_1, \ldots, \theta_m$ are parameters of interest, for example, mean treatment differences or differences in proportions. Let $\widehat{\theta}_i$ denote an estimate of θ_i and assume that $\widehat{\theta}_i$ is normally distributed with mean θ_i and standard deviation σ_i. The estimated standard error of θ_i is denoted by s_i, $i = 1, \ldots, m$. Further, z_x denotes the $(1 - x)$-quantile of the standard normal distribution.

Bonferroni procedure

A one-sided $100(1 - \alpha)\%$ confidence interval for $\theta_1, \ldots, \theta_m$ is given by $(\widetilde{L}_i, \infty)$, $i = 1, \ldots, m$, where

$$\widetilde{L}_i = \widehat{\theta}_i - z_{\alpha/m} s_i.$$

Holm procedure

One-sided simultaneous confidence intervals for the Holm procedure were developed by Strassburger and Bretz (2008) and Guilbaud (2008). Based on the general results presented in Section 2.6.5, the lower limits of one-sided

$100(1-\alpha)\%$ confidence intervals are given by

$$\widetilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \widehat{\theta}_i - z_{\alpha/(m-r)} s_i & \text{if } i \notin R, \\ \max(0, \widehat{\theta}_i - z_{\alpha/m} s_i) & \text{if } R = I, \end{cases}$$

where R is the index set of rejected hypotheses and r is the number of rejected hypotheses. Here the first case applies to the hypotheses rejected by the Holm procedure ($i \in R$) when the procedure retains some of the hypotheses ($R \neq I$). The second case applies to the hypotheses retained by the Holm procedure ($i \notin R$) and the third case corresponds to scenarios when all hypotheses are rejected by the procedure ($R = I$). It is worth noting that the lower limit for a parameter is set to 0 whenever the Holm procedure rejects the corresponding hypothesis of no treatment effect. The only exception is when the procedure rejects all hypotheses. In this case the lower limits associated with rejected hypotheses can be greater than 0.

Fixed-sequence procedure

Hsu and Berger (1999) constructed simultaneous confidence intervals associated with the fixed-sequence procedure in one-sided parametric problems. Based on the general results from Section 2.6.5, we obtain the lower limits of the one-sided $100(1-\alpha)\%$ simultaneous confidence intervals for θ_i, $i = 1, \ldots, m$,

$$\widetilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \widehat{\theta}_i - z_\alpha s_i & \text{if } i = i^* \text{ and } R \neq I, \\ \min_{i \in I}\{\widehat{\theta}_i - z_\alpha s_i\} & \text{if } R = I, \end{cases}$$

where i^* denotes the first hypothesis in the sequence not rejected by the procedure when the procedure retains some of the hypotheses ($R \neq I$) and R is the index set of rejected hypotheses. Note that the hypotheses H_j with $j > i^*$ are not tested and therefore no confidence intervals are available for the associated parameters. Further, it follows from this definition that the lower limits of simultaneous confidence intervals for the fixed-sequence procedure are similar to the Holm-adjusted limits in the sense that they are set to 0 if the corresponding hypothesis is rejected unless the fixed-sequence procedure rejects all hypotheses.

Fallback procedure

An extension of the method for setting up simultaneous confidence intervals for Bonferroni-based closed testing procedures proposed by Strassburger and Bretz (2008) can be used to define simultaneous confidence intervals for

the fallback procedure. The lower limits of one-sided $100(1-\alpha)\%$ confidence intervals are derived as follows. First, for any non-empty index set $J \subseteq I$, let

$$\alpha_i(J) = \begin{cases} 0 & \text{if } i \notin J, \\ \alpha(i - \ell_i(J))/m & \text{if } i \in J, \end{cases}$$

where $\ell_i(J)$ is the largest index in J that is smaller than i if i is not the smallest index in J and $\ell_i(J) = 0$ if i is the smallest index in J. Similarly, for any non-empty index set $J \subseteq I$ and $i \notin J$, let

$$\alpha_i^*(J) = \frac{1}{m - |J|}\left(\alpha - \sum_{j \in J} \alpha_j(J)\right),$$

where $|J|$ is the number of elements in J.

The lower limits are given by

$$\widetilde{L}_i = \begin{cases} \min_{J \subseteq A} \max(0, \widehat{\theta}_i - z_{\alpha_i^*(J)} s_i) & \text{if } i \in R \text{ and } R \neq I, \\ \widehat{\theta}_i - z_{\alpha_i(A)} s_i & \text{if } i \in A, \\ \max(0, \widehat{\theta}_i - z_{\alpha/m} s_i) & \text{if } R = I, \end{cases}$$

where A and R are the index sets of retained and rejected hypotheses, respectively. These lower limits take advantage of the fact that the fallback procedure is not α-exhaustive and are uniformly sharper than those based on the general method presented in Section 2.6.5. Unlike the lower limits for the Holm and fixed-sequence procedures, the fallback-adjusted lower limits are not automatically set to 0 for parameters corresponding to rejected hypotheses of no treatment effect.

Dose-finding trial example

Table 2.3 shows the lower limits of one-sided 97.5% simultaneous confidence intervals for the four mean treatment differences under Scenario 1 in the dose-finding trial example (the unadjusted lower limits are presented in Table 2.1). The limits are computed for the Bonferroni, Holm, fixed-sequence and fallback procedures (as before, the weights are assumed to be equal in the fallback procedure).

Table 2.3 illustrates key properties of simultaneous confidence intervals. First of all, comparing the lower limits displayed in Table 2.3 to the adjusted p-values presented in Table 2.2 under Scenario 1, it is easy to see that the lower limit is less than 0 if the procedure fails to reject the corresponding hypothesis of no treatment effect at the 0.025 level.

The Holm-adjusted lower limits are sharper than those for the Bonferroni procedure when the latter fails to reject a hypothesis of no treatment effect

TABLE 2.3: Lower limits of one-sided 97.5% simultaneous confidence intervals for the mean dose-placebo treatment differences in the dose-finding trial example (Scenario 1)

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Bonferroni	−0.71	−0.47	−0.05	0.20
Holm	−0.34	−0.10	0.00	0.00
Fixed-sequence	0.07	0.07	0.07	0.07
Fallback	0.00	0.00	0.00	0.20

(see the lower limits for the D1-Placebo, D2-Placebo and D3-Placebo tests). However, when both procedures reject a hypothesis, the Holm-adjusted lower limit is less informative than the Bonferroni-adjusted lower limit. Consider, for example, the D4-Placebo test. In this case the Bonferroni-adjusted lower limit is positive and thus provides information about the likely magnitude of the treatment difference whereas the Holm-adjusted lower limit is simply equal to 0.

Further, the lower limits for the fixed-sequence procedure are positive and constant across the four dose-placebo comparisons because the procedure rejects all hypotheses in Scenario 1. The fallback procedure also rejects all hypotheses and thus the associated lower limits are nonnegative. However, the first three lower limits are set to 0 and only one lower limit is positive (D4-Placebo test).

2.7 Parametric multiple testing procedures

In certain situations, for example, in dose-finding clinical trials with normally distributed outcomes, it is possible to improve the power of p-value-based procedures by taking advantage of parametric assumptions about the joint distribution of the test statistics. Multiple testing procedures that rely on these assumptions are known as parametric procedures. The most well-known parametric procedure is the Dunnett procedure (Dunnett, 1955) developed for problems with multiple dose-control comparisons. This single-step procedure is described in this section along with other parametric procedures such as the stepwise Dunnett procedures and parametric Shaffer procedure.

The following setting will be used throughout this section. Consider a dose-finding clinical trial designed to compare m doses or regimens of a treatment to a placebo. For simplicity, a balanced one-way layout will be assumed; i.e.,

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where y_{ij} is the response of the jth patient in the ith treatment group, $i =$

$0, \ldots, m$ ($i = 0$ denotes the placebo group) and $j = 1, \ldots, n$. The errors, ε_{ij}, $i = 0, \ldots, m$, $j = 1, \ldots, n$, are normally distributed with mean 0 and common standard deviation σ.

