Xinping Cui, Thorsten Dickhaus, Ying Ding, Jason C. Hsu (Eds.)

Handbook of Multiple Comparisons

To our families
and friends.

Contents

Foreword

This handbook will treat the topics of multiple comparisons, simultaneous and selective inference from a variety of different perspectives. The handbook will be useful for (i) researchers, (ii) students / lecturers, (iii) practitioners. The need for such a systematic treatment of the field originates from the relevance of multiple comparisons in many applications (medicine, industry, economics), and from the diversity of approaches and developments, which shall be described here in a coherent manner.

Preface

This handbook has three parts. The first part deals with general methodology, the second part with applications in medicine, and the third part with further topics.

List of Figures

List of Tables

Contributors

Xinping Cui
University of California
Riverside, California

Thorsten Dickhaus
University of Bremen
Bremen, Germany

Ying Ding
University of Pittsburgh
Pittsburgh, Pennsylvania

Jason C. Hsu
Ohio State University
Columbus, Ohio

Symbols

Symbol Description

HR	Hazard Ratio	RR	Relative Respoonse
RCT	Randomized Controlled Trial	SME	Subgroup Mixable Estimation
RoM	Ratio of Medians		

Part I

General Methodology

1

Partitioning for Confidence Sets, Confident Directions, and Decision Paths

Helmut Finner

German Diabetes Center

Szu-Yu Tang

Pfizer Worldwide Research and Development

Xinping Cui

University of California, Riverside

Jason C. Hsu

The Ohio State University

CONTENTS

Partitioning is a fundamental principle in multiple comparisons. In this chapter, we discuss and illustrate three applications of the partitioning principle corresponding to three motivations.

1.1 Motivations for Partitioning

Partitioning, as the name implies, refers to a partitioning of the entire *parameter* space.

Some scientific problems naturally partition the parameter space. In comparing seven treatments for a disease, for example, if treatment one is the best, then treatment two is not; if treatment three is the best, then treatment seven is not, and so forth. Such natural partitioning provides the first of the three motivations listed below for developing the Partitioning Principle (PP).

Section 1.2 Some scientific problems naturally partition.

Section 1.3 By providing associated confidence sets, partitioning can reduce multiplicity adjustment while guaranteeing control of the directional error rate.

Section 1.4 Partitioning can formulate multiple testing problems so that decision-making automatically follow desirable paths.

Both the Partitioning Principle and the Closed Testing Principle can control the familywise error rate (FWER) in testing multiple hypotheses, while keeping multiplicity adjustment only to the extent that it is needed. (See Chapter 1 of this Handbook for descriptions of the Closed Testing Principle and FWER.) A second motivation for PP, stated in both Takeuchi (1973, 2010) and in Stefansson et al. (1988), is to be able to derive confidence sets associated with multiple tests. A motivation for that, in turn, is making decisions based on confidence sets naturally controls the directional error rate. In contrast, we will cite examples in Section 1.3.5 of multiple tests that control the FWER in testing *equality* null hypotheses but do not control the directional error rate, tests without clearly associated confidence sets because the union of their null hypotheses make up only a (small) part of the entire parameter space.

A third motivation for PP is that some decision-making processes naturally have paths. To assess the efficacy of a medicine, in some (but not all) therapeutic areas it is natural to test doses from high to low in that order.

As another example, efficacy in the primary endpoint would be tested before the secondary endpoint, because efficacy in the secondary endpoint is relevant only if there is efficacy in the primary endpoint. While gate-keeping methods impose rules on closed tests to keep decision-making on paths, the Partitioning Principle can transparently partition the parameter space to channel decision-making onto desirable decision paths.

1.2 Multiple Comparisons with the Best: A Scientific Problem Which Naturally Partitions

Whether a drug starts as an active compound and gets metabolized and eliminated from the body, or starts as an inactive compound and gets metabolized to an active form, patients in subgroups separated by polymorphism of a gene metabolizing the drug might derive differential benefit from that drug. So one might wonder which subgroup or subgroups of patients derive maximum benefit or practically maximum benefit from the drug, and which other subgroups do not.

Most drugs are "soft drugs", active compounds that, after performing their activity, are metabolized into inactive form that is then excreted from the body. Other drugs, such as tamoxifen and clopidogrel, are "pro-drugs", inactive compounds needing to be metabolized to their active form.

The cytochrome P450 family of enzymes (abbreviated as CYP) is associated with the metabolism of many drugs (Nebert 2002). Perhaps the two most prominent genes in the P450 family are 2D6 and 2C19. Efficacy of some high profile drugs have been reported to be impacted by polymorphism in 2D6 and 2C19. For example, Schroth et al. (2009), Hoskins, Carey and McLeod (2009), and Abraham et al. (2010) discuss differential efficacy of tamoxifen for patients with variants of the CYP2D6 gene. Mega et al. (2010), Paré et al. (2010), and Holmes et al. (2011) compare efficacy of Plavix (clopidogrel) for patients with variants of the CYP2C19 gene. This is not surprising, because patients in different subpopulations defined by such polymorphisms will not metabolize the drug at exactly the same rate and therefore will not derive exactly the same benefit from that drug. As John W. Tukey (1992) said:

> Our experience with the real world teaches us – if we are willing learners – that, provided we measure to enough decimal places, no two 'treatments' ever have identically the same long-run value.

Polymorphisms in P450 genes are annotated by the so-called *star-allele* nomenclature. CYP2D6 has more than 90 alleles. CYP2C19 is somewhat less polymorphic. Its major alleles are *1, *2, *3, and *17, with *1 being normal (wild-type), *2 and *3 being loss-of-function, and *17 being gain-of-function alleles.

Paré et al. (2010) obtained 5059 samples from a randomized, double-blind, placebo controlled trial with 12562 patients and studied the effect of CYP2C19 genotype on the efficacy of clopidogrel as measured by cardiovascular outcomes. Paré et al. (2010) classified the population into five metabolizer subgroups according to their CYP2C19 genotype, as shown in Table 1.1. However, one can easily imagine classifying the population into finer subgroups, separating *1/*17 from *17/*17 for example, and separating *2/*17 from *3/*17.

Metabolizer	Alleles
Poor	*2/*2 or *2/*3 or *3/*3
Intermediate	*1/*2 or *1/*3
Extensive	*1/*1
Ultra	*1/*17 or *17/*17
Unknown	*2/*17 or *3/*17

TABLE 1.1: Metabolizer Subgroups Defined by CYP2C19 Polymorphism

So let us say we have k patient subpopulations. Identifying the following three kinds of subgroups will be very useful.

$S^{>}$ The subgroup deriving *the maximum* efficacy from the drug

$S_{<}$ Subgroups deriving *less than* maximum efficacy from the drug

S^{δ} Subgroups deriving *practically the maximum* efficacy

With finite amount of data, we cannot identify $S^{>}$, $S^{<}$, and S^{δ} with 100% confidence. But we can certainly identify these subgroups with $100(1-\alpha)\%$ confidence (in a confidence sets sense). One possibility is to compare every subgroup with every other subgroup, i.e., do an all-pairwise comparisons, and deduce information about $S^{>}, S^{<}, S^{\delta}$. However, since comparisons among bad treatments are not of primary interest, one might ask whether confident $S^{>}, S^{<}, S^{\delta}$ subgroup identifications is possible without deducing the information from all pairwise comparisons. Surprisingly, the answer is "yes", to a large extent.

Let us call the subgroup that receives the most efficacy the "best" subgroup, and think about testing the hypotheses

H_{01}: The 1^{st} subgroup is the best

H_{02}: The 2^{nd} subgroup is the best

$\vdots$

H_{0k}: The k^{th} subgroup is the best

There is only one best subgroup. If the 1^{st} group receives the most efficacy, then the 2^{nd} group does not, and so forth. So exactly one of the hypotheses is true, all others are false.

Therefore, one cannot make more than one type-I error in testing these k null hypotheses, and no multiplicity adjustment is needed for testing these k hypotheses simultaneously. That is, if each of $H_{0i}, i = 1, \ldots, k$, is tested at level-α, FWER is controlled at level-α.

However, testing the i^{th} hypothesis involves comparing the i^{th} subgroup with the other $k-1$ subgroups. And the $k-1$ comparisons are 1-sided, because no subgroup can be better than the "best". So, there is a 1-sided multiplicity adjustment of $k-1$ within each of the k tests. It is less than the the $k(k-1)$ multiplicity adjustment for 1-sided all-pairwise comparisons, or the $k(k-1)/2$ multiplicity adjustment for 2-sided all-pairwise comparisons.

Note that the null hypotheses $H_{0i}, i = 1, \ldots, k$, essentially partition the parameter space. This formulation of multiple comparisons is called Multiple Comparisons with the Best (MCB). The MCB formulation is convenient for us to explain how confidence sets for multiple comparisons can be constructed.

1.2.1 Confidence sets associated with multiple tests

Let Θ denote the parameter space. The connection between a family of tests for a parameter and a confidence set for that parameter is given by the following theorem.

Theorem 1.1 (Lehmann 1986, p. 90, Casella and Berger 2001, p. 421) *Let Θ denote the parameter space and let $\hat{\boldsymbol{\theta}}$ be a random vector whose distribution depends on $\boldsymbol{\theta} \in \Theta$. If $\{\phi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in \Theta\}$ is a family of tests such that*

$$P_{\boldsymbol{\theta}}\{\phi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = 0\} \geq 1 - \alpha$$

for each $\boldsymbol{\theta} \in \Theta$, then

$$C(\hat{\boldsymbol{\theta}}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = 0, \boldsymbol{\theta} \in \Theta\}$$

is a level $100(1-\alpha)\%$ confidence set for $\boldsymbol{\theta}$.

One of the earliest use of this correspondence is by Fieller (1964), to get a confidence interval for the *ratio* of two Normal means. His size-α test for each hypothesized value of the true ratio of means is a clever linear combination of the estimated means, and these tests are then pivoted to obtain the confidence set.

The key point to note is that in order to obtain a confidence set, there needs to be a test for each parameter value of the parameter space, that is, the *family* of tests should partition the parameter space. Note that if each test is actually of *size* α, $P_{\boldsymbol{\theta}}\{\phi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = 0\} = 1 - \alpha$, then the confidence set $C(\hat{\boldsymbol{\theta}})$ is *exact*, that is, it has confidence level exactly equal to $1 - \alpha$. Depending on the choice of the family of tests, the confidence set may or may not be convex.

With Θ denoting the parameter space, let $\mathcal{X}$ denote the sample space. Given $x \in \mathcal{X}$, once the family of tests have been executed, confidence bounds for each parameter can be deduced by calculating its minimum and maximum values in the confidence set.

Lemma 1.1 (Projection-Lemma) *Suppose $C(x)$ is a level $1-\alpha$ confidence set for the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$. Given $x \in \mathcal{X}$, define, for $i \in \{1, \ldots, p\}$,*

$$\begin{aligned} U_i(x) &= \sup\{\eta : \exists\ \boldsymbol{\theta} \in C(x) : \theta_i = \eta\}, \\ L_i(x) &= \inf\{\eta : \exists\ \boldsymbol{\theta} \in C(x) : \theta_i = \eta\}. \end{aligned}$$

Then $D_i(x) = [L_i(x), U_i(x)], i = 1, \ldots, p$, constitute level $(1-\alpha)$ simultaneous confidence intervals for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

Note that even if the confidence set $C(x)$ is *exact*, the simultaneous confidence intervals $(D_1, \ldots, D_p)$ may be conservative, that is, their confidence level may be greater than $1 - \alpha$. The confidence set $D = D_1 \times D_2 \times \ldots \times D_p$ is of course convex.

Now consider a partition $\Theta_1, \ldots, \Theta_M$ of the parameter space, that is,

$$\bigcup_{m=1}^{M} \Theta_m = \Theta$$

and

$$\Theta_i \cap \Theta_j = \emptyset \text{ for all } i \neq j.$$

If parameters in each Θ_m is tested by a *different* family of tests, then these families of tests can be pivoted separately in eacn Θ_m and then combined to yield a confidence set for $\boldsymbol{\theta}$, leading naturally to a Partitioning Principle for multiple comparisons confidence set construction. Letting $\mathcal{C}_{1-\alpha}(\Theta)$ denote all possible level $1 - \alpha$ confidence sets $C(x)$ for the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$, the following Partition-Projection corollary gives a formal guideline for calculating the final confidence bounds.