The testing problem is formulated in terms of the m treatment-placebo comparisons; i.e., the hypotheses $H_i: \theta_i = 0$, $i = 1, \ldots, m$, are tested against the one-sided alternatives $K_i: \theta_i > 0$, $i = 1, \ldots, m$, where $\theta_i = \mu_i - \mu_0$. Let t_i be the t statistic for testing H_i; i.e.,

$$t_i = \frac{\bar{y}_i - \bar{y}_0}{s\sqrt{2/n}},$$

where s is the pooled sample standard deviation.

2.7.1 Single-step Dunnett procedure

The single-step Dunnett procedure can be thought of as a set of two-sample tests for $H_1, \ldots, H_m$ adjusted for multiplicity. However, unlike p-value-based procedures described in Section 2.6, the Dunnett procedure is based on the joint distribution of the test statistics and thus accounts for the correlation among the test statistics.

It is easy to show that, under H_i, t_i follows the standard univariate t distribution with $\nu = 2(n-1)$ df and thus the regular (unadjusted) critical value for t_i is $t_\alpha(\nu)$; i.e., the $(1-\alpha)$-quantile of the t distribution. Similarly, the Dunnett-adjusted critical value for $t_1, \ldots, t_m$, denoted by $u_\alpha(m, \nu)$, is the $(1-\alpha)$-quantile of the distribution of the maximum of t-distributed random variables with $\nu = (m+1)(n-1)$ df. In other words, $u_\alpha(m, \nu) = F^{-1}(1 - \alpha|m, \nu)$, where $F(x|m, \nu)$ is the cumulative distribution function of the one-sided Dunnett distribution; i.e.,

$$F(x|m, \nu) = P\{\max(t_1, \ldots, t_m) \leq x\},$$

where the probability is evaluated under the overall null hypothesis $H_0 : \theta_1 = \ldots = \theta_m = 0$. The Dunnett procedure rejects H_i if $t_i \geq u_\alpha(m, \nu)$, $i = 1, \ldots, m$.

Dunnett-adjusted critical values are smaller than Bonferroni-adjusted critical values. Therefore, the use of the Dunnett procedure leads to more powerful inferences compared to the Bonferroni procedure. As an illustration, we will use Scenario 1 in the dose-finding trial example introduced in Section 2.5.3. The Bonferroni-adjusted critical value in this problem is $t_{\alpha/m}(\nu)$ with $\alpha = 0.025$, $m = 4$ and $\nu = 2(n-1) = 152$, i.e., 2.53. As was stated in Section 2.6.1, only one test statistic is greater than this critical value (Dose D4 is superior to placebo). The Dunnett-adjusted critical value is given by $u_\alpha(m, \nu)$ with $\alpha = 0.025$, $m = 4$ and $\nu = (m+1)(n-1) = 380$. The critical value is 2.45 and the Dunnett procedure detects two significant dose-placebo comparisons in Scenario 1 (Doses D3 and D4 are superior to placebo).

It is important to note that the Dunnett procedure can be applied to any problem in which the test statistics for multiple dose-placebo comparisons

asymptotically follow a multivariate normal distribution. For example, this procedure can be used in clinical trials with categorical outcomes provided the proportions are not so different as to cause serious heteroscedasticity problems (Chuang-Stein and Tong, 1995). However, it is generally preferable to fit the model desired and use the multiplicity adjustments that follow from that model specifically. Hothorn, Bretz and Westfall (2008) gave examples that include binary and other parametric non-normally distributed cases.

2.7.2 Stepwise Dunnett procedures

The Dunnett procedure defined in Section 2.7.1 is similar to the Bonferroni procedure in that it also has a single-step structure and is not α-exhaustive (for this reason, the Dunnett procedure can be thought of as a parametric version of the Bonferroni procedure). This implies that one can develop more powerful parametric procedures by applying the closure principle. This section introduces two stepwise versions of the Dunnett procedure: a step-down procedure analogous to the Holm procedure and a step-up procedure analogous to the Hochberg procedure. Both procedures are uniformly more powerful than the single-step Dunnett procedure.

Step-down Dunnett procedure

A step-down procedure which serves as a parametric extension of the Holm procedure presented in Section 2.6.2 was developed by Naik (1975) and Marcus, Peritz and Gabriel (1976). Recall that the Holm procedure is defined using ordered p-values and, to define the step-down Dunnett procedure, we will use ordered test statistics $t_{(1)} > \ldots > t_{(m)}$ and associated hypotheses $H_{(1)}, \ldots, H_{(m)}$.

The step-down Dunnett procedure is a sequentially rejective procedure that first assesses if there is sufficient evidence to reject $H_{(1)}$ under the overall null hypothesis, i.e., all doses are no different from placebo. If $H_{(1)}$ cannot be rejected, testing stops. Otherwise, the next hypothesis in the sequence, $H_{(2)}$, is tested under the assumption that the remaining $m-1$ hypotheses are true and so on.

Defining $u_\alpha(i,\nu)$, $i = 1, \ldots, m$, as the $(1-\alpha)$-quantile of the i-variate t distribution with $\nu = (m+1)(n-1)$ df, the step-down version of the Dunnett procedure is implemented using the following algorithm:

- Step 1. If $t_{(1)} \geq c_1$, where $c_1 = u_\alpha(m,\nu)$, reject $H_{(1)}$ and go to the next step. Otherwise retain all hypotheses and stop.
- Steps $i = 2, \ldots, m-1$. If $t_{(i)} \geq c_i$, where $c_i = u_\alpha(m-i+1,\nu)$, reject $H_{(i)}$ and go to the next step. Otherwise retain $H_{(i)}, \ldots, H_{(m)}$ and stop.
- Step m. If $t_{(m)} \geq c_m$, where $c_m = u_\alpha(1,\nu)$, reject $H_{(m)}$. Otherwise retain $H_{(m)}$.

The step-down Dunnett procedure uses the critical value associated with the single-step Dunnett procedure at the first step, i.e., $c_1 = u_\alpha(m, \nu)$. Further, $c_1 > c_2 > \ldots > c_m$ and thus the other hypotheses are tested using successively sharper critical values. This implies that the step-down procedure rejects as many (and potentially more) hypotheses than the single-step Dunnett procedure. In addition, the step-down Dunnett procedure is uniformly more powerful than the Holm procedure.

The step-down Dunnett procedure defined above assumes a balanced one-way layout. The step-down procedure in the general unbalanced case was considered by Bofinger (1987) and Dunnett and Tamhane (1991).

Scenario 1 in the dose-finding trial example given in Section 2.5.3 will be used to illustrate the step-down testing algorithm. The ordered t statistics in this scenario are given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1$$

and the critical values at Steps 1 through 4 are equal to 2.45, 2.36, 2.22 and 1.97, respectively. The first ordered test statistic, $t_{(1)} = 2.64$, is greater than the corresponding critical value, 2.45, and thus the hypothesis $H_{(1)}$ is rejected (Dose D4 is superior to placebo). The next ordered statistic, $t_{(2)} = 2.46$, is compared to 2.36 and is again significant (Dose D3 is superior to placebo). However, the other two hypotheses of no treatment effect are retained since $t_{(3)} = 2.17$ is less than 2.22.

Step-up Dunnett procedure

A step-up version of the Dunnett procedure was proposed by Dunnett and Tamhane (1992). It is conceptually similar to the step-up Hochberg and Rom procedures described in Section 2.6.9.

The step-up testing algorithm is set up as follows. The ordered t statistics $t_{(1)} > \ldots > t_{(m)}$ are compared to suitably defined critical values $c_1, \ldots, c_m$ in a stepwise fashion starting with the least significant test statistic, i.e., $t_{(m)}$. At each step, all remaining hypotheses are rejected if the test statistic is greater or equal to the corresponding critical value. Specifically, testing is performed as follows:

- Step 1. If $t_{(m)} < c_1$, retain $H_{(m)}$ and go to the next step. Otherwise reject all hypotheses and stop.
- Steps $i = 2, \ldots, m-1$. If $t_{(m-i+1)} < c_i$, retain $H_{(m-i+1)}$ and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step m. If $t_{(1)} < c_m$, retain $H_{(1)}$ and reject it otherwise.

Dunnett and Tamhane (1992) showed that the step-up Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure as well as

the Hochberg procedure which serves as an example of a nonparametric step-up procedure. However, the step-up Dunnett procedure does not uniformly dominate the step-down Dunnett procedure in terms of power. The step-up procedure tends to be more powerful than the step-down Dunnett procedure when most of the true mean treatment-control differences in a dose-finding study are positive.