Corollary 1.1 (Partition-Projection Corollary) *Let $\{\Theta_1, \ldots, \Theta_M\}$ be a partition of the parameter space Θ. For each $m \in \{1, \ldots, M\}$, let $\tilde{C}_m(x)$ be a level $1 - \alpha$ confidence set for $\boldsymbol{\theta}$. Given $x \in \mathcal{X}$, $i \in \{1, \ldots, p\}$, $m \in M_+(x) =$*

$\{m \in M : \tilde{C}_m(x) \neq \emptyset\}$, *define*

$$\begin{aligned} U_{im}(x) &= \sup\{\eta : \exists\ \boldsymbol{\theta} \in \tilde{C}_m(x) : \theta_i = \eta\}, \\ L_{im}(x) &= \inf\{\eta : \exists\ \boldsymbol{\theta} \in \tilde{C}_m(x) : \theta_i = \eta\}, \\ U_i(x) &= \sup_{m \in M_+(x)} U_{im}(x), \\ L_i(x) &= \inf_{m \in M_+(x)} L_{im}(x), \end{aligned}$$

then $(D_1, \ldots, D_p)$ *with* $D_i(x) = [L_i(x), U_i(x)]$ *constitute level* $(1-\alpha)$ *simultaneous confidence intervals for* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

When the family of distributions is a location family of distributions, one can start with one or more tests for a particular hypothesized parameter value and employ equivariance to generate the family of tests for the entire parameter space. In the presence of an unknown (nuisance) scale parameter, usually this hypothesized parameter value is chosen so that a statistic whose distribution depends on neither the location parameters nor the scale parameter (i.e., a pivotal quantity) is available. Pivoting within each subspace then taking their union gives the confidence set.

Theorem 1.2 (Pivoting-Partitioning Confidence Set Construction) *Suppose the distribution of* $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ *does not depend on* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. *Consider a partition* $\Theta_1, \ldots, \Theta_M$ *of the parameter space. If each* $\phi_m(\hat{\boldsymbol{\theta}}) = \phi_m(\hat{\theta}_1, \ldots, \hat{\theta}_p)$ *is a level-*α *test for*

$$H_0 : \theta_1 = \cdots = \theta_p = 0$$

with acceptance region $A_m, m = 1, \ldots, M$, *then a level* $100(1-\alpha)\%$ *confidence region for* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ *is*

$$C(\hat{\theta}_1, \ldots, \hat{\theta}_p) = \bigcup_{m=1}^{M} \left(\{-\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} : \boldsymbol{\theta} \in A_m\} \cap \Theta_m \right).$$

Proof *The test*

$$\phi(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \phi_m(\hat{\theta}_1 - \theta_1^0, \ldots, \hat{\theta}_p - \theta_p^0)$$

is a level-α *test for*

$$H_{\boldsymbol{\theta}^0} : \theta_1 = \theta_1^0, \ldots, \theta_p = \theta_p^0.$$

Therefore, by Corollary 1.1, a level $100(1-\alpha)\%$ *confidence set for* $\boldsymbol{\theta}$ *is*

$$\begin{aligned} C(\hat{\boldsymbol{\theta}}) &= \{\boldsymbol{\theta}^0 : H_{\boldsymbol{\theta}^0} \text{ is accepted } \} && (1.1) \\ &= \bigcup_{m=1}^{M} \{\boldsymbol{\theta}^0 : \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 \in A_m \text{ and } \boldsymbol{\theta}^0 \in \Theta_m\} && (1.2) \\ &= \bigcup_{m=1}^{M} \left(\{-\boldsymbol{\theta} + \hat{\boldsymbol{\theta}} : \boldsymbol{\theta} \in A_m\} \cap \Theta_m \right). && (1.3) \end{aligned}$$

Stefansson et al. (1988) used this Partitioning Principle to construct a confidence set for the step-down version of Dunnett's method, as well as MCB confidence sets which we will show in Section 1.2.2. (See the introductory Chapter 1 for a general description of step-wise methods.) Other examples of uses of the Partitioning Principle to construct multiple comparison confidence sets include Hayter and Hsu (1994), Finner (1994), Finner and Strassburger (2007), and Strassburger and Bretz (2008).

1.2.2 Partition confidence set for MCB

Suppose that under the ith treatment a random sample $Y_{i1}, Y_{i2}, \ldots, Y_{in}$ of size n is taken, where the observations between the treatments are independent. Then under the usual normality and equality of variances assumptions, we have the balanced one-way model (1.4)

$$Y_{ia} = \mu_i + \epsilon_{ia}, \quad i = 1, \ldots, k, \quad a = 1, \ldots, n, \tag{1.4}$$

where μ_i is the effect of the ith treatment, $i = 1, \ldots, k$, and $\epsilon_{11}, \ldots, \epsilon_{kn}$ are i.i.d. normal errors with mean 0 and unknown variance σ^2. We use the notation

$$\hat{\mu}_i = \bar{Y}_i = \sum_{a=1}^{n} Y_{ia}/n,$$

$$\hat{\sigma}^2 = MSE = \sum_{i=1}^{k}\sum_{a=1}^{n}(Y_{ia} - \bar{Y}_i)^2/[k(n-1)]$$

for the sample means and the pooled sample variance.

Suppose we partition of the parameter space by $\Theta_1, \ldots, \Theta_m$ where $\Theta_i = \{\mu_i > \max_{j \neq i} \mu_j\}$, i.e., Θ_i is the part of the parameter space where the i^{th} subgroup is the best.

Within Θ_i we want to test the null hypothesis

$$H_{0i} : \text{the } i^{th} \text{ subgroup is the best}$$

If a larger treatment effect is better, then that null hypothesis becomes

$$H_{0i} : \quad \mu_i > \max_{j \neq i} \mu_j$$

Figure 1.1 shows, for the case of $k = 3$, H_{0i} : the i^{th} subgroup is the best for $i = 2, 3$. The shaded area in Figure 1.1a is Θ_2, while the shaded area in Figure 1.1b is Θ_3.

For every parameter value $(\mu_1, \ldots, \mu_k)$ in Θ_i, Dunnett's size-α test for that parameter value being true has acceptance region

$$A_i = \left\{\hat{\mu}_i - \mu_i > \hat{\mu}_j - \mu_j - d\hat{\sigma}\sqrt{2/n} \text{ for all } j \neq i\right\}$$

where d is the quantile that makes the test size-α.

(a) Null space for $\mu_2 > \max_{j=1,3} \mu_j$ (b) Null space for $\mu_3 > \max_{j=1,2} \mu_j$

FIGURE 1.1: Examples of partitioned MCB null hypotheses

Following the pivoting Theorem 1.2, the parameters within each Θ_i that are not rejected are

$$\left\{\mu_i - \mu_j > \hat{\mu}_i - \hat{\mu}_j - d\hat{\sigma}\sqrt{2/n} \text{ for all } j \neq i\right\} \bigcap \Theta_i$$

Therefore, an exact $100(1-\alpha)$ confidence set is

$$C(\hat{\mu}_1, \ldots, \hat{\mu}_k, \hat{\sigma}) = \bigcup_{i=1}^{k} \left(\left\{\mu_i - \mu_j > \hat{\mu}_i - \hat{\mu}_j - d\hat{\sigma}\sqrt{2/n} \text{ for all } j \neq i\right\} \cap \Theta_i \right).$$

Figure 1.2a shows, for the case of $k = 3$, a particular example of an exact MCB confidence set. Location of "$\times$" is the point estimate of $(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3)$. It is such that $\hat{\mu}_3$ is somewhat larger than $\hat{\mu}_2$ but much larger than $\hat{\mu}_1$. Thus H_{01} : the 1^{st} subgroup is the best is rejected so that

$$\left\{\mu_i - \mu_j > \hat{\mu}_i - \hat{\mu}_j - d\hat{\sigma}\sqrt{2/n} \text{ for all } j \neq i\right\} \cap \Theta_i = \emptyset,$$

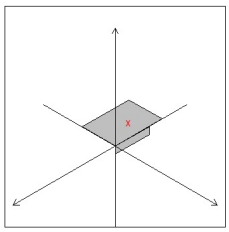
but neither H_{02} nor H_{03} is rejected.

As we will discuss in more detail in the next section, the natural parameters for MCB are $\mu_i - \max_{j \neq i} \mu_j, i = 1, \ldots, k$. Contour of constant $\mu_i - \max_{j \neq i} \mu_j$ is a "$\vee$", with the tip of the $\vee$ on the i^{th} axis. Figure 1.2b shows, given an exact MCB confidence set such as the one in Figure 1.2a, how one would deduce lower and upper confidence bound for $\mu_3 - \max_{j \neq 3} \mu_j$.

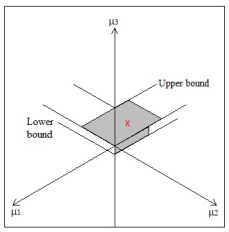
In the next section, we will relate MCB confidence sets derived algebraically in Hsu (1981) and Hsu (1984) with the ones obtained by using the geometrical pivoting technique above.

1.2.3 Multiple Comparisons with the Best

Earlier parameterization of MCB was in terms of "comparisons with the best". Thus, if a larger treatment effect is better, then even though which treatment



(a) Exact MCB confidence set



(b) Bounds for $\mu_3 - \max_{j=1,2} \mu_j$

FIGURE 1.2: Deducing MCB confidence bounds from its exact confidence set

is best is unknown, one could define the parameters of primary interest to be

$$\max_{j=1,\ldots,k} \mu_j - \mu_i, i = 1, \ldots, k, \tag{1.5}$$

the difference between the (unknown) true best treatment effect and each of the k treatment effects. This was the parametrization in Hsu (1981, 1982) and Edwards and Hsu (1983).

However, it turns out to be advantageous to parameterize MCB as "comparison with the best of the *others*". Suppose a larger treatment effect (e.g. survival time) implies a better treatment. Then the parameters

$$\mu_i - \max_{j \neq i} \mu_j, \ i = 1, \ldots, k \tag{1.6}$$

contain all the information that the parameters (1.5) contain, for if

$$\mu_i - \max_{j \neq i} \mu_j > \ 0,$$

then treatment i is the best treatment. On the other hand, if

$$\mu_i - \max_{j \neq i} \mu_j < \ 0,$$

then treatment i is not the best treatment. Further, even if the ith treatment is not the best, but nevertheless

$$\mu_i - \max_{j \neq i} \mu_j > -\delta$$

where δ is a small positive number, then the ith treatment is at least close to the best.

Note that whereas the range of the parameters (1.5) is $[0, \infty)$, the range of the parameters (1.6) is $(-\infty, \infty)$. Thus, the reference point to which confidence

intervals for the parameters (1.6) should be compared is the usual 0. This is one advantage of the parametrization (1.6) over the parametrization (1.5). Another advantage is, as will be shown in Section 1.2.4, lower and upper confidence bounds on the (1.6) parameters correspond to Indifference Zone Selection and Subset Selection respectively. Starting with Hsu (1984), MCB parameterization switched to (1.6) from (1.5).

Contrasts such as those for all-pairwise comparisons (MCA)

$$\mu_i - \mu_j,\ i \neq j$$

and for multiple comparisons with a control (MCC)

$$\mu_i - \mu_k,\ i \neq k,$$

would be straight lines in Figure 1.2b, but multiple comparisons with the best (MCB) parameters (1.6) have "$\vee$" shaped contours, as shown in that figure.

Hsu (1984) showed that the closed intervals

$$[-(\hat{\mu}_i - \max_{j \neq i} \hat{\mu}_j - d\hat{\sigma}\sqrt{2/n})^-, (\hat{\mu}_i - \max_{j \neq i} \hat{\mu}_j + d\hat{\sigma}\sqrt{2/n})^+], i = 1, \ldots, k \quad (1.7)$$

form a set of $100(1-\alpha)\%$ simultaneous confidence intervals for $\mu_i - \max_{j \neq i} \mu_j$. Here $-x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$.