The critical values in the step-up procedure are defined in such a way that the FWER is controlled at the α level. The following recursive algorithm can be used to obtain the critical values in a balanced one-way layout. Let $T_1, \ldots, T_m$ be random variables with the same joint distribution as $t_1, \ldots, t_m$ under the global null hypothesis. The critical value c_1 is found from

$$P(T_1 \geq c_1) = \alpha.$$

Further, given $c_1, \ldots, c_{i-1}$, the critical value c_i is chosen so that

$$P(T_{(1)} \geq c_1 \text{ or } T_{(2)} \geq c_2 \text{ or } \ldots \text{ or } T_{(i)} \geq c_i) = \alpha,$$

where $T_{(1)} < \ldots < T_{(i)}$. Note that c_1 is simply the $(1-\alpha)$-quantile of the univariate t distribution with $\nu = (m+1)(n-1)$ df and thus the step-up Dunnett procedure is similar to the Hochberg procedure in that it also rejects all hypotheses if the least significant p-value is no greater than α.

Calculation of critical values for the step-up procedure in the general unbalanced case was considered by Dunnett and Tamhane (1995) and Grechanovsky and Pinsker (1999). An efficient algorithm for computing the critical values was proposed by Kwong and Liu (2000).

The step-up algorithm will be illustrated using Scenarios 1 and 2 in the dose-finding trial example from Section 2.5.3. To carry out the step-up Dunnett procedure, we will use the critical values given in Table 2 of Dunnett and Tamhane (1992) with $\nu = \infty$. The critical values are given by $c_1 = 1.96$, $c_2 = 2.22$, $c_3 = 2.36$ and $c_4 = 2.45$.

Considering Scenario 1, the ordered test statistics are given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1.$$

At the first step of the algorithm, the least significant test statistic, $t_{(4)} = 2.01$, is compared to c_1. Since the test statistic is greater than the critical value, the step-up Dunnett procedure rejects all hypotheses of no treatment effect and thus all doses are declared superior to Placebo.

The ordered test statistics in Scenario 2 are again given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1.$$

Since the least significant test statistic, $t_{(4)} = 1.80$, is no greater than c_1, the step-up Dunnett procedure retains the hypothesis $H_{(4)}$. At the next step, the test statistic, $t_{(3)} = 1.89$, is less than the corresponding critical value c_2 and thus $H_{(3)}$ is also retained. Lastly, $t_{(2)} = 2.38$ exceeds the critical value c_3 and, as a consequence, the step-up Dunnett procedure rejects the two remaining hypotheses (Doses D3 and D4 are superior to Placebo).

2.7.3 Extended Shaffer-Royen procedure

Westfall and Tobias (2007) discussed the extended Shaffer-Royen procedure that serves as a parametric extension of Shaffer's Method 2 described in Section 2.6.2 to account for logical dependencies among hypotheses. When the hypotheses are formulated in terms of dose-placebo contrasts that are not logically related, the procedure reduces precisely to the step-down Dunnett method described above. More generally, the extended Shaffer-Royen procedure is a truncated closed testing procedure similar to the Shaffer's Method 2 procedure. Note that the latter procedure uses the Bonferroni test for each intersection hypothesis H_I while the parametric procedure uses the distribution of the maximum test statistic for the intersection hypothesis $H_I = \cap_{i\in I} H_i$, i.e., $\max_{i\in I} t_i$, to make the method more powerful. "Royen" appears in the name of the procedure since Royen (1989) first applied it to the problem of testing all pairwise comparisons. The extended Shaffer-Royen procedure can be used to test arbitrary contrasts when the contrasts are logically related, as occurs, for example, when the multiple contrasts represent subgroups (see the example in Hochberg and Westfall, 2000).

2.7.4 Adjusted p-values and simultaneous confidence intervals

In this section we will introduce algorithms for computing adjusted p-values and associated simultaneous confidence intervals for the Dunnett-based parametric procedures.

Single-step Dunnett procedure

The adjusted p-values for individual hypotheses are found using the multivariate t distribution. Specifically, the adjusted p-value for H_i is $\widetilde{p}_i = 1 - F(t_i|m,\nu)$, where $F(x|m,\nu)$ is the cumulative distribution function of the one-sided Dunnett distribution with $\nu = (m+1)(n-1)$ defined in Section 2.7.1. In other words, the adjusted p-value for H_i is found from

$$t_i = u_{\widetilde{p}_i}(m,\nu).$$

The lower limits of one-sided $100(1-\alpha)\%$ simultaneous confidence intervals for the mean treatment differences $\theta_i = \mu_i - \mu_0$, $i = 1,\ldots,m$, are defined as follows:

$$\widetilde{L}_i = \widehat{\theta}_i - u_\alpha(m,\nu)s_i,$$

where s_i is the standard error of $\widehat{\theta}_i$, i.e., $s_i = s\sqrt{2/n}$.

Step-down Dunnett procedure

The adjusted p-values for the step-down Dunnett procedure are found using the following algorithm (Dunnett and Tamhane, 1992). First, define

$\gamma_1, \ldots, \gamma_m$ as follows:

$$t_{(i)} = u_{\gamma_i}(m - i + 1, \nu), \quad i = 1, \ldots, m,$$

where $\nu = (m + 1)(n - 1)$. The adjusted p-values are given by

$$\widetilde{p}_{(i)} = \begin{cases} \gamma_i & \text{if } i = 1, \\ \max(\widetilde{p}_{i-1}, \gamma_i) & \text{if } i = 2, \ldots, m. \end{cases}$$

Simultaneous confidence intervals for the step-down procedure were derived by Bofinger (1987) and Stefansson, Kim and Hsu (1988). The lower limits of one-sided $100(1-\alpha)\%$ simultaneous confidence intervals for $\theta_i = \mu_i - \mu_0$, $i = 1, \ldots, m$, are derived using the Stefansson-Kim-Hsu method defined below. The underlying algorithm is similar to the algorithm used in the calculation of lower simultaneous confidence limits for the step-down version of the Bonferroni procedure, e.g., Holm procedure (see Section 2.6.11):

$$\widetilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \widehat{\theta}_i - c_{r+1} s_i & \text{if } i \notin R, \\ \max(0, \widehat{\theta}_i - c_m s_i) & \text{if } R = I, \end{cases}$$

where R is the index set of rejected hypotheses, $I = \{1, \ldots, m\}$ and r is the number of rejected hypotheses. As in Section 2.6.11, the first case defines the lower limits for the hypotheses rejected by the step-down procedure ($i \in R$) when some other hypotheses are retained ($R \neq I$). Note that the lower limits for the treatment differences asserted to be significant by the step-down procedure ($i \in R$) are automatically set to 0 (unless all hypotheses are rejected, i.e., $R = I$). When confidence limits for the step-down procedure are equal to 0, they may be less informative than positive limits for the treatment differences that are found significant by the single-step Dunnett procedure. Further, the second case applies to the hypotheses retained by the step-down procedure ($i \notin R$) and the third case defines the lower limits when all hypotheses are rejected ($R = I$).

Step-up Dunnett procedure

Adjusted p-values for the step-up Dunnett procedure are defined in Dunnett and Tamhane (1992) and Grechanovsky and Pinsker (1999). The algorithm is computationally intensive and calculation of adjusted p-values for this procedure will not be discussed in this book. Further, a method for constructing simultaneous confidence intervals for the step-up Dunnett procedure has not been developed yet.

TABLE 2.4: Adjusted p-values for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted p-values that are significant at the 0.025 level.

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
	Scenario 1			
Single-step Dunnett	0.0715	0.0493	0.0242*	0.0152*
Step-down Dunnett	0.0280	0.0280	0.0190*	0.0152*
	Scenario 2			
Single-step Dunnett	0.1090	0.0909	0.0297	0.0238*
Step-down Dunnett	0.0535	0.0535	0.0238*	0.0238*
	Scenario 3			
Single-step Dunnett	0.0523	0.0351	0.0191*	0.0994
Step-down Dunnett	0.0298	0.0278	0.0191*	0.0329

TABLE 2.5: Lower limits of one-sided 97.5% simultaneous confidence intervals for the mean dose-placebo treatment differences in the dose-finding trial example (Scenario 1).