While the derivation in Hsu (1984) was algebraic, if one were to deduce from the exact MCB confidence set not just confidence bounds for $\mu_3 - \max_{j \neq 3} \mu_j$ as in Figure 1.2b but for $\mu_i - \max_{j \neq i} \mu_j$ for all $i = 1, \ldots, k$, then indeed one would get the simultaneous confidence intervals (1.7).

1.2.4 Connection with ranking and selection

Multiple comparison with the best, an early example of what is called *selective inference*, has its roots in ranking and selection, which has two principal formulations: *subset selection*, and *indifference zone selection*.

Let $(1), \ldots, (k)$ denote the unknown indices such that

$$\mu_{(1)} \leq \mu_{(2)} \leq \cdots \leq \mu_{(k)}.$$

In other words, (i) is the *anti-rank* of μ_i among $\mu_1, \ldots, \mu_k$. For example, suppose $k = 3$ and $\mu_2 < \mu_3 < \mu_1$; then $(1) = 2$, $(2) = 3$, $(3) = 1$. For the balanced one-way model (1.4), Subset Selection, due to Gupta (1956; 1965), gives the set

$$G = \left\{ i : \hat{\mu}_i - \max_{j \neq i} \hat{\mu}_j + d\hat{\sigma}\sqrt{2/n} > 0 \right\}$$

as a $100(1-\alpha)\%$ confidence set for (k).[1] Subset Selection basically infers

[1] In traditional Subset Selection literature, when there are multiple indices for $\operatorname{argmax}_i \mu_i$, arbitrarily one such index is "tagged" to be (k).

treatments with indices not in G are not the best treatment. Assuming $\mu_{(k)} > \mu_{(k-1)}$, the subset selection confidence statement

$$\inf_{\boldsymbol{\mu},\sigma^2} P_{\boldsymbol{\mu},\sigma^2}\{(k) \in G\} \geq 1-\alpha$$

is implied by the confidence statement associated with constrained MCB upper bounds

$$\inf_{\boldsymbol{\mu},\sigma^2} P_{\boldsymbol{\mu},\sigma^2}\left\{\mu_i - \max_{j\neq i}\mu_j \leq \left(\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j + d\hat{\sigma}\sqrt{2/n}\right)^+\right\} \geq 1-\alpha,$$

since a non-positive upper bound on $\mu_i - \max_{j\neq i}\mu_j$ implies $i \neq (k)$.

Indifference zone selection, due to Bechhofer (1954), has a *design* aspect and an *inference* aspect.

Traditional Ranking and Selection inferences were on *indices* that correspond to different rankings of the means of the populations; they were not directly on the values of means themselves. Let $[1],\ldots,[k]$ denote the random indices such that

$$\hat{\mu}_{[1]} < \hat{\mu}_{[2]} < \ldots < \hat{\mu}_{[k]}.$$

(Since $\hat{\mu}_i$ are continuous random variables, ties occur in them with probability zero.) In other words, $[i]$ is the anti-rank of $\hat{\mu}_i$ among $\hat{\mu}_1,\ldots,\hat{\mu}_k$. For example, suppose $k=3$ and $\hat{\mu}_2 < \hat{\mu}_1 < \hat{\mu}_3$; then $[1]=2$, $[2]=1$, $[3]=3$. To understand the explanation of the connection between traditional Ranking and Selection inferences and modern MCB inference below, it is important to keep in mind that $[1],\ldots,[k]$ are *random* variables.

For the balanced one-way model (1.4) with σ known, the design aspect of indifference zone selection sets the common sample size n to be the smallest integer such that

$$d\sigma\sqrt{2/n} \leq \delta^\star, \tag{1.8}$$

where $\delta^\star(>0)$ represents the quantity of *indifference* to the user, that is, treatment means within $\delta^\star$ of the best are considered to be equivalent to the best.

Once data has been collected, the inference aspect of indifference zone selection then 'selects' the $[k]$th treatment as the best treatment. The indifference zone confidence statement is that if $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k)$ is in the so-called *preference zone*

$$\{\mu_{(k)} - \mu_{(k-1)} > \delta^\star\},$$

then with a probability of at least $1-\alpha$ the true best treatment will be selected. In other words, the indifference zone confidence statement is

$$\inf_{\mu_{(k)}-\mu_{(k-1)}>\delta^\star} P_{\boldsymbol{\mu},\sigma^2}\{\mu_{[k]} = \mu_{(k)}\} = 1-\alpha. \tag{1.9}$$

This confidence statement is implied by the confidence statement

$$\inf_{\boldsymbol{\mu},\sigma^2} P_{\boldsymbol{\mu},\sigma^2}\{-\delta^\star \leq \mu_{[k]} - \max_{j\neq[k]}\mu_j\} \geq 1-\alpha \tag{1.10}$$

because, for $\boldsymbol{\mu}$ in the preference zone, the only treatment mean μ_i with

$$-\delta^\star \le \mu_i - \max_{j\neq i} \mu_j$$

is $\mu_{(k)}$. While Fabian (1962) gave a direct proof of (1.10), we can see that the confidence statement (1.10) is implied by the confidence statement

$$\inf_{\boldsymbol{\mu},\sigma^2} P_{\boldsymbol{\mu},\sigma^2}\{-d\sigma\sqrt{2/n} \le \mu_{[k]} - \max_{j\neq[k]} \mu_j\} \ge 1-\alpha$$

because

$$d\sigma\sqrt{2/n} \le \delta^\star$$

by indifference zone sample size design (1.8). This last confidence statement in turn is implied by

$$\inf_{\boldsymbol{\mu},\sigma^2} P_{\boldsymbol{\mu},\sigma^2}\{-(\hat{\mu}_{[k]} - \max_{j\neq[k]} \hat{\mu}_j - d\sigma\sqrt{2/n})^- \le \mu_{[k]} - \max_{j\neq[k]} \mu_j\} \ge 1-\alpha,$$

because

$$\hat{\mu}_{[k]} - \max_{j\neq[k]} \hat{\mu}_j \ge 0$$

and

$$-d\sigma\sqrt{2/n} < 0.$$

The last confidence bound now looks familiar. It is one of the constrained lower MCB confidence bounds on $\mu_i - \max_{j\neq i}\mu_j,\ i = 1,\ldots,k$, namely, the one on $\mu_{[k]} - \max_{j\neq[k]}\mu_j$, in the special case of σ known, which can be thought of as the case where the degree of freedom of $\hat{\sigma}$ equals infinity.

In essence, the design aspect of indifference zone selection guarantees a desired accuracy of the MCB lower bound for $\mu_{[k]} - \max_{j\neq[k]}\mu_j$, so that after experimentation, the useful level $1-\alpha$ confidence statement

$$-\delta^\star \le \mu_{[k]} - \max_{j\neq[k]} \mu_j \tag{1.11}$$

can be made with probability one. For single-stage experiments, this can be achieved only when σ is known. When σ is unknown and must be estimated, there is a positive probability that the lower confidence bound

$$\hat{\mu}_{[k]} - \max_{j\neq[k]} \hat{\mu}_j - d\hat{\sigma}\sqrt{2/n}$$

on $\mu_{[k]} - \max_{j\neq[k]}\mu_j$ is less than $-\delta^\star$. However, in analogy with sample size computation based on power of tests, one can design a single-stage experiment so that, with a probability $1-\beta < 1-\alpha$, the lower bound on $\mu_{[k]} - \max_{j\neq[k]}\mu_j$ will be greater than $-\delta^\star$.

Book-length discussions of ranking and selection include Gibbons, Olkin and Sobel (1977), Gupta and Panchapakesan (1979), and Bechhofer, Santner and Goldsman (1995).

Having shown subset selection and indifference zone selection correspond to upper and lower MCB confidence bounds, a most important observation to make at this point is that, since the MCB confidence intervals are guaranteed to cover the parameters $\mu_i - \max_{j \neq i} \mu_j,\ i = 1, \ldots, k$, simultaneously with a probability of at least $1 - \alpha$, subset selection inference and indifference zone selection inference can be given simultaneously with the guarantee that both inferences are correct with a probability of at least $1 - \alpha$.

1.3 Partitioning for Confidence Sets and Confident Directions

Both the Partitioning Principle, as well as the Closed Testing Principle, can be applied to reduce multiplicity adjustment in multiple testing. Idea for both principles is, to control FWER, multiplicity need only be adjusted to the extent that multiple null hypotheses may be simultaneously true. (See Chapter 1 of this Handbook for descriptions of the Closed Testing Principle and FWER.)

If a multiple test is a partition test, then (so long as it partitions the entire parameter space) generally it can be pivoted to have an associated confidence set (using Theorem 1.1 and Corollary 1.1). Why this is important is if multiple testing decision-making is "compatible" with a confidence set, then the directional error rate is controlled.

In testing a new treatment Rx against a control C, merely rejecting the null hypothesis that there is no difference is not useful. One has to make a decision: either Rx is better than C, or Rx is worse than C. In practice, this decision is made according to what the data indicates, upon a rejection of an equality null. If one infers Rx is better than C but in fact Rx is worse than C, then this *directional error* should be counted as an error, and the rate of this incorrect decision should be controlled.

The problem of comparing Rx versus C, for 2-sided inference, is often formulated as a test of equality. There are two issues with that, the first is, as Tukey (1991) said, "Statisticians classically asked the wrong question – and were willing to answer with a lie, one that was often a downright lie ... All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking 'Are the effects different' is foolish. What we should be answering first is 'Can we tell the direction in which the effects of A differ from the effects of B?' " So an equality null is almost surely false. Ding et al. (2018) in fact reported their observation that all equality nulls are statistically false in genome-wide association studies (GWAS), see the discussions in Sections 6.4 and 7.1 of Chapter 15 in this Handbook on testing for SNPs (single nucleotide polymorphisms) testing as well. Both in Ding et al. (2018) and in Chapter 15 of Handbook, instead of tests of equalities, confidence intervals are given instead. Chapter

14 on bioinfomatics and genomics in this Handbook also find exact equalities nulls in medical imaging to be false. Making extensive use of the Partitioning Principles, inferences given in Chapter 14 are also in the form of confidence intervals.

The other issue is, in general, we cannot assume a multiple test that controls the FWER for testing

$$\begin{aligned} H_1^{=} &: \eta_1 = 0 \\ H_2^{=} &: \eta_2 = 0 \\ H_3^{=} &: \eta_3 = 0 \end{aligned} \tag{1.12}$$

automatically controls the directional error rate, because directional errors are not counted in the Type I error definition for testing equality null hypotheses. To control the directional error rate in 2-sided testing, instead of testing equality nulls (1.12), the formulation by Finner (1999) is to test *pairs* of 1-sided nulls:

$$\begin{aligned} H_1^{\leq} &: \eta_1 \leq 0, \quad H_1^{\geq} : \eta_1 \geq 0 \\ H_2^{\leq} &: \eta_2 \leq 0, \quad H_2^{\geq} : \eta_2 \geq 0 \\ H_3^{\leq} &: \eta_3 \leq 0, \quad H_3^{\geq} : \eta_3 \geq 0. \end{aligned} \tag{1.13}$$

Controlling the FWER of testing the null hypotheses (1.13) would indeed control the directional error rate, because direction errors are counted as Type I errors.

Shaffer (1980) proved that, if the test statistics are independent and certain distributional assumptions are satisfied, then some step-down methods for testing equality nulls do control the directional error rate. Providing a unifying framework, Finner (1999) gave conditions under which tests controlling the FWER of testing equality nulls actually control the FWER of testing the paired 1-sided nulls. Intricacy of their proofs, as well as the distributional assumptions required, make clear that control of directional error rate should not be taken for granted in general.

However, to control the FWER of testing a set of null hypotheses using a $100(1-\alpha)\%$ confidence set is simple: reject a null hypothesis H_{0i} if the confidence set[2] does not contain any parameter point in H_{0i}. This is true regardless of whether the null hypotheses are equalities like (1.12), or 1-sided inequalities, or pairs of 1-sided inequalities like (1.13). Therefore, multiple tests based on confidence sets have the advantage that they can control the directional error rate when the problem is properly formulated to do so.

The multiple tests that we recommend in this chapter, such as the step-down version of 1-sided Dunnett's test, are tests with associated confidence sets.