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Single-step Dunnett	−0.64	−0.40	0.02	0.27
Step-down Dunnett	−0.31	−0.07	0	0

Dose-finding trial example

Table 2.4 lists adjusted p-values produced by the single-step and step-down Dunnett procedures in the dose-finding trial example (see Section 2.5.3). It is easy to verify that the adjusted p-values for the single-step Dunnett procedure are uniformly smaller than those associated with the Bonferroni procedure (Bonferroni-adjusted p-values are displayed in Table 2.2). The single-step Dunnett procedure finds two significant dose-placebo comparisons in Scenario 1, one significant comparison in Scenario 2 and one significant comparison in Scenario 3. The step-down Dunnett procedure is uniformly superior to the single-step Dunnett procedure as well as the nonparametric step-down procedure, i.e., Holm procedure.

Further, the lower limits of one-sided 97.5% simultaneous confidence intervals for the mean treatment differences in the dose-finding trial example are displayed in Table 2.5 (assuming Scenario 1). Note that the lower confidence limits for the single-step Dunnett procedure are uniformly sharper than those for the Bonferroni procedure (Table 2.3). Further, the lower confidence limits associated with the step-down Dunnett procedure are greater than the lower confidence limits for the single-step Dunnett procedure when the latter does not reject a hypothesis and, when both procedures reject a hypothesis, the opposite is true.

2.7.5 Multiple comparisons in general linear models

In this section we extend the single-step Dunnett procedure from Section 2.7.1 and describe a general approach to the problem of constructing multiple testing procedures in general linear models that account for the stochastic dependencies among the test statistics. The general theory is covered, among others, by Hochberg and Tamhane (1987), Hsu (1996, Chapter 7) and Bretz, Hothorn and Westfall (2008).

Consider the common linear model

$$Y = X\beta + \varepsilon,$$

where Y is an $n \times 1$ response vector, X is a fixed and known $n \times p$ design matrix, β is an unknown $p \times 1$ parameter vector and ε is an $n \times 1$ vector of independent normally distributed errors with mean 0 and unknown variance σ^2. The least square unbiased estimates of β and σ are given by

$$\widehat{\beta} = (X'X)^- X'Y \text{ and } s^2 = \frac{(Y - X\widehat{\beta})'(Y - X\widehat{\beta})}{\nu},$$

respectively, where $\nu = n - \text{rank}(X)$ is the error degrees of freedom and $(X'X)^-$ is some generalized inverse of $X'X$.

Let C denote a constant $p \times m$ matrix which describes the experimental questions of interest. Each column c_i, $i = 1, \ldots, m$, of C defines a single experimental comparison of interest. Without loss of generality, we consider the associated one-sided testing problem

$$H_i : \ \theta_i \leq 0, \quad i = 1, \ldots, m,$$

where $\theta_i = c_i'\beta$.

The hypotheses are tested using the test statistics

$$t_i = \widehat{\theta}_i / s_i, \quad i = 1, \ldots, m,$$

where $s_i = s\sqrt{c_i'(X'X)^- c_i}$, $i = 1, \ldots, m$.

It can be shown that the joint distribution of $t_1, \ldots, t_m$ is multivariate t with ν degrees of freedom and correlation matrix $DC'(X'X)^- CD$, where $D = \text{diag}(c_i'(X'X)^- c_i)^{-1/2}$. In the asymptotic case $\nu \to \infty$ or if σ is known, the corresponding multivariate normal distribution can be used instead. Let u_α denote the critical value derived from the multivariate normal or t distribution. Then, H_i is rejected if $t_i \geq u_\alpha$. Equivalently, adjusted p-values $\widetilde{p}_i$ can be calculated from the multivariate normal or t distribution and we reject H_i if $\widetilde{p}_i \leq \alpha$. Numerical integration methods to calculate the multivariate normal and t probabilities required for the computation of critical values and adjusted p-values are described by Genz and Bretz (2002, 2009). Finally, one-sided simultaneous confidence intervals for $\theta_1, \ldots, \theta_m$ with simultaneous coverage probability $1 - \alpha$ are given by

$$(-\infty, \widehat{\theta}_i + u_\alpha s_i], \quad i = 1, \ldots, m,$$

To illustrate this framework, we revisit the single-step Dunnett test considered in Section 2.7.1 for comparing m treatments with a control. Here, $p = m + 1$ and $\beta = (\mu_0, \ldots, \mu_m)'$ is the parameter vector. The index $i = 0$ denotes the placebo control to which the remaining m treatment arms are compared. The associated C matrix is

$$C_{m+1\times m} = \begin{pmatrix} -1 & -1 & \ldots & -1 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Thus, for the ith vector

$$c_i' = (-1, 0, \ldots, 1, 0, \ldots, 0),$$

we obtain the pairwise mean treatment differences $\theta_i = \mu_1 - \mu_0$, $i = 1, \ldots, m$. It is easy to show that the resulting m treatment-control comparisons are exactly those considered in Section 2.7.1.

This general framework includes many standard parametric multiple testing procedures beyond the Dunnett procedure, such as the Tukey procedure for all pairwise comparisons, the trend tests of Williams (1971) and Marcus (1976) (see Chapter 3) and other procedures listed in Bretz, Genz and Hothorn (2001). The framework enables one to include covariates and/or factorial treatment structures in classical regression and ANOVA applications. An extension to more general parametric and semi-parametric models relying on standard asymptotic multivariate normal assumptions was provided by Hothorn, Bretz and Westfall (2008), which allows a unified treatise of multiple comparisons for generalized linear models, mixed models, survival models, etc. Note that in this section we focused on single-step procedures. Using the results from Section 2.7.2, more powerful closed testing procedures can be constructed based on the parametric framework described here. These procedures account for the stochastic dependencies among the test statistics and exploit inherent logical constraints for general parametric models. These methods are all implemented in the *multcomp* package reviewed in Section 2.9 and described in Bretz, Hothorn and Westfall (2010).

2.8 Resampling-based multiple testing procedures

Resampling is a general term that encompasses the bootstrap, permutation analysis, and parametric simulation-based analyses. The general method extends parametric methods described in Section 2.7.

The resampling-based methods

- make fewer assumptions about the data-generating process (for example, normality does not have to be assumed) and yield more robust multiple testing procedures,
- utilize data-based distributional characteristics, including discreteness and correlation structure which yield more powerful procedures.

At the same time, one needs to be aware of some drawbacks of the resampling-based approach. In particular, resampling-based methods

- are often approximate, requiring large sample sizes and/or simulations to justify their use (except for permutation-based methods, which are exact even for small sample sizes; see Section 2.8.4),
- can be computationally difficult,
- are based on careful modeling of the data-generating process requiring very complex models in some cases such as survival analysis.

This section begins with the description of general principles used in the construction of resampling-based multiple testing procedures (Section 2.8.1). Sections 2.8.2–2.8.3 give examples of parametric and non-parametric resampling-based procedures and Section 2.8.4 discusses exact permutation-based procedures.

2.8.1 Closed resampling-based procedures

The closure principle introduced in Section 2.3.3 provides a convenient, flexible and powerful foundation to describe resampling-based multiple testing procedures. Consider m hypotheses of interest denoted by $H_1, \ldots, H_m$. As in Section 2.7, let $T_1, \ldots, T_m$ denote the random variables representing the test statistics associated with these hypotheses and $t_1, \ldots, t_m$ denote the observed test statistics (realizations of $T_1, \ldots, T_m$).

In general, any α-level test may be used to test intersection hypotheses in the closed family induced by $H_1, \ldots, H_m$. Previous sections have discussed the use of procedures based on the minimum p-value (e.g., Bonferroni procedure) or maximum test statistic (e.g., Dunnett procedure). However, there are many other choices. In an ANOVA setting arising in dose-finding studies, F-statistics may be used to test intersection hypotheses. Similarly, in multicenter, subgroup, or other types of analyses where it is desired to combine data across locations, Fisher combination tests or other meta-analytic tests may be used. The choice of test statistic to use should primarily be based on power considerations. Once a powerful test statistic is chosen, resampling can be used to ensure that the test is robust to violations of distributional and/or dependence assumptions.