[2]which can be 2-sided simultaneous confidence intervals for 2-sided testing, or 1-sided simultaneous confidence bounds for 1-sided testing

1.3.1 A dose-response motivating example

Schnell et al. (2017) described an Alzheimer Disease (AD) study which compared three doses of a new treatment (doses 1, 2, and 3) with a negative control (dose 0, a placebo). There was an active control (a Standard of Care or SoC, denoted as dose 4 for convenience) in the study as well. Response in this study is *improvement* in ADAS-Cog11 (baseline ADAS-Cog11 minus final ADAS-Cog11) after 24 weeks (a relatively short duration for an AD study). For illustration purpose, we look at the male subset of the data (Sex = 1).

We can model the n_i observed *improvements* from baseline $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ under the i^{th} dose, including dose 0, as a one-way model:

$$Y_{ir} = \mu_i + \epsilon_{ir}, \quad i = 0, \ldots, k, \quad r = 1, \ldots, n_i, \tag{1.14}$$

where μ_i is the mean improvement given the i^{th} dose, $i = 0, \ldots, k$ with $k = 4$, and $\epsilon_{01}, \ldots, \epsilon_{kn_k}$ are i.i.d. normal errors with mean 0 and variance σ^2 unknown. This model differs from (1.4) in that sample sizes may be different. We use the notation

$$\hat{\mu}_i = \bar{Y}_i = \sum_{r=1}^{n_i} Y_{ir}/n_i,$$
$$\hat{\sigma}^2 = MSE = \sum_{i=0}^{k}\sum_{r=1}^{n_i}(Y_{ir} - \bar{Y}_i)^2 / \sum_{i=0}^{k}(n_i - 1)$$

for the sample means and the pooled sample variance.

Of interests are:

- Is there verification that the active control (dose 4) is better than the placebo (dose 0)?
- Which of Doses 1, 2, 3 are better than the placebo (Dose 0)?

Let $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_k)$, and for $i = 1, \ldots, 4$, let

$$H_{0i} : \mu_i \leq \mu_0$$

be the null hypotheses that Dose i is not better than the placebo, and let

$$\Theta_i = \{\boldsymbol{\mu} : \mu_i - \mu_0 \leq 0\}$$

be the corresponding subspace of the parameter space.

Each of the null hypothesis can be tested using a 1-sided t-test, or a lower confidence bound on $\mu_i - \mu_0$. Dunnett's (1955) method adjusts for the multiplicity of testing four null hypotheses, providing four simultaneous confidence bounds, taking correlation of the point estimates into account, producing the results in Table 1.2. From the lower confidence bounds, or the t adjusted p-value, we can infer that Dose 4 (the active control) and Dose 2 are better than the placebo.

However, there are two reasons for thinking beyond Dunnett's method:

Dose	Estimated improvement	$\|t\|$ adjusted p-value	95% 1-sided lower confidence bound	t adjusted p-value
4	5.667	0.0061	1.806	0.0030
3	3.605	0.2076	−0.675	0.1040
2	3.710	0.0991	0.006	0.0496
1	1.757	0.6911	−1.914	0.3663

TABLE 1.2: Analysis of the Alzheimer data set from single-step Dunnett's Method with Dose 0 as the Control

1. In parts of the parameter space where not all four null hypotheses are true, adjustment for multiplicity conceptually can be less than four;
2. In testing to make decision about which doses are better than the placebo, it may seem natural to follow a path of testing in the order of active control → high dose → medium dose → low dose.

Both reasons lead to partitioning of the parameter space, different partitioning. The remainder of this section describes how to partition to reduce multiplicity, whereas in Section 1.4 we will describe how to partition to channel decision-making to follow paths.

1.3.2 Partitioning 1-sided tests without paths

Following Finner and Strassburger (2002), any family of hypotheses $\mathcal{H} = \{H_i : i \in I\}$ generates a natural partition which is the coarsest partition with the property that each H_i can be represented as a disjoint union of the members of the partition.

If $\boldsymbol{\theta} \in \Theta$ is the 'true' parameter, then H_i is said to be *true* if $\boldsymbol{\theta} \in H_i$. The index set $I(\boldsymbol{\theta}) = \{i \in I : H_i \ni \boldsymbol{\theta}\}$ will denote the set of all indices of true null hypotheses if $\boldsymbol{\theta}$ is the true parameter.

Let $\mathcal{J} = \{J : J \subseteq I\}$ and $\Theta_J = \{\boldsymbol{\theta} \in \Theta : I(\boldsymbol{\theta}) = J\}$, $J \subseteq I$. The natural partition is defined by

$$\Theta_{\mathcal{J}} = \{\Theta_J : J \in \mathcal{J}\}.$$

In words, Θ_J consists of parameter points for which $H_i, i \in J$, are true but $H_j, j \notin J$, are false.

Note that $\Theta_\emptyset$ is one of the partitioning subspaces. That is, one of the J

is the empty set $\emptyset$, so $\Theta_\emptyset$ consists of parameter points for which all the null hypotheses are false. While these parameter points are not involved in any test of the original null hypotheses $\mathcal{H} = \{H_i : i \in I\}$, it is better for each parameter in $\Theta_\emptyset$ to be formally tested as well since the construction of confidence set associated with testing $\mathcal{H}$ by Theorem 1.1 expects every parameter point in the parameter space be tested.

Thereby, the closure $\overline{\mathcal{H}}$ of a family $\mathcal{H}$ generates the same natural partition as $\mathcal{H}$. Supposing that $\mathcal{H} = \overline{\mathcal{H}}$, one may set $\Theta_i = H_i \cap (\bigcup_{j:H_j \subset H_i} H_j)^c$ for $i \in I$ and $J_p = \{i \in I : \Theta_i \neq \emptyset\}$. Then the natural partition generated by $\mathcal{H}$ is given by

$$\Theta(J_p) = \{\Theta_i : i \in J_p\},$$

i. e., $\Theta(J_p) = \Theta_{\mathcal{J}}$. If $J_p = I$, then each hypothesis H_i can be identified with Θ_i and vice versa. Note that J_p may be much smaller than I.

So in testing k null hypotheses

$$H_{0i} : \theta_i \leq 0,\ i = 1, \ldots, k, \tag{1.15}$$

for each $I \subseteq \{1, \ldots, k\}, I \neq \emptyset$, form $H^\star_{0I} : \theta_i \leq 0$ for all $i \in I$ and $\theta_j > 0$ for $j \notin I$. There are 2^k parameter subspaces and $2^k - 1$ hypotheses to be tested.

In the Alzheimer Disease study, $\theta_i = \mu_i - \mu_0,\ i = 1, \ldots, 4$. So with $k = 4$ in the Alzheimer study, there are sixteen combinations of these four null hypotheses being true or false. Partitioning divides the parameter space $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ into sixteen disjoint subspaces as depicted in Table 1.3, with a ✓representing a null hypothesis being true and an ✗ representing that null hypothesis being false.

To see explicitly the 16 parameter subspaces, for each of the 16 rows in Table 1.3, we take intersection of Θ_i and their complements where a ⊙ represents Θ_i and a ⊗ represents $\Theta_i^{\mathfrak{c}}$. For example, the second row, with H_{01} and H_{02} and H_{03} being true while H_{04} being false, that combination corresponds to $\boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \Theta_4^{\mathfrak{c}}$ of the parameter space. So the sixteen partitioning parameter subspaces are

$$\begin{aligned}
\Theta_{\{1,2,3,4\}} &= \{\theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \text{ and } \theta_3 \leq 0 \text{ and } \theta_4 \leq 0\} \\
\Theta_{\{1,2,3\}} &= \{\theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \text{ and } \theta_3 \leq 0 \text{ and } \theta_4 > 0\} \\
&\cdots \\
\Theta_{\{1,2\}} &= \{\theta_1 \leq 0 \text{ and } \theta_2 \leq 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 > 0\} \\
\Theta_{\{1,3\}} &= \{\theta_1 \leq 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 \leq 0 \text{ and } \theta_4 > 0\} \\
&\cdots \\
\Theta_{\{3\}} &= \{\theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 \leq 0 \text{ and } \theta_4 > 0\} \\
\Theta_{\{4\}} &= \{\theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 \leq 0\} \\
\Theta_\emptyset &= \{\theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 > 0\}
\end{aligned}$$

with the corresponding partitioning null hypotheses being

$$
\begin{aligned}
H^\star_{\{1,2,3,4\}}: &\quad \theta_1 \le 0 \text{ and } \theta_2 \le 0 \text{ and } \theta_3 \le 0 \text{ and } \theta_4 \le 0\\
H^\star_{\{1,2,3\}}: &\quad \theta_1 \le 0 \text{ and } \theta_2 \le 0 \text{ and } \theta_3 \le 0 \text{ and } \theta_4 > 0\\
&\quad \cdots\\
H^\star_{\{1,2\}}: &\quad \theta_1 \le 0 \text{ and } \theta_2 \le 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 > 0\\
H^\star_{\{1,3\}}: &\quad \theta_1 \le 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 \le 0 \text{ and } \theta_4 > 0\\
&\quad \cdots\\
H^\star_{\{3\}}: &\quad \theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 \le 0 \text{ and } \theta_4 > 0\\
H^\star_{\{4\}}: &\quad \theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 \le 0\\
H^\star_{\emptyset}: &\quad \theta_1 > 0 \text{ and } \theta_2 > 0 \text{ and } \theta_3 > 0 \text{ and } \theta_4 > 0
\end{aligned}
$$

In general, for each i, partition testing would infer $\theta_i > 0$ if and only if all $H^\star_{0I}$ with $i \in I$ are rejected, because H_{0i} is the union of $H^\star_{0I}$ with $i \in I$. To infer Dose i to be better than the placebo is to rule out the possibility that $\mu_i - \mu_0 \le 0$, which means $\boldsymbol{\mu}$ does not belong to any of the partition null hypotheses that contain $\mu_i - \mu_0 \le 0$, which in turn means all rows that contain a ✓for H_{0i} is rejected.

These 16 parameter subspaces obviously partition the parameter space, that is, the true $\boldsymbol{\mu}$ is in exactly one of these 16 partition null hypotheses. It is impossible for Dose 2 to be better than the placebo and at the same time to be worse than the placebo, for example. Therefore, each of the partitioned null hypothesis can be tested at level-α while controlling FWER at level-α, no multiplicity adjustment is needed (even though there are 16 partitioning null hypotheses). There is multiplicity adjustment within the test for most rows in Table 1.3 though, because partitioning null hypothesis such as $H^\star_{\{1,2,3\}}: \boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \Theta_4^{\mathbf{c}}$ implies the three null hypotheses $H_{0i}: \theta_i \le 0,\ i = 1, \ldots, 3$ are simultaneously true. However, for this particular partitioning null hypothesis, the extent to which multiplicity needs to be adjusted is three, not four. For testing $H^\star_{\{2,3\}}$, multiplicity adjustment is two, not four, for example. So compared to Dunnett's (1955) single-step method, which in essence adjusts for a multiplicity of four for all $H^\star_I$, partition testing potentially reduces multiplicity adjustment.

How to test any null hypothesis is not unique (one can flip a coin, for example, but that would be silly). Regardless of how each partitioning null hypothesis is tested, we can invoke Theorem 1.1 to obtain a corresponding confidence set.

Weak partition testing makes use the fact that a level-α test for the *intersection* null hypothesis

$$H_{\{1,2,3\}}: \boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3 \tag{1.16}$$

(which does not specify whether $\boldsymbol{\mu} \in \Theta_4$ or not) actually is also a level-α test for the *partitioning* null hypothesis

$$H^\star_{\{1,2,3\}}: \boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \Theta_4^{\mathbf{c}} \tag{1.17}$$

for example. This is because a level-α test for (1.16) would not reject with a probability greater than α when $\boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3$ regardless of the value of μ_4, so it would not reject with a probability greater than α when $\boldsymbol{\mu}$ is in the subset $\boldsymbol{\mu} \in \Theta_1 \cap \Theta_2 \cap \Theta_3 \cap \Theta_4^{\mathbf{c}}$ with $\mu_4 > \mu_0$ in particular. Weak partition testing tests each of the 15 partitioning null hypotheses such as (1.17) at level-α by testing its corresponding intersection null hypothesis (1.16) at level-α.