While power is the main concern for choice of a test statistic, expediency becomes important when the number of hypotheses m is large. There are

$2^m - 1$ intersection hypotheses in the closed family and, if m is large, it is computationally impossible to test every single intersection. However, the computational burden can be eased dramatically if the following assumptions are made:

- For each non-empty index set $I \subseteq \{1, \ldots, m\}$, the intersection hypothesis $H_I = \cap_{i \in I} H_i$ is tested using the maximum statistic $t_{\max}(I) = \max_{i \in I} t_i$.
- The *subset pivotality condition* (Westfall and Young, 1993) is met, i.e., for each non-empty index set I, the distribution of $T_{\max}(I)$ under H_I is identical to the distribution of $T_{\max}(I)$ under the global null hypothesis H_N.

The benefit of the two assumptions is that stepwise resampling-based procedures similar to the Holm and stepwise Dunnett procedures can be constructed. In other words, we need to test only m hypotheses corresponding to the ordered test statistics $t_{(1)} > \ldots > t_{(m)}$ rather than all $2^m - 1$ intersection hypotheses. Further, resampling can be done simultaneously under the global null hypothesis rather than separately for each intersection hypothesis.

To illustrate, let $H_{(1)}, \ldots, H_{(m)}$ denote the hypotheses associated with the ordered test statistics. The step-down resampling-based procedure is defined as follows:

- Step 1. Reject $H_{(1)}$ if

$$P(\max(T_1, \ldots, T_m) \geq t_{(1)}) \leq \alpha$$

  and go to the next step. Otherwise retain $H_{(1)}, \ldots, H_{(m)}$ and stop.

- Steps $i = 2, \ldots, m-1$. Reject $H_{(i)}$ if

$$P(\max(T_i, \ldots, T_m) \geq t_{(i)}) \leq \alpha$$

  and go to the next step. Otherwise retain $H_{(i)}, \ldots, H_{(m)}$ and stop.

- Step m. Reject $H_{(m)}$ if

$$P(T_m \geq t_{(m)}) \leq \alpha.$$

  Otherwise retain $H_{(m)}$ and stop.

The probabilities in this algorithm are computed under the global null hypothesis. The step-down procedure is identical to the full-blown closed testing procedure based on the examination of each individual intersection hypothesis in the closed family. Adjusted p-values are the probabilities shown at each step, adjusted for monotonicity as in the case of the parametric step-down Dunnett procedure.

It is worth noting that, in general, we do not need to use resampling to implement this step-down procedure. This will be the case when the probabilities at Steps 1 through m can be computed directly. However, when direct computations are not feasible, resampling methods are used to obtain these probabilities.

In addition to the use of the maximum test statistics and subset pivotality condition, we also need to assume that there are no logical constraints among the hypotheses to be able to construct a step-down procedure identical to the original closed testing procedure. If there are logical constraints, power can be improved by restricting attention only to intersection hypotheses consistent with the constraints. However, in this case the computational shortcuts disappear and we are back in the position of having to evaluate the tests for all intersection hypotheses. The step-down procedure defined above can still be used, though, as it provides a conservative approximation to the closed testing procedure.

Sections 2.8.2–2.8.4 illustrate the general step-down method in some special cases.

2.8.2 Step-down Dunnett procedures based on parametric and nonparametric resampling

The step-down Dunnett procedure for parametric problems arising in dose-finding trials was presented in Section 2.7.2. We will show in this section that the step-down procedure defined in Section 2.8.1 is obtained (at least in the simulation limit) via parametric normal resampling. In addition, we will show how to extend this method simply using bootstrap resampling for cases where the normality assumption is violated.

Consider the setting introduced in Section 2.7. In particular, assume that the responses follow the ANOVA model defined in that section and assume that the errors are independent, identically distributed random variables with mean 0 and variance σ^2. Further, consider the same set of hypotheses, i.e., $H_i : \mu_i = \mu_0$, $i = 1, \ldots, m$, and define the closed family associated with $H_1, \ldots, H_m$. The intersection hypothesis H_I is tested using the test statistic $t_{\max}(I) = \max_{i \in I} t_i$, where t_i is the t-statistic for comparing the ith group to placebo. The p-value for this intersection hypothesis is given by

$$p_I = P(T_{\max}(I) \geq t_{\max}(I)),$$

where $t_{\max}(I)$ is defined based on the observed test statistics $t_1, \ldots, t_m$ and the probability is computed under the global null hypothesis.

If the errors in the ANOVA model are normally distributed, we can use the multivariate t distribution to calculate the p-value for each intersection hypothesis. In this case, the step-down procedure introduced in Section 2.8.1 will simplify to the step-down Dunnett procedure. Alternatively, as shown below, the same step-down algorithm results exactly (in the simulation limit) from parametric resampling and these parametric resampling procedures suggest

natural extensions to nonparametric resampling (also known as bootstrap-based procedures).

Parametric resampling

To set up a step-down procedure based on parametric resampling, consider the ANOVA model with normally distributed errors and note that the distribution of T_i, $i \in I$, does not depend on μ_j for $j \notin I$ when H_I is true, i.e., subset pivotality holds. Therefore, we can simulate T_i by parametrically resampling the data as follows:

- Step 1. Generate a resampled data set

$$y_{ij}^* = 0 + \varepsilon_{ij}^*, \quad i = 0, \ldots, m, \quad j = 1, \ldots, n,$$

  where the ε_{ij}^* are independent, identically distributed normal variables with mean 0 and variance s^2. Actually, one can use any variance, since the distribution of T_i is also free of σ^2. However, use of s^2 clarifies the connection to the nonparametric resampling algorithm described later in this section.

- Step 2. Compute the statistics

$$T_i^* = \frac{\bar{y}_i^* - \bar{y}_0^*}{s^* \sqrt{2/n}}, \quad i = 1, \ldots, m,$$

  where s^* is the pooled sample standard deviation computed from the resampled data set.

Repeat the two steps B times (B needs to be large, e.g., $B = 100,000$). The probability for the ith step in the step-down algorithm is approximately (within binomial simulation error) the proportion of the B samples where

$$\max(T_i^*, \ldots, T_m^*) \geq t_{(i)}, \quad i = 1, \ldots, m.$$

Nonparametric resampling

The parametric resampling method generalizes very easily to models in which the distribution of random errors is unknown. However, one does not simply "resample the data and hope for the best." Instead care is needed to specify a model and resampling scheme that is appropriate for the given model.

Consider again the ANOVA model and assume that the random errors follow a general distribution with mean zero and finite variance, rather than a normal distribution. This model is called a *location-shift model*. As before, define the closed family associated wtih $H_1, \ldots, H_m$. The p-value for the intersection hypothesis H_I is again given by

$$p_I = P(T_{\max}(I) \geq t_{\max}(I)),$$

TABLE 2.6: Adjusted p-values produced by the bootstrap-based step-down Dunnett procedure for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted p-values that are significant at the 0.025 level.

Scenario	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
1	0.0279	0.0279	0.0189*	0.0151*
2	0.0534	0.0534	0.0238*	0.0238*
3	0.0299	0.0278	0.0191*	0.0330

where the probability is computed under the global null hypothesis based on the true distribution F. What is different is that the p-value now depends on the underlying distribution F, which is unknown. If we knew F, we could use the parametric resampling algorithm described above with the ε_{ij}^*'s simulated from this distribution. Since F is unknown, its estimate, denoted by $\widehat{F}$, will be used and the ε_{ij}^*'s will be generated from $\widehat{F}$ in the algorithm. The resulting p-value is only approximate, since $\widehat{F}$ is not equal to the true distribution F. However, typically (though not automatically) with larger sample sizes, $\widehat{F}$ becomes closer to F. Still, in any finite-sample multiple testing problem, the fact that the procedure is approximate means that simulation studies are generally needed to assess the adequacy of the approximation.