Still, tests for the intersection null hypotheses are not unique. They could be F-tests, or max-T/min-P tests, for example. Technically, it would not be wrong to use an F-test to test $H_{\{1,2,3,4\}}$ and use max-T/min-P tests for the remaining intersection null hypotheses, for example, *so long as all 15 tests are executed without taking shortcuts*. What has caused confusion was that two legacy multiple tests, Holm's step-down method and Hochberg's step-up method, appear to execute only k tests based on the ordered p-values. In reality, those k tests are shortcuts to all $2^k - 1$ tests (see Huang and Hsu 2007). Without that realization, there were some incorrect shortcutting early on (see Chapters 3 and 4 of Hsu 1996). So we will use the analysis of the Alzheimer study to illustrate when and how to take legitimate shortcuts in executing a partition test.

Holm's step-down method adjusts for multiplicity within the test for each partitioning null hypothesis $H_I^\star$ by the Bonferroni inequality, while Hochberg's step-up method adjusts for multiplicity using a conservative modification of Simes' equality (see Huang and Hsu 2007). Neither method takes the correlations among the test statistics into account. Under model (1.14) though, joint distribution of the test statistics is readily computable. We thus illustrate partition testing using Dunnett's method to test each $H_I^\star$, to take the joint distribution of the test statistics into account. See Chapter 3 on multivariate methods in this Handbook for a comprehensive discussion of multiple tests that take joint distribution into account.

Conditions for taking shortcuts

To take step-down shortcuts, pretty much the form of the test for each partitioning null hypothesis $H_I^\star$ needs to be a maximum T (maxT) or, equivalent in form, a minimum p-value (minP) test. F-tests, which are based on sums of squares, do not allow shortcuts.

Shortcut condition 1 For the individual null hypothesis $H_{0i}^\star$ that has the largest test statistic value or, equivalent in form, the smallest p-value in testing a partitioning null hypothesis $H_I^\star$ with $I \ni i$, it remains the null hypothesis having the largest test statistic and the smallest p-value in testing any other partitioning null hypothesis $H_J^\star$ with a smaller set $J, J \subset I$;

Shortcut condition 2 For this individual null hypothesis $H_{0i}^\star$, its adjusted p-value in testing $H_J^\star$ with any $J \subset I$ is no larger than its adjusted p-value in testing $H_I^\star$.

Even with a maxT/minP test for each $H_I^\star$, there is subtlety involved in

executing a multiple test to meet the shortcut conditions, which we will illustrate in executing a step-down version of Dunnett's method. (See Chapter 1 of this Handbook for a description of what are called step-wise methods.)

1.3.3 Confident decision-making based on step-down Dunnett's method

Even with the null hypotheses being the 1-sided (1.15), that is, with the intention being to infer which doses are better than the placebo, current practice is still to execute the testing as 2-sided.

There is a perception that controlling the FWER of 2-sided testing of equality nulls at level-α controls the FWER of testing 1-sided nulls at level-$\alpha/2$. This perception is slightly wrong if the equality nulls are tested by a confidence intervals method, but can be quite wrong if the equality nulls are tested by a method based on p-values without an associated confidence set. Basically, if the 1-sided method on which 2-sided testing is based has an associated confidence set, then the 1-sided FWER (including the directional error rate) is (to a close approximation) $\alpha/2$. But such is not necessarily the case if 2-sided testing is based on p-values for testing equalities.

The single-step Dunnett's method produces confidence intervals. In general, if we use the lower confidence bounds of $100(1-\alpha)\%$ simultaneous 2-sided confidence intervals to test 1-sided nulls (in any particular $H^{\star}_{0I}$), the 1-sided Type I error rate (for testing that $H^{\star}_{0I}$) is close to but not exactly $100(\alpha/2)\%$, as can be seen by the fact that the 2-sided $|t|$ adjusted $p-$values are not exactly twice the 1-sided t adjusted p-values in Table 1.2. With equal-tailed confidence intervals and $\pm$ symmetry in the joint distribution, it actually can be shown that this practice is slightly liberal. However, our experience has been that the liberalism is mostly slight, of not a big concern. So, to reflect current practice, we test the 1-sided null hypotheses (1.15) using the appropriate "side" of a method which has an associated confidence set.

Based on the partitioning principle, Chapter 3 of Hsu (1996) derived simultaneous confidence bounds for a step-down version of 1-sided Dunnett's method, using the Partition Projection corollary 1.1.[3] So, using the lower bounds of 90% 2-sided Dunnet simultaneous confidence intervals to test each H_{0I}, the 1-sided FWER including the directional error rate for testing (1.15) will be (approximately) 5%. However, in Section 1.3.5, we will explain the danger of using a method based on p-values for tests of equality nulls without an associated confidence set.

We also note (in passing) that, if one were to view the execution of step-down Dunnett's method controlling FWER for testing the equality nulls (1.12) at level-α as intended for 2-sided inference, then it would be non-trivial to prove that the 2-sided directional error rate is controlled at level-α (i.e., it

[3]Technically, the derivation in Hsu (1996) was for a balanced one-way design, but the idea generalizes.

would be non-trivial to prove the FWER for testing the *paired* 1-sided nulls (1.13) is controlled at level-α). The reason for that is such proofs typically assume there is balance in the design (such as equal sample sizes), or even independence among the test statistics. Neither is true is in our real life Alzheimer study example.

Table 1.4 facilitates the execution of partition testing using the 2-sided Dunnett's method for testing each partitioning $H_I^\star$. To be clear, our intended inferences are 1-sided so the problem is formulated as testing the 1-sided null hypotheses (1.15). We use the lower confidence bounds of 90% 2-sided confidence intervals to test each each $H_{0I}^\star$ at (approximately) 5%, so that the FWER of testing the 1-sided nulls (1.15) is approximately 5%.

The trade-off between the single-step Dunnett's method and its step-down version is that, while the step-down version potentially can infer more doses to be better than the placebo (the negative control), it gives up the ability to give strictly positive lower bounds. That is, the step-down version infers Dose i to be better than the placebo by giving the inference $\mu_i - \mu_0 > 0$. Instead of going through details of the derivation, we give an intuitive explanation of why this is so.

In accordance with Theorem 1.1, to provide positive lower confidence bounds for $\mu_i - \mu_0$, one has to test for possible positive values of $\mu_i - \mu_0$, such as $\mu_i - \mu_0 = 0.01, 0.02, \ldots$. Dunnett's method does that, testing each parameter configuration $(\mu_i - \mu_0 = \mu_i^\star - \mu_0^\star, i = 1, \ldots, k)$, for all possible values of $\mu_i^\star - \mu_0^\star$ positive and negative, and then applying the pivoting Theorem 1.1 to get the lower bounds.

Both closed testing and partition testing can potentially infer more doses to be better than the placebo than Dunnett's method. (See the introductory Chapter 1 for a description of closed testing.) How they do that is by *not* testing for possible positive values of $(\mu_i - \mu_0, i \notin J)$, in parts of the parameter space where $(\mu_i - \mu_0, i \notin J)$ are positive (thus reducing multiplicity adjustment in testing $H_J^\star$ comparing to testing $H_{\{1,\ldots,k\}}^\star$). To wit, $(\mu_i - \mu_0, i \notin J) > 0$ in $H_J^\star$ with $J \subset I$, so closed testing and (weak) partition testing do not bother testing for $(\mu_i - \mu_0, i \notin J)$ in testing $H_J^\star$, thus reducing multiplicity adjustment but giving up the ability to provide positive lower bounds for them. However, the partitioning version of the step-down 1-sided Dunnett's method is a confidence set method, so it controls the directional error rate. That is the important point.

1.3.4 Executing step-down Dunnett's method for the Alzheimer study

The subtlety in execution alluded to after stating the Shortcut conditions is that tests for all $H_I^\star, I \subseteq \{1, \ldots, k\}$, should be executed by fitting the entire data to the model. The reason for this is if, instead, testing for $H_J^\star$ is done by fitting only data involved in $H_J^\star$, then even with point estimates for treatment

effects remaining the same, estimates for σ^2 would differ for different J, and shortcut condition 2 can be violated.

Suppose, for example, $\hat{\sigma}^2$ for $H^\star_{\{1,2,3\}}$ is computed based on Dose 1 and 2 and 3 data, while $\hat{\sigma}^2$ for $H^\star_{\{1,2\}}$ is computed based on Dose 1 and 2 data. Then the two $\hat{\sigma}^2$ would differ in value and in degrees of freedom, and p-value of H_{02} adjusted for $H^\star_{\{1,2\}}$ may be larger than the p-value of H_{02} adjusted for $H^\star_{\{1,2,3\}}$, so shortcut condition 2 may not hold.

We thus fit the entire data set to the model (1.14) throughout our demonstration of how using Dunnett's method to test each $H^\star_I$ has some shortcuts, facilitated by the adjusted p-values displayed in Table 1.4, which are to be compared with 0.10 for 1-sided FWER $\approx 5\%$.

Note that the SAS codes in Program 12.9 of Dmitrienko et al. (2007) and the codes in Program 14.5 of Westfall et al. (2011) are meant for studies with balanced designs only. If one so desires, one can follow the concept demonstrated below to write his/her own codes for studies that are not perfectly balanced and/or have covariates, fitting the entire data set to a model, and specifying contrasts for each $H^\star_I$ that needs to be tested.

Step 1 Dose 4 has the smallest p-value. Its p-value adjusted for testing $H^\star_{\{1,2,3,4\}}$ is 0.0061, so $H^\star_{\{1,2,3,4\}}$ is rejected at the 2-sided $\alpha = .10$ level. The adjusted p-value for Dose 4 in testing any $H^\star_J$ with $J \subset \{1,2,3,4\}$ would be smaller than 0.0030, so all $H^\star_J$ with $J \subset \{1,2,3,4\}$ would be rejected as well. We thus know all eight of the partitioning null hypotheses $H^\star_I$ with $4 \in I$ are rejected. Therefore, we can infer Dose 4 (the active control) to be better than the placebo, with a confidence bound of $\mu_4 - \mu_0 > 0$.

Step 2 Dose 2 has the second smallest p-value. Its p-value adjusted for testing $H^\star_{\{1,2,3\}}$ is 0.0783, so $H^\star_{\{1,2,3\}}$ is rejected at the 2-sided $\alpha = .10$ level. The adjusted p-value for Dose 2 in testing any $H^\star_J$ with $J \subset \{1,2,3\}$ would be smaller than 0.0783, so all $H^\star_J$ with $J \subset \{1,2,3\}$ would be rejected as well. We know from earlier that all $H^\star_I$ with $4 \in I$ are rejected, and we now know that among the remaining $H^\star_I$, those with $2 \in I$ are rejected. Therefore, we can infer Dose 2 to be better than the placebo, with a confidence bound of $\mu_2 - \mu_0 > 0$.

Step 3 Dose 3 has the third smallest p-value. Its p-value adjusted for testing $H^\star_{\{1,3\}}$ is 0.1217, so $H^\star_{\{1,3\}}$ fails to be rejected at the 2-sided $\alpha = .10$ level. Thus, we are unable to infer Dose 3 to be better than the placebo.[4] At this point, we might as well stop, not bother testing $H^\star_{\{1\}}$ or $H^\star_{\{3\}}$, because even if either is rejected, we cannot infer either dose to be better than the placebo, because $H^\star_{\{1,3\}}$ fails to be rejected.

So, for the Alzheimer study, 1-sided Dunnett's method and its step-down

[4]Technically, one could compute a lower confidence bound for $\mu_3 - \mu_0$ by projection, but it would be negative (< 0), so is not reported here.

version come to the same inference, both inferring Doses 4 and 2 to be better than the placebo. The single-step Dunnett's method in fact provides more information, in giving positive lower bounds of 1.806 and 0.006 for $\mu_4 - \mu_0$ and $\mu_2 - \mu_0$ (instead of the lower bounds of zero by its step-down version). However, one can see that there is the possibility that the Dose 3 p-value adjusted for $H_{\{1,3\}}$ (instead of adjusted for $H_{\{1,2,3,4\}}$ by the single-step Dunnett's method) can potentially be small enough ($< .10$ instead of being 0.1217) to allow the step-down version to infer Dose 3 to be better than the placebo, had the data turned out a bit differently. Such is the trade-off between single-step and step-down, the potential of more doses inferred to be better than the placebo versus an inability to give strictly positive lower confidence bounds.

1.3.5 Testing equality null hypotheses may not control the directional error rate

Besides reminding readers that multiple tests that have associated confidence sets automatically control the directional error rate, below we describe some real life situations in which (not confidence set based) multiple tests that control the FWER of testing equality nulls may *not* control the directional error rate.

An earlier realization of this danger was documented in Hsu and Berger (1999). In the setting of dose-response studies, a Type I incorrect decision is erroneously inferring a minimum effective dose (MED) that is lower than the true MED. Hsu and Berger (1999) showed that most of the so-called contrasts tests (that were popular then) that technically control the FWER of testing equality null hypotheses have inflated Type I incorrect decision rates.

Let us say we test for efficacy of Rx versus C, and there is a biomarker dividing the patients into a g^+ and a g^- subgroup. We are interested in answering the questions

Q^+ : Is efficacy η_{g^+} in the g^+ subgroup > 0?

Q^- : Is efficacy η_{g^-} in the g^- subgroup > 0?

$Q^\pm$: Is efficacy $\eta_{g^\pm}$ in the overall population $\{g^+, g^-\} > 0$?

If one formulates these questions properly as testing the three 1-sided null hypotheses

$$\begin{aligned} H^+_{\leq} : \eta_{g^+} \leq 0 \quad vs. \quad K^+_{>} : \eta_{g^+} > 0 \\ H^-_{\leq} : \eta_{g^-} \leq 0 \quad vs. \quad K^-_{>} : \eta_{g^-} > 0 \\ H^\pm_{\leq} : \eta_{g^\pm} \leq 0 \quad vs. \quad K^\pm_{>} : \eta_{g^\pm} > 0 \end{aligned} \tag{1.18}$$

then it is possible that

3 true All three nulls are true;

2 true Two out of three are true (e.g., $H^-_{\leq}$ and $H^{\pm}_{\leq}$)

1 true Only one of the nulls is true (e.g. $H^-_{\leq}$)

0 true None is true.

For example, it is certainly possible that Rx is better than C by an amount δ (> 0) in g^+ but worse in g^- by an amount more than δ, in which case not all three null hypotheses are true ($H^+_{\leq}$ is false) but two out of the three nulls ($H^-_{\leq}$ and $H^{\pm}_{\leq}$) are true. So multiplicity adjustment in step-down testing would go from three for [3 true] to two for [2 true] to one for [1 true].

On the other hand, if one formulated these questions as testing three equality null hypotheses,

$$\begin{aligned} H^+_{=} &: \eta_{g^+} = 0 \\ H^-_{=} &: \eta_{g^-} = 0 \\ H^{\pm}_{=} &: \eta_{g^{\pm}} = 0 \end{aligned} \tag{1.19}$$

then any two of the null hypotheses being true implies the third is true[5]. Therefore, *if it is not the case that all three null hypotheses are true, then at most one of the null hypothesis is true.* So, if a test for all three equality null hypotheses (1.19) being true is rejected, then one can go straight to testing the individual nulls with no multiplicity adjustment. In other words, multiplicity adjustment in testing would decrease from three directly to one, too drastic a jump to control the directional error rate. A calculation in Han et al. (2020) shows a multiple test which controls the FWER of testing the 2-sided null (1.19) at 10% has a 1-sided (directional) error rate at least 6.4%.

What Shaffer (1980) and Finner (1999) showed was it is non-trivial to prove a multiple test controlling the FWER of testing the equality nulls (1.12) at α actually controls the directional error rate testing the *pairs* of 1-sided nulls (1.13). What we are cautioning here is a counter-example exists that a 2-sided multiple test controlling the FWER for testing the equality nulls (1.12) at α does not control the directional error rate testing the 1-sided nulls (1.15) at level-$\alpha/2$.

It may also be useful to mention how perspective on taking advantage of logical relationships among equality nulls has evolved. A simple example of such logical relationships similar to but different from our example above is in traditional all-pairwise comparisons. If the comparison of three means μ_1, μ_2, μ_3 is formulated as testing the three equality nulls:

$$\begin{aligned} H_{12} &: \mu_1 = \mu_2 \\ H_{23} &: \mu_2 = \mu_3 \\ H_{31} &: \mu_3 = \mu_1 \end{aligned}$$

[5]provided efficacy measure is logic-respecting as defined in Section 4 of Chapter 13 on Subgroups Analysis in this Handbook, so that $\eta_{g^{\pm}}$ is a weighted average of η_{g^+} and η_{g^-} say

then clearly any two nulls being true implies the third is true. Shaffer (1986) proposed to take advantage of such relationships to reduce multiplicity adjustment. Westfall (1997) and Westfall and Tobias (2007) followed up with computer algorithms for implementation, as `TYPE=LOGICAL` in the `STEPDOWN` option of the `LSMEANS` and `MSMESTIMATE` statements of SAS. However, further follow-up by Westfall et al. (2013) revealed that making use of logical relationships *in testing equality nulls* may cause the directional error rate to not be controlled. So perspective has turned from being positive toward the negative. For 2-sided inference when there are logical relationships among the parameters, a safe approach is to use confidence set methods such as Tukey's (1953) method for all-pairwise comparisons and the methods in Ding et al. (2016) and Lin et al. (2019) for inference on efficacy in subgroups and their mixtures (methods which are described in Chapter 13).

Finally, we point out that an extreme example of a test that is not capable of controlling the directional error rate is the log-Rank test used in survival analysis everyday. It can be thought of as testing infinitely many equality nulls (1.12) between Rx and C, that the survival probabilities are exactly equal at all time points or, equivalently, that the expected survival times are exactly equal for all quantiles. In multiple comparisons, such a null hypothesis is called a *Complete* null, where *all* the null hypotheses are true.[6] Controlling the Type I error rate of testing a *complete* null is termed *weak* control, which may be insufficient to control the incorrect decision rate. Section 8 of Chapter 13 on Subgroups Analysis in this Handbook contains a realistic example showing that a level-5% log-Rank test can have an incorrect (directional) decision rate exceeding 15%, so the danger of the log-Rank test testing a very restrictive equality null can harm patients is real. Instead of reporting p-values based on the log-Rank test while reporting confidence interval based on the Wald test, we suggest reporting confidence sets (which can of course be used to test hypotheses) or employ tests with compatible confidence sets.

1.4 Partition To Follow Decision Paths

In therapeutic areas such as diabetes and hypertension, higher doses generally give larger effects.[7] However, in psychiatric areas such as schizophrenia, true response to increasing dose of a drug, as measured by reduction in Positive and Negative Syndrome Scale (PANSS) for example, may first increase then decrease. See Arvanitis et al. (1997) for an example.

[6]The *complete* null is also called the *global* null. See Chapter 1 of this Handbook.

[7]But too high a dose can be dangerous. For diabetic patients, injecting too much insulin can cause blood sugar level to drop too low and result in hypoglycemia. Too much diuretics for treating hypertension may cause the blood pressure to be too low resulting in syncope (fainting).

In our Alzheimer study example, Dose 2 (medium dose) seems to be a bit more effective than Dose 3 (high dose). Whether that is real, or due to variability in a finite sample, is hard to tell.

In any case, whether the thinking is higher doses correspond to larger effects, or it is awkward to state there is evidence of efficacy at the medium dose but not at the high dose, it is not uncommon for the analysis plan of a clinical study to have a *decision path*, testing for efficacy from high dose to low dose. Testing for efficacy proceeds along the path, stopping as soon as efficacy fails to be established.

How this differs from testing without a path is, testing along a *single* path, FWER is controlled without multiplicity adjustment. A common misconception is this validity depends on an assumption that the true response is monotonically non-decreasing as dose increases. Actually, it is valid without any assumption on the response curve. The simplest proof of this validity is to "ask the questions", by partitioning.

1.4.1 The decision path principle: asking the right questions

A sequence of potential inferences is a *decision path*.

Decision Path Principle: Null hypotheses should be formulated so that decision-making naturally follows decision paths.

Implicitly used in Hsu and Berger (1999), this principle was explicitly stated in Liu and Hsu (2009). Applying this principle changes how the null hypotheses are formulated, by *asking the right questions*.

Suppose dose i is considered effective if $\mu_i > \mu_0 + \Delta$. To logically infer dose k is effective by the rejection of a null hypothesis, the null hypothesis tested has to be H_{0k} : Dose k is ineffective. To logically infer doses k and $k-1$ are effective by the rejection of the null hypotheses H_{0k} and $H_{0(k-1)}$, the union of the null hypotheses H_{0k} and $H_{0(k-1)}$ needs to include the possibilities dose k is ineffective and/or dose $k-1$ is ineffective.

Consider testing the null hypotheses

$$H^{\downarrow}_{0k}: \text{ Dose } k \text{ is ineffective}$$
$$H^{\downarrow}_{0(k-1)}: \text{ Dose } k \text{ is effective but dose } k-1 \text{ is ineffective}$$
$$\vdots$$
$$H^{\downarrow}_{0i}: \text{ Doses } i+1,\ldots,k \text{ are effective but dose } i \text{ is ineffective}$$
$$\vdots$$
$$H^{\downarrow}_{01}: \text{ Doses } 2,\ldots,k \text{ are effective but dose 1 is ineffective}$$

Statistically, the null hypotheses are:

$$\begin{gathered} H_{0k}^{\downarrow} : \mu_k \leq \mu_0 + \Delta \\ H_{0(k-1)}^{\downarrow} : \mu_{k-1} \leq \mu_0 + \Delta < \mu_k \\ \vdots \\ H_{0i}^{\downarrow} : \mu_i \leq \mu_0 + \Delta < \min\{\mu_{i+1}, \ldots, \mu_k\} \\ \vdots \\ H_{01}^{\downarrow} : \mu_1 \leq \mu_0 + \Delta < \min\{\mu_2, \ldots, \mu_k\} \end{gathered} \tag{1.20}$$

Together with

$$H_{00}^{\downarrow} : \mu_0 + \Delta < \min\{\mu_1, \ldots, \mu_k\} \tag{1.21}$$

the null hypotheses (1.20) partition the parameter space, so no multiplicity adjustment is needed in testing them.

For any integer i, if $H_{0j}^{\downarrow},\ j = i, \ldots, k$, are all rejected, then the logical inference is doses $i, \ldots, k$ are all efficacious: $\mu_j > \mu_1 + \Delta,\ j = i, \ldots, k$.

For example, suppose

$$H_{04}^{\downarrow} : \text{ Dose 4 is not efficacious}$$

is rejected. Then obviously one can infer Dose 4 is efficacious.

Suppose

$$H_{04}^{\downarrow} : \text{ Dose 4 is not efficacious}$$

and

$$H_{03}^{\downarrow} : \text{ Dose 4 is efficacious but dose 3 is not efficacious}$$

are both rejected, then since the union of $H_{04}^{\downarrow}$ and $H_{03}^{\downarrow}$ is "either dose 4 or dose 3 is not efficacious," the rejection of $H_{04}^{\downarrow}$ and $H_{03}^{\downarrow}$ implies "both dose 4 and dose 3 are efficacious."

On the other hand, if

$$H_{04}^{\downarrow} : \text{ Dose 4 is not efficacious}$$

is rejected,

$$H_{03}^{\downarrow} : \text{ Dose 4 is efficacious but dose 3 is not efficacious}$$

fails to be rejected, but

$$H_{02}^{\downarrow} : \text{ Doses 4 and 3 are efficacious but dose 2 is not efficacious}$$

is rejected, then still the only useful inference remains Dose 4 is efficacious. So one might as well stop testing when $H_{03}^{\downarrow}$ fails to be rejected. By asking

the right questions (1.20), partition testing automatically follows the decision path.

Level-α tests for each $H_{0i}^{\downarrow},\ i = 1, \ldots, k$, are of course not unique. Note, however, a level-α test for

$$H_{0i} : \mu_i \leq \mu_0 + \Delta \tag{1.22}$$

is also a level-α test for

$$H_{0i}^{\downarrow} : \mu_i \leq \mu_0 + \Delta < \min\{\mu_{i+1}, \ldots, \mu_k\}$$

For example, a test which rejects no more than 5% of the time when dose 3 is ineffective, regardless of whether dose 4 is effective, will reject no more than 5% of the time in particular when dose 3 is ineffective and dose 4 is effective. So the simplest level-α test for $H_{0i}^{\downarrow}$ is to use a one-sided two-sample size-α t-test comparing μ_i with μ_0 for each $H_{0i}^{\downarrow}$.