There are a variety of ways to obtain p-values in the nonparametric algorithm, including various flavors of simple and smoothed bootstraps. As an example, consider the procedure based on the basic bootstrap method. It is virtually identical to the parametric resampling-based procedure defined above. The only difference is that the ε_{ij}^*'s are sampled with replacement from the sample random errors $e_{ij} = y_{ij} - \bar{y}_i$, $i = 0, \ldots, m$, $j = 1, \ldots, n$. Note that, as in the parametric case, the distribution of T_i, $i \in I$, is free of μ_j, $j \notin I$, under H_I, so one can simulate data with mean 0, i.e., subset pivotality holds again. The probabilities in the nonparametric algorithm are approximated by the proportions of the B samples in which

$$\max(T_i^*, \ldots, T_m^*) \geq t_{(i)}, \quad i = 1, \ldots, m.$$

To illustrate the bootstrap-based step-down procedure (i.e., the procedure based on nonparametric resampling), consider the dose-finding trial example given in Section 2.5.3. Adjusted p-values produced by the bootstrap-based procedure with $B = 5,000,000$ bootstrap samples are shown in Table 2.6. These adjusted p-values differ little from the adjusted p-values produced by the parametric step-down Dunnett procedure (see Table 2.4).

Bootstrap-based simultaneous confidence intervals

To develop the simultaneous confidence intervals for a general nonparametric case, consider again the ANOVA model in which the errors ε_{ij} are a random sample from an unspecified distribution F having finite variance σ^2. Further, consider the lower limits of one-sided simultaneous confidence intervals for the mean treatment differences $\theta_i = \mu_i - \mu_0$, $i = 1, \ldots, m$, associated with the single-step Dunnett procedure defined in Section 2.7.4. The lower limit for θ_i is given by $\widetilde{L}_i = \widehat{\theta}_i - u_\alpha s_i$, where u_α is the Dunnett-adjusted critical value. The analog to this critical value is $u_\alpha(F)$, where

$$P(\theta_i \geq \widehat{\theta}_i - u_\alpha(F)s_i \text{ for all } i) = 1 - \alpha,$$

or, equivalently,

$$P(\max(T_1, \ldots, T_m) \leq u_\alpha(F)) = 1 - \alpha.$$

Note that the joint distribution of $T_1, \ldots, T_m$ does not depend on the parameters $\mu_1, \ldots, \mu_m$.

Since F is unknown, the critical value $u_\alpha(F)$ is estimated using a bootstrap method, e.g., using the basic bootstrap method defined above. Note that, since F is estimated, the resulting critical values are doubly approximate $\widehat{u}_\alpha(\widehat{F})$, with approximation due to simulation error resulting from B simulations (this error can be reduced with greater B) and due to the approximation of F via $\widehat{F}$ (this error is reduced by increasing the sample size n).

As an illustration, the one-sided 0.025-level bootstrap-based critical value for Scenario 1 in the dose-finding trial example from Section 2.5.3 is 2.45 (this value was computed using $B = 1,000,000$ bootstrap samples). The bootstrap-based critical value is equal to the Dunnett critical value $u_{0.025}(4, 380) = 2.45$, shown in Section 2.7.1, and the bootstrap-based simultaneous confidence intervals are identical to those displayed in Table 2.5.

While the example shows no difference between the parametric and resampling-based approaches, there are cases where resampling matters. These include multiple endpoints, where resampling provides a convenient way to incorporate correlation structure, and adverse events, where sparseness is automatically incorporated to allow much greater power.

2.8.3 Step-down resampling-based procedures for multivariate linear models

Nonparametric resampling-based procedures (bootstrap-based procedures) are more compelling in complex models where there are no standard methods for handling non-normal data and/or complex dependence structure. As in Westfall and Young (1993), consider the general multivariate regression model

$$Y = X\beta + \varepsilon,$$

where Y is a $n \times v$ matrix of response variables, X is a full rank $n \times b$ design matrix, β is a $b \times v$ matrix of regression parameters, and ε is a $n \times v$ matrix of random error terms, all with mean zero. This model subsumes basic ANOVA as well as analysis of covariance models arising in clinical trials when univariate ANOVA models are expanded to include covariates, e.g., demographic and clinical characteristics. Note that n does not have to be greater than v, so the model can be used also for gene expression data where v is typically much larger than n.

Assume that the rows ε_i of ε are independent and identically distributed according to some multivariate distribution F. Specific tests of interest in such models are usually one-dimensional (or single-degree-of-freedom) tests which may be formulated with respect to hypotheses defined as

$$H_i : \ \theta_i = 0, \quad i = 1, \ldots, m,$$

where $\theta_i = c_i' \beta d_i$, $i = 1, \ldots, m$, and c_i and d_i are vectors of constants that specify the hypotheses of interest.

Commonly-used test statistics and unadjusted p-values for each H_i are obtained from the matrix of least squares estimates

$$\widehat{\beta} = (X'X)^{-1} X'Y$$

and the usual unbiased error variance-covariance matrix estimate

$$S = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})/(n - b).$$

The statistic $T_i = \widehat{\theta}_i / s_i$, where

$$\widehat{\theta}_i = c_i' \widehat{\beta} d_i \text{ and } s_i = \sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i},$$

is commonly used to test H_i. When the error distribution F is multivariate normal and H_i is true, T_i has the t distribution with $n - b$ degrees of freedom.

Note that when H_i is true,

$$T_i = \frac{c_i' \widehat{\beta} d_i}{\sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i}} = \frac{c_i' (X'X)^{-1} X' \varepsilon d_i}{\sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i}},$$

showing that subset pivotality holds. Note also that

$$S = \frac{(Y - X\widehat{\beta})'(Y - X\widehat{\beta})}{n - b} = \frac{\varepsilon' \{I - X(X'X)^{-1} X'\} \varepsilon}{n - b},$$

so the distribution of the test statistics is completely determined by the distribution of ε, the design matrix and the chosen comparisons.

It is desirable to account for the correlations between the variables in the multiple comparisons procedure. This can be done via parametric resampling, where error vectors are simulated from a multivariate normal distribution with estimated covariance matrix, or by nonparametric bootstrap-based resampling of error vectors. A simple bootstrap algorithm to test the intersection hypothesis H_I is as follows:

- Step 1. Generate a resampled data set

$$Y^* = 0 + \varepsilon^*,$$

where the rows ε_i^* of ε^* are chosen with replacement from the sample error vectors $\{e_i^*\}$, where e_i^* is the ith row of the sample error matrix $e = Y - X\widehat{\beta}$. Because the test statistics do not depend on β under H_I, there is no need to include $X\widehat{\beta}$ in the resampled data set, unlike other bootstrap procedures.

- Step 2. Compute the statistics $\widehat{\beta}^* = (X'X)^{-1}X'Y^*$, $S^* = (Y^* - X\widehat{\beta}^*)'(Y^* - X\widehat{\beta}^*)/(n-b)$ and

$$T_i^* = \frac{c_i'\widehat{\beta}^* d_i}{\sqrt{d_i' S^* d_i c_i'(X'X)^{-1}c_i}}.$$

Steps 1 and 2 are repeated B times. The bootstrap p-value $\widehat{p}_I$ is defined as the proportion of the B samples in which $T_{\max}^*(I) \geq t_{\max}(I)$, where $T_{\max}^*(I) = \max_{i\in I} T_i^*$ and $t_{\max}(I) = \max_{i\in I} t_i$. Again, as described in Section 2.8.1, testing can be performed sequentially, based on ordered observed test statistics.

As an example, consider the Phase II clinical trial described in Westfall et al. (1999, Section 11.3). In this trial, the efficacy of a treatment was evaluated using four endpoints labeled Y1 through Y4 (Endpoint Y4 is reverse-scored). Table 2.7 displays the raw and adjusted p-values produced by the bootstrap-based procedure with $B = 1,000,000$ bootstrap samples and Holm procedure (again, the adjusted p-values are monotonicity-enforced, as with the step-down Dunnett procedure). The main competitor for the bootstrap-based step-down procedure is the Holm step-down procedure, but the latter loses power because it does not account for correlations. Unlike the previous examples with the Dunnett procedure where there was virtually no difference between the bootstrap method and the parametric counterpart, a clear advantage of the bootstrap method is seen in this example.

2.8.4 Permutation-based procedures

Permutation-based resampling provides an ideal method for constructing multiple testing procedures with multivariate two-sample data, including multiple endpoints and adverse events, since

- the conditions giving rise to subset pivotality can be relaxed relative to the location-shift models presented above,

- the resulting multiple comparisons method is exact, with no finite-sample error incurred from estimating F via $\widehat{F}$,

TABLE 2.7: One-sided raw p-values and adjusted p-values produced by the Holm and bootstrap-based procedures for Endpoints Y1–Y4. The asterisk identifies the adjusted p-values that are significant at the 0.025 level.