With this choice of test for $H_{0i}^{\downarrow},\ i = 1, \ldots, k$, since the null hypotheses partition the parameter space, Hsu and Berger (1999) could apply the Partition-Project corollary 1.1 to give the confidence bounds version of the inference:

Step 1

If $\hat{\mu}_k - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_0} > \Delta$,

then infer $\mu_k - \mu_0 > \Delta$ and go to Step 2;

else infer $\mu_k - \mu_0 > \hat{\mu}_k - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_0}$ and stop.

Step 2

If $\hat{\mu}_{k-1} - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_0} > \Delta$,

then infer $\mu_{k-1} - \mu_0 > \Delta$ and go to Step 3;

else infer $\mu_{k-1} - \mu_0 > \hat{\mu}_{k-1} - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_0}$ and stop.

$\vdots$

Step k

If $\quad \hat{\mu}_1 - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_1 + 1/n_0} > \Delta$

then infer $\mu_1 - \mu_0 > \Delta$ and go to Step $k+1$;

else infer $\mu_1 - \mu_0 > \hat{\mu}_1 - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_1 + 1/n_0}$ and stop.

Step $k+1$

Infer $\min_{i=1,\ldots,k} \mu_i - \mu_0 > \min_{i=1,\ldots,k}\{\hat{\mu}_i - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_0}\}$ and stop.

Note the Step $k+1$ confidence bound is from pivoting an Intersection-Union Tests (IUT) for (1.21).

Closed duplicate testing to stay on a decision path

Whether there are decision paths or not, closed testing would test all intersections of the null hypotheses in (1.15). To stay on a (single) decision path, what closed testing does (including the Graphical Approach) is to test all the intersection null hypotheses that make up each of $H_{0i}^{\downarrow}$ by one and the same pair-wise t test, rejecting if $\hat{\mu}_i - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_0} > \Delta$. (Chapter 5 of this Handbook is on the Graphical approach.) In this scheme, testing $H_{04}^{\downarrow}$ corresponds to testing all eight rows with $\circledcirc$ for Θ_4 in Table 1.3 by the same test which rejects when $\hat{\mu}_4 - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_4 + 1/n_0} > \Delta$. Testing $H_{03}^{\downarrow}$ corresponds to testing the four rows with $\circledcirc$ for Θ_3 but $\otimes$ for Θ_4 in Table 1.3 by the same test which rejects when $\hat{\mu}_3 - \hat{\mu}_0 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_3 + 1/n_0} > \Delta$, and so forth. Bauer et al. (1998) explains this (redundant) closed testing scheme as:

> "Now for strictly ordered null hypotheses, every level α-test can be formally considered as a level α-test for all intersections with null hypotheses at a lower hierarchical order."

1.4.2 Making decisions along a path for the Alzheimer study

We take $\Delta = 0$ and fit the entire data to the model. Unlike the multivariate case, in the univariate case, comparing 2-sided $|t|$ p-values to 10% corresponds exactly to comparing 1-sided t p-values to 5%. So, using the 2-sided $|t|$ p-values in Table 1.5 which are computed without multiplicity adjustment, we have

Step 1

Is $\hat{\mu}_4 - \hat{\mu}_0 - t_{.05,\nu}\hat{\sigma}\sqrt{1/n_4 + 1/n_0} > 0$?

Yes since the 2-sided $|t|$ p-value for Dose 4 $= 0.0016$

So infer $\mu_4 - \mu_0 > 0$ and go to Step 2;

Step 2

Is $\hat{\mu}_3 - \hat{\mu}_0 - t_{.05,\nu}\hat{\sigma}\sqrt{1/n_3 + 1/n_0} > 0$?

No since the 2-sided $|t|$ p-value for Dose 3 $= 0.0665$

So stop.

So, for the particular case of this Alzheimer study, at the 1-sided 5% level, making decisions along the Dose $4 \to 3 \to 2 \to 1$ path infers only Doses 4 to be better than the placebo, while the single-step and the step-down Dunnett's method infer Doses 4 and 2 to be better than the control.

The methods we have presented do not assume response has any particular form as a function of dose. Though, for the Minimum Effective Dose (MED) problem to be meaningful, there is the tacit assumption that if a dose is efficacious, then all higher doses are efficacious as well. In the presence of a $\Delta > 0$ defining a clinically meaningful difference, this assumption actually is not as strong as the one that efficacy is monotonically increasing in dose (see Dmitrienko et al. 2007). In addition, see Chapter 11 on dose-finding in this Handbook for the MCP-Mod approach which models response as a continuously valued function of dose.

1.4.3 Partitioning when there are multiple decision paths

Let μ_{ij} denote the mean response in dose group i for endpoint j, $i = 0, 1, ..., k$, $j = 1, ..., m$, where $i = 0$ denotes the placebo group, while $i = 1, \ldots, k$ are additional doses. And $j = 1$ denotes the primary endpoint, $j = 2$ is the secondary endpoint, $j = 3$ is the tertiary endpoint, and so forth. Define $\theta_{ij} = \mu_{ij} - \mu_{0j}$ to be the difference in mean efficacy measurement between dose group i and the placebo group for endpoint j.

Assuming a larger measurement indicates a better treatment, the statistical inference of interest is to test, for each dose endpoint combination,

$$H_{0ij} : \theta_{ij} \leq \delta_j, \quad i = 1, ..., k, \; j = 1, ..., m. \tag{1.23}$$

Efficacy claim of the new experimental drug is based on the primary endpoint alone, additional claims on secondary endpoints are of interest only if the primary endpoint has shown efficacy.

Liu and Hsu (2009) showed how to apply Decision Path Principle by using partition testing to situations in which there is an ordering among the endpoints (in terms of sequence of potential inferences), but there is no such ordering among the doses. The lower ordered secondary endpoints, in this situation, are required to be tested only if higher ordered ones are proven efficacious.

Decision paths are thus within, but not between, doses. Figure 1.3 illustrates such paths with k doses and m endpoints.

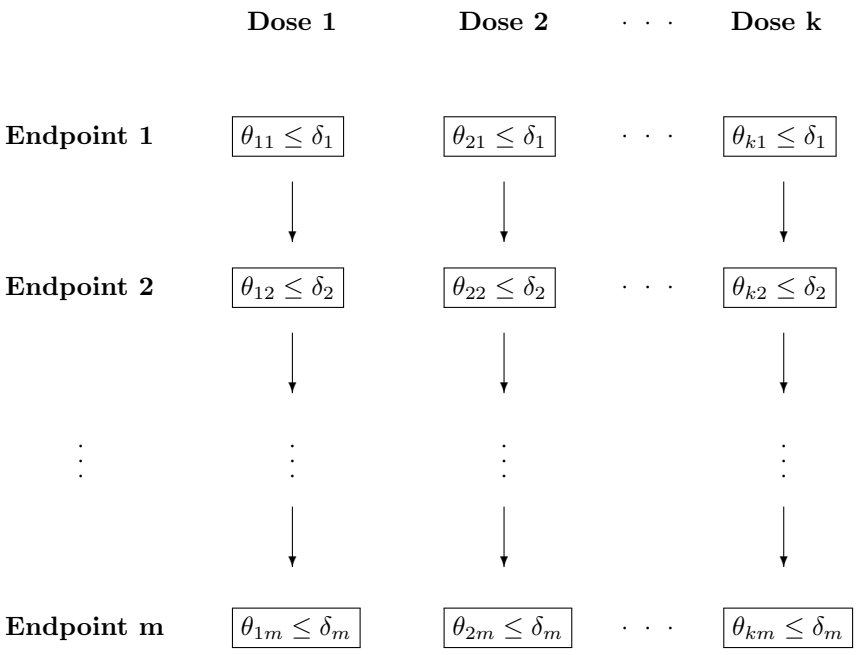


FIGURE 1.3: Decision paths for k doses m endpoints, with one path for each dose going from endpoint 1 to endpoint m.

If a single secondary endpoint is involved, the decision path in Figure 1.3 becomes Figure 1.4 shown below. For notation conveniences, instead of using the second subscript to index the endpoint, we use superscripts ^{P} and ^{S} to denote primary and secondary endpoints.

Given decision paths in Figure 1.4, the parameter space is partitioned in two stages:

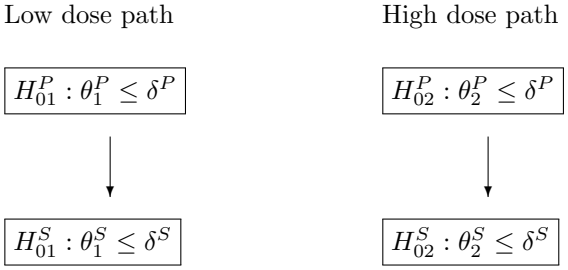


FIGURE 1.4: Decision paths for low and high doses.

Path partition: Partition within each path.

Disjointness partition: Further partition by taking intersections to make hypotheses between paths disjoint.

Path partition is within each dose i. Starting with the primary endpoint, we test $H_{0i}^P : \theta_i^P \leq \delta^P$. If it is rejected, then efficacy in the primary endpoints has been established.

Then follow the path to the secondary endpoint. But, instead of testing $H_{0i}^S : \theta_i^S \leq \delta^S$, we make it disjoint with H_{0i}^P and test $H_{0i}^{\star S} : \theta_i^S \leq \delta^S$ and $\theta_i^P > \delta^P$. If both H_{0i}^P and $H_{0i}^{\star S}$ are rejected, then we logically conclude efficacy in both the primary and the secondary endpoints.

Whereas the reason for path partitioning is inference in the secondary endpoint is irrelevant unless efficacy in the primary endpoint is established, disjointness partitioning is for proper multiplicity adjustment.

Multiplicity adjustment is needed (only) to the extent that two or more hypotheses can be true simultaneously. Ask the question, "is it possible that high dose primary lacks efficacy and simultaneously there is efficacy in high dose primary but not in high dose secondary?" The answer is "no", so there is no need to adjust for multiplicity in testing H_{0i}^P and $H_{0i}^{\star S}$. However, "is it possible that high dose primary and low dose secondary lack efficacy?" "Yes it is", so that particular multiplicity of two needs to be adjusted for.

To figure out which combination of the path-partitioned null hypotheses can be true simultaneously, between paths, connect an *edge* between subspaces of the parameter space that are not disjoint, as illustrated in Figure 1.5. Edges represent hypotheses that can be true simultaneously. (There is no edge between $\{\theta_1^S > \delta^S$ and $\theta_1^P > \delta^P\}$ and $\{\theta_2^S > \delta^S$ and $\theta_2^P > \delta^P\}$, since the intersection of these two hypotheses, the ideal situation of having efficacy in all doses and endpoints, need not be tested.) Take intersections of connected subspaces to form new hypotheses. The resulting set of hypotheses, as presented in Table 1.6, partitions the parameter space. Therefore, so long as each partition hypothesis is tested at level α, the FWER is controlled strongly at level α. Inferences on which of the $m \times k$ combinations of dose and endpoint

are efficacious are then obtained by collating the results from the $(m+1)^k - 1$ tests.