Endpoint	Raw	Adjusted p-value	
	p-value	Holm	Bootstrap
Y1	0.0060	0.0242*	0.0186*
Y2	0.0071	0.0242*	0.0186*
Y3	0.0993	0.0993	0.0988
Y4	0.0095	0.0242*	0.0186*

- the resulting methods can be exceptionally more powerful than parametric counterparts when data are sparse, in particular when the data are binary.

Consider the following two-sample problem. Let $Y_i = (Y_{i1}, \ldots, Y_{in})$ denote the multivariate v-dimensional data vectors in the ith sample, $i = 1, 2$. One might assume the general location-shift model

$$Y = X\beta + \varepsilon$$

described in Section 2.8.3, where the matrix X has two columns of indicator (dummy) variables, and where the v-dimensional rows of ε are independent identically distributed variables, but this is somewhat restrictive. Instead, following Westfall and Troendle (2008), we assume a family of distributions for

$$(Y_1, Y_2) = (Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots, Y_{2n})$$

with minimal assumptions. To introduce the assumptions, for any $I \subseteq \{1, \ldots, m\}$, let Y_{ij}^I denote the subvector of Y_{ij} with elements in I and

$$(Y_1^I, Y_2^I) = (Y_{11}^I, \ldots, Y_{1n}^I, Y_{21}^I, \ldots, Y_{2n}^I).$$

The null hypotheses tested will be

$$H_i: \text{ the distribution of } (Y_1^{\{i\}}, Y_2^{\{i\}}) \text{ is exchangeable.}$$

The null hypothesis says that the treatment difference in the two groups has no effect whatsoever on the ith variable. This is a natural null hypothesis for binary and nominal data. However, with interval data, if there is interest only in differences in means and not standard deviations, the permutation test may not be appropriate.

Since intersections are required for closed testing, and since multivariate permutation tests will be used for the intersections, we need to make only the

following assumption about the model. Assume that, if for $I, J \subseteq \{1, \ldots, m\}$ the distribution of (Y_1^I, Y_2^I) is exchangeable in its $2n$ elements and the distribution of (Y_1^J, Y_2^J) is exchangeable in its $2n$ elements, then the distribution of $(Y_1^{I\cup J}, Y_2^{I\cup J})$ is also exchangeable in its $2n$ elements. In particular, the assumption implies that

$$\cap_{i\in I} H_i = H_I : \text{ the distribution of } (Y_1^I, Y_2^I) \text{ is exchangeable,}$$

for any subset I.

Like all assumptions, this one may be questionable, but it should be noted that

- the model is substantially more general than the multivariate location-shift model, which is a special case of this model,
- it is perhaps not unrealistic to assume, e.g., that if there is no difference in treatment effect for each of variables $\{1, 2\}$, then the joint distribution of $(Y_1^{\{1,2\}}, Y_2^{\{1,2\}})$ is exchangeable,
- unlike most statistical models, no assumption of independence is needed.

To define a closed testing procedure for this two-sample problem, define a test statistic $t_i = t_i(Y_1^{\{i\}}, Y_2^{\{i\}})$ for each variable, with larger values suggesting non-exchangeability, and test each intersection hypothesis H_I using the maximum test statistic $t_{\max}(I) = \max_{i\in I} t_i$. The test statistics can be quite general, and are often defined in terms of the p-values themselves. The exact permutation p-value for the test of H_I is p_I which is defined as the proportion of the $(2n)!$ permutations of the data vectors

$$(y_{11}^I, \ldots, y_{1n}^I, y_{21}^I, \ldots, y_{2n}^I)$$

that yield $\max_{i\in I} T_i^* \geq \max_{i\in I} t_i$. Since the subset pivotality condition is satisfied and maximum test statistics are used, the shortcut methods described in Section 2.8.1 apply here as well. Complete enumeration of the $(2n)!$ permutations is not usually feasible, so p_I is instead typically approximated as follows:

- Step 1. Generate a resampled data set Y_{ij}^*, $i = 1, 2$, $j = 1, \ldots, n$, by sampling *without replacement* from the observed vectors $\{y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2n}\}$.
- Step 2. Compute the statistics T_i^* from the Y_{ij}^*.

Repeat the two steps B times and define the exact permutation p-value p_I as the proportion of the B samples where $\max_{i\in I} T_i^* \geq \max_{i\in I} t_i$.

In contrast to the approximate p-values for the bootstrap-based procedures in Sections 2.8.2–2.8.3, the permutation p-values are exact when the $(2n)!$ permutations are completely enumerated. The algorithm above can approximate

TABLE 2.8: Two-sided raw p-values and adjusted p-values produced by the Holm and permutation-based procedures for five adverse events. The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

Adverse event	Raw p-value	Adjusted p-value	
		Holm	Permutation
AE1	0.0008	0.0210*	0.0021*
AE8	0.0293	0.7912	0.1340
AE6	0.0601	1.0000	0.2615
AE5	0.2213	1.0000	0.6279
AE10	0.2484	1.0000	0.9276

p_I with arbitrary precision. Thus, the permutation-based procedure is effectively exact, incorporates relevant correlation between variables, and can be used even when v is much larger than $2n$, e.g., for gene expression data.

See Puri and Sen (1971) for further details on multivariate permutation tests. Further details and applications of resampling-based testing are given in Westfall and Young (1993). Resampling is also quite useful in constructing multiple testing procedures that control generalized definitions of the family-wise error rate, e.g., gFWER or FDP defined in Section 2.2.2. Van der Laan et al. (2004) gave methods to adapt procedures that control the FWER to control the gFWER or FDP by enlarging the rejection set. Korn et al. (2004) showed how more powerful procedures can be obtained in a straightforward fashion.

As an illustration, we will consider the adverse event data set provided by Westfall et al. (1999, Section 12.2). There are two groups, control and treatment, with 80 patients in each group, and $m = 28$ adverse event variables (binary indicators) per patient. Null hypotheses are that the adverse events are exchangeable in the combined sample, tested using Fisher exact upper tailed p-values, with smaller p-values indicating more adverse events in the treatment group. Raw p-values, Holm step-down p-values and permutation-adjusted step-down p-values for the five most significant adverse events labeled AE1, AE8, AE6, AE5 and AE10 are shown in Table 2.8. The adjustment is performed using the minimum p-value method, which is identical to the method based on the maximum test statistic, where the test statistics q_j's are defined by $q_j = 1 - p_j$ and p_j's are the unadjusted p-values from the permutation test.

There is a substantial benefit in using permutation-based step-down testing rather than the Holm method as is seen by comparing the smallest adjusted p-values in Table 2.8. This occurs because the permutation-based method takes advantage of the discreteness of the data. Permutational methods implicitly exploit sparseness. If a particular variable has a permutational distribution

that does not contain any large test statistic values, it is effectively removed from the maximum in the calculation of maximum test statistics (Westfall and Wolfinger, 1997).

2.9 Software implementation

This section describes software implementation of multiple testing procedures described in this chapter with emphasis on SAS and R.

2.9.1 Multiple comparison procedures in SAS

PROC MULTTEST supports a host of popular p-value-based procedures described in Section 2.6 and resampling-based procedures introduced in Section 2.8. This includes the calculation of adjusted p-values for the Bonferroni, Holm, Hochberg and Hommel procedures as well as simultaneous confidence intervals and corresponding adjusted p-values for resampling-based procedures.

P-value-based and parametric procedures are available in other SAS procedures. PROC GLM and PROC MIXED support adjusted p-values and simultaneous confidence intervals for the Bonferroni and single-step Dunnett procedures (covariate-adjusted if needed) in linear and mixed-linear models. Further, PROC GLIMMIX completely subsumes PROC MIXED and PROC GLM and enables the user to implement the p-value-based and parametric procedures supported by these two procedures as well as other multiple testing procedures, e.g., it supports the extended Shaffer-Royen analysis (Section 2.7.3). In cases where the exact distributions are unavailable, these SAS procedures use simulation (often with control-variate variance reduction to improve accuracy) to obtain critical values and adjusted p-values. Monte Carlo errors can be made negligible by specifying simulation sizes in the millions or even billions.