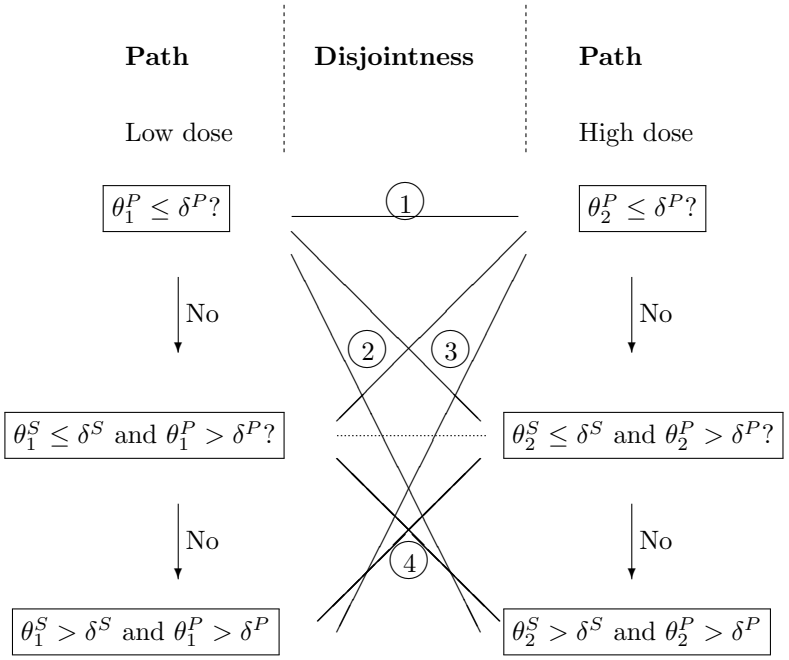


FIGURE 1.5: Graphical representation of two stages of partitioning in the setting of Figure 1.4.

Specifically, the inference $\theta_{ij} > \delta_j$ is made if all hypotheses implying that $\theta_{ij} \leq \delta_j$ could be true are rejected. This includes partitioning hypotheses which do not explicitly state an inequality for θ_{ij}. For example, the hypothesis which states $\theta_1^P \leq \delta^P$ and $\theta_2^P \leq \delta^P$ includes the possibility that $\theta_i^S \leq \delta^S$ (as well as the possibility $\theta_i^S > \delta^S$), and must be rejected before inference on any secondary endpoint is given.

1.4.3.1 Insights from the path-partitioning principle

The path-partitioning principle is most useful in giving insights into the structure of multiple testing when there are multiple doses and decision paths. To actually execute multiple testing when there are decision paths, the Graphical Approach described in Chapter 5 is perhaps more convenient. We thus focus on explaining the *insights*.

Multiplicity adjustment: Inclusion of secondary endpoints in the analysis may necessitate multiplicity adjustment in inference on the primary endpoint.

For example, in Table 1.6, one rejects the hypothesis H_{01}^P if

$$\{\theta_1^P \leq \delta^P \text{ and } \theta_2^P \leq \delta^P\}, \tag{1.24}$$

$$\{\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S \leq \delta^S\}, \tag{1.25}$$

$$\text{and } \{\theta_1^P \leq \delta^P \text{ and } \theta_2^P > \delta^P \text{ and } \theta_2^S > \delta^S\} \tag{1.26}$$

are all rejected. Thus, inference on low dose of the primary endpoint (θ_1^P) may need multiplicity adjustment, to account for the possibility that high dose of the secondary endpoint may lack efficacy ($\theta_2^S \leq \delta^S$ in (1.25)). The mere presence of high dose secondary endpoint necessitates multiplicity adjustment in testing for efficacy in low dose primary, at least initially.

Note however, this multiplicity adjustment is removed if data shows efficacy in the secondary endpoint at high dose (i.e., if the partition hypothesis (1.25) is rejected). In fact, partitioning makes transparent that, if efficacy at a dose has been established for all endpoints, then that dose needs no longer be included in multiplicity adjustment.

This last realization from Liu and Hsu (2009), which some have come to phrase as "there is no need to leave money on the table", explains a key difference from Figure 2 in Bretz et al. (2009), to Figure 1 in Bretz et al. (2011) and Figure 6.3(a) in Chapter 5 on Graphical Approach in this Handbook, that there is an arrow from the bottom node of each path to the top node of the other path in the latter two figures, as the Graphical Approach to this problem has evolved.

Appearance: Whether one readily sees it or not, inference on the primary endpoint may depend on observations on a secondary endpoint (at a different dose), because initially one must account for possibilities such as efficacy is lacking for the primary endpoint at high dose and for the secondary endpoint at low dose. However, this multiplicity can be removed if data indicates otherwise.

One can give the appearance that such dependence does not occur, by choosing not to remove the multiplicity adjustment, even if data indicates it can be. This is the approach taken in Dmitrienko et al. (2006), Xu et al. (2009), and Dmitrienko and Tamhane (2011), to ensure "inference made in primary endpoints not affected by the inference made in secondary endpoints". We feel there is no need for such loss of power, for the sake of appearance.

1.4.4 Controlling FWER may be too simplistic for primary-secondary endpoint problems

The original proof (in Hsu and Berger 1999) that no multiplicity adjustment is needed to control FWER in testing along a (single) path was in the setting of dose-response studies. In the setting of testing high dose first, with evidence of efficacy in high dose, one then tests low dose, there may be one or two Type I errors. If neither dose is effective, then inferring high dose is effective only commits one Type I error, while inferring both high dose and low dose as

effective commits two Type I errors. FWER, being the probability of making at least one Type I error, counts one or two Type I errors as the same. With the principal purpose of Hsu and Berger (1999) being to show that most of the contrast methods that pool information across doses popular then do not control incorrect decision rate, they viewed this oversimplification of FWER control as an acceptable first approximation, since those two Type I errors are of the same kind: "too low is too low".

However, the ordered endpoints setting is different. In testing Primary and Secondary endpoints in sequence, unconditional FWER refers to, over many studies each with a Primary and a Secondary endpoint, roughly how many percent of the studies have incorrect efficacy claim in either Primary or Secondary or both. However, since testing Primary is for approval, while Secondary testing is for additional labeling claim, we might consider conditional Type I error rate on testing for efficacy in the Secondary, conditional on inferring efficacy in the Primary. This conditional error rate is, over many drug submissions which get approval (and therefore have drug labels), roughly how many percent of the labels have incorrect additional claims (beyond indication), and might offer useful additional information toward sound decision-making.

1.5 Key Messages of This Chapter

- In multiple comparisons, error rate controls are useful if they translate to controlling the probability of making *incorrect decisions*.
- One can be confident that the *directional* error rate is controlled if the null hypotheses of a multiple test partition the entire parameter space, but not if the null hypotheses are mere equalities.
- If the null hypotheses partition the entire parameter space, then that multiple test can be pivoted to give a compatible confidence set.
- Using a compatible confidence set to execute a multiple test guarantees the directional error rate is controlled.
- Partitioning is also useful in formulating null hypotheses to channel multiple tests onto pre-specified decision paths.

Θ_1	Θ_2	Θ_3	Θ_4	H_{01}	H_{02}	H_{03}	H_{04}
◎	◎	◎	◎	✓	✓	✓	✓
◎	◎	◎	⊗	✓	✓	✓	✗
◎	◎	⊗	◎	✓	✓	✗	✓
◎	⊗	◎	◎	✓	✗	✓	✓
⊗	◎	◎	◎	✗	✓	✓	✓
◎	◎	⊗	⊗	✓	✓	✗	✗
◎	⊗	◎	⊗	✓	✗	✓	✗
◎	⊗	⊗	◎	✓	✗	✗	✓
⊗	◎	◎	⊗	✗	✓	✓	✗
⊗	◎	⊗	◎	✗	✓	✗	✓
⊗	⊗	◎	◎	✗	✗	✓	✓
◎	⊗	⊗	⊗	✓	✗	✗	✗
⊗	◎	⊗	⊗	✗	✓	✗	✗
⊗	⊗	◎	⊗	✗	✗	✓	✗
⊗	⊗	⊗	◎	✗	✗	✗	✓
⊗	⊗	⊗	⊗	✗	✗	✗	✗

TABLE 1.3: Partition testing of four null hypotheses

Dose	Adjusted for $\{1,2,3,4\}$	Adjusted for $\{1,2,3\}$	Adjusted for $\{1,3\}$	For $\{1\}$
4	0.0061	-	-	-
3	0.2076	0.1673	0.1217	-
2	0.0991	0.0783	-	-
1	0.6911	0.6021	0.4802	0.2949

TABLE 1.4: Adjusted 2-sided $|t|$ $p-$values facilitating execution of step-down Dunnett's method for the Alzheimer study, to be compared with 0.10 for 1-sided FWER $\approx 5\%$.

Dose	$\|t\|$ p-value for $\{4\}$	$\|t\|$ p-value for $\{3\}$	$\|t\|$ p-valued for $\{2\}$	$\|t\|$ p-value for $\{1\}$
4	0.0016	-	-	-
3	-	0.0665	-	-
2	-	-	0.0295	-
1	-	-	-	0.2949

TABLE 1.5: Unadjusted 2-sided $|t|$ $p-$values facilitating execution of decision-path method for the Alzheimer study, to be compared with .10 for 1-sided FWER $= 5\%$.

TABLE 1.6: Partition hypotheses following decision paths in Figure 1.4.

Index[8]	Partition hypothesis		Rejection rule		
1	$\boldsymbol{\theta_1^P \leq \delta^P}$ and $\boldsymbol{\theta_2^P \leq \delta^P}$		$t_1^P > c_1$ or $t_2^P > c_1$		
2	$\boldsymbol{\theta_1^P \leq \delta^P}$and $\theta_2^P > \delta^P$	and $\boldsymbol{\theta_2^S \leq \delta^S}$	$t_1^P > c_1$		or $t_2^S > c_2$
	$\boldsymbol{\theta_1^P \leq \delta^P}$and $\theta_2^P > \delta^P$	and $\theta_2^S > \delta^S$	$t_1^P > c_3$		
3	$\theta_1^P > \delta^P$ and $\boldsymbol{\theta_2^P \leq \delta^P}$and $\boldsymbol{\theta_1^S \leq \delta^S}$			$t_2^P > c_1$ or $t_1^S > c_2$	
	$\theta_1^P > \delta^P$ and $\boldsymbol{\theta_2^P \leq \delta^P}$and $\theta_1^S > \delta^S$			$t_2^P > c_3$	
	$\theta_1^P > \delta^P$ and $\theta_2^P > \delta^P$ and $\boldsymbol{\theta_1^S \leq \delta^S}$and $\boldsymbol{\theta_2^S \leq \delta^S}$				$t_1^S > c_1$ or $t_2^S > c_1$
4	$\theta_1^P > \delta^P$ and $\theta_2^P > \delta^P$ and $\boldsymbol{\theta_1^S \leq \delta^S}$and $\theta_2^S > \delta^S$				$t_1^S > c_3$
	$\theta_1^P > \delta^P$ and $\theta_2^P > \delta^P$ and $\theta_1^S > \delta^S$ and $\boldsymbol{\theta_2^S \leq \delta^S}$				$t_2^S > c_3$

Note: [8]the index column corresponds to the labels of edges in Figure 1.5.

Bibliography

Abraham, J. E., Maranian, M. J., Driver, K. E., Platte, R., Kalmyrzaev, B., Baynes, C., Luccarini, C., Shah, M., Ingle, S., Greenberg, D., Earl, H. M., Dunning, A. M., Pharoah, P. D., , and Caldas, C. (2010). CYP2D6 gene variants: association with breast cancer specific survival in a cohort of breast cancer patients from the United Kingdom treated with adjuvant tamoxifen. *Breast Cancer Research*, 12:R64.

Arvanitis, L. A., Miller, B. G., and the Seroquel Trial 13 Study Group (1997). Multiple fixed doses of "Seroquel" (Quetiapine) in patients with acute exacerbation of Schizophrenia: A comparison with Haloperidol and placebo. *Biological Psychiatry*, 42:233–246.

Bauer, P., Rohmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17:2133–2146.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 25:16–39.

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. John Wiley & Sons, New York.

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

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal*, 53:894–913.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Thomson Learning, Pacific Grove, CA, 2nd edition.

Ding, Y., Li, Y. G., Liu, Y., Ruberg, S. J., and Hsu, J. C. (2018). Confident inference for snp effects on treatment efficacy. *Ann. Appl. Statist.*, 12(3):1727–1748.

Ding, Y., Lin, H.-M., and Hsu, J. C. (2016). Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine*, 35:1580–1594.

Dmitrienko, A., Fritsch, K., Hsu, J., and Ruberg, S. (2007). *Pharmaceutical Statistics Using SAS: A Practical Guide*, chapter Design and Analysis of Dose-Ranging Clinical Studies, pages 273–311. SAS Institute, Inc.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*, 5:19–28.

Dmitrienko, A. and Tamhane, A. C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine*, 30:1