The following is a list of SAS programs that were used to perform multiplicity adjustments in the examples included in this chapter. The programs can be downloaded from the book's Web site (http://www.multxpert.com).

- Program 2.1 computes adjusted p-values for the Bonferroni, Holm, fixed-sequence, fallback, Hommel and Hochberg procedures in the dose-finding trial example (Section 2.6.10).
- Program 2.2 calculates lower limits of one-sided simultaneous confidence intervals for the Bonferroni, Holm, fixed-sequence and fallback procedures in the dose-finding trial example (Section 2.6.11).
- Program 2.3 computes adjusted p-values for the single-step and step-

down Dunnett procedures in the dose-finding trial example (Section 2.7.4).

- Program 2.4 derives lower limits of one-sided simultaneous confidence intervals for the single-step and step-down Dunnett procedures in the dose-finding trial example (Section 2.7.4).

- Program 2.5 implements the resampling-based procedures discussed in Section 2.8.

2.9.2 Multiple comparison procedures in R

R is a language and environment for statistical computing and graphics (Ihaka and Gentleman, 1996). It provides a wide variety of statistical and graphical techniques, and is highly extensible. The latest version of R is available at the Comprehensive R Archive Network (CRAN), which can be accessed from

http://www.r-project.org/

In this section we illustrate briefly the use of the *multcomp* package, which provides a variety of multiple comparison procedures for the linear and other (semi-)parametric models described in Section 2.7.5. The most recent version of the *multcomp* package is available at CRAN under the contributed packages section. For a detailed discussion of multiple comparison procedures in R we refer to Bretz, Hothorn and Westfall (2010).

We consider Scenario 1 in the dose-finding trial example from Section 2.5.3 to illustrate some of the capabilities of the *multcomp* package. We first load the *multcomp* package with the command

```
R> library(multcomp)
```

Suppose that we have a data frame `data` containing the individual observations, where the variables `resp` and `dose` denote the response and the dose group, respectively. To analyse the data, we fit a standard analysis-of-variance model with the `aov` function,

```
R> data.aov <- aov(resp ~ dose, data = data)
```

Assume that we want to perform the single-step Dunnett procedure from Section 2.7.1. The `glht` function from *multcomp* takes the fitted response model `data.aov` to perform the multiple comparisons. To be more precise, we can call

```
R> data.mc <- glht(data.aov, linfct = mcp(dose = "Dunnett"),
+    alternative = "less")
```

In this statement, we used the `mcp` function for the `linfct` argument to specify the comparison type we are interested in. Since we are interested in the Dunnett procedure, we pass over the argument `Dunnett`. Other pre-defined comparison types are also available in *multcomp*; alternatively, the constant matrix C introduced in Section 2.7.5 can be specified manually, to exactly determine the experimental questions of interest. Note that this is a one-sided testing problem and we are interested in showing an increase in the mean HDL cholesterol level; therefore we pass the `alternative = "less"` argument to `glht`.

A detailed summary of the results is available from the `summary` method associated with the `glht` function:

```
R> summary(data.mc)
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = resp ~ dose, data = data)

Linear Hypotheses:
           Estimate Std. Error t value p value
1 - 0 <= 0    2.899      1.445   2.006  0.0714 .
2 - 0 <= 0    3.140      1.445   2.173  0.0493 *
3 - 0 <= 0    3.561      1.445   2.465  0.0242 *
4 - 0 <= 0    3.813      1.445   2.639  0.0151 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

The output shows the observed mean differences, standard errors and t statistics for Scenario 1 in Table 2.1. In the last column, entitled `p value`, the adjusted p-values for the single-step Dunnett procedure are reported, which coincide with the values reported for Scenario 1 in Table 2.4.

In addition, we can compute one-sided 97.5% simultaneous confidence intervals by using the `confint` method associated with the `glht` function:

```
R> data.ci <- confint(data.mc, level = 0.975)
R> data.ci
         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = resp ~ dose, data = data)

Estimated Quantile = 2.4513
97.5% family-wise confidence level
```

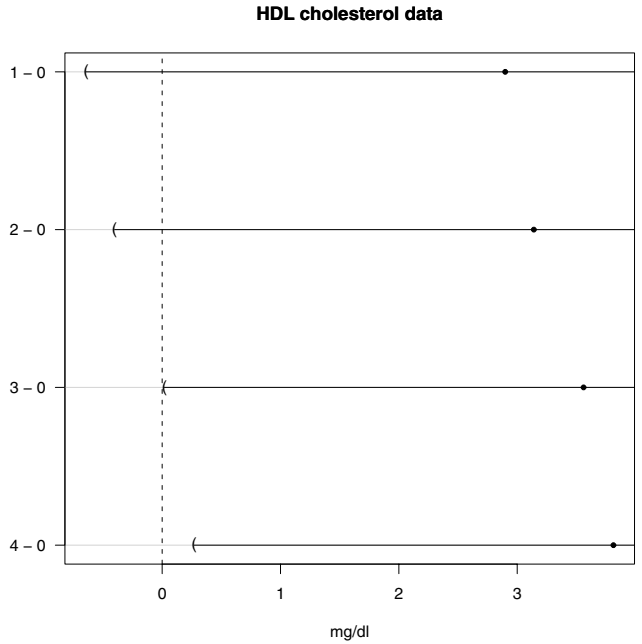

FIGURE 2.6: One-sided 97.5% simultaneous confidence intervals for the dose-finding trial example (Section 2.5.3, Scenario 1).

```
Linear Hypotheses:
           Estimate lwr      upr
1 - 0 <= 0  2.89870 -0.64324      Inf
2 - 0 <= 0  3.14026 -0.40168      Inf
3 - 0 <= 0  3.56104  0.01909      Inf
4 - 0 <= 0  3.81299  0.27104      Inf
```

The simultaneous lower limits match with the values displayed in Table 2.5. We can also display the confidence intervals graphically with

```
R> plot(data.ci, main = "HDL cholesterol data", xlab = "mg/dl")
```

see Figure 2.6 for the resulting plot.

So far we have illustrated only the single-step Dunnett procedure accounting for the correlation among the test statistics. As described in Section 2.7.2, the step-down Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure. Using *multcomp*, we can perform the step-down Dunnett procedure by calling

```
R> summary(data.mc, test = adjusted(type = "free"))
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = resp ~ dose, data = data)

Linear Hypotheses:
           Estimate Std. Error t value p value
1 - 0 <= 0    2.899      1.445   2.006  0.0280 *
2 - 0 <= 0    3.140      1.445   2.173  0.0280 *
3 - 0 <= 0    3.561      1.445   2.465  0.0190 *
4 - 0 <= 0    3.813      1.445   2.639  0.0152 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- free method)
```

The option `type = "free"` leads to a step-down procedure under the free combination condition, which incorporates correlations. In this example we used the Dunnett contrasts, and the results from the previous call coincide with the values reported for Scenario 1 in Table 2.4. If the hypotheses are restricted, truncated closed testing procedures based on Westfall (1997) and Westfall and Tobias (2007) can be performed with the `type = "Westfall"` option; see also Section 2.7.3. In combination with the parametric procedures described in Section 2.7.5, the *multcomp* package thus provides powerful stepwise multiple testing procedures for a large class of parametric models, including generalized linear models, mixed models, and survival models; see Bretz, Hothorn and Westfall (2010) for further details.

The *multcomp* package also implements some of the p-value-based multiple comparison procedures described in Section 2.6. To be more precise, *multcomp* provides an interface to the multiplicity adjustments implemented by the `p.adjust` function from the `stats` package. Given a set of (raw) p-values, the `p.adjust` function provides the resulting adjusted p-values using one of several methods, including the Bonferroni, Holm, Hochberg and Hommel procedures. In order to perform, for example, the Bonferroni procedure, one can call `summary(data.mc, test = adjusted(type = "bonferroni"))`.

Acknowledgements

James Troendle's research was supported in part by the Intramural Research Program of the National Institute of Child Health and Human Development. Ajit C. Tamhane's research was supported by grants from the National Heart, Lung and Blood Institute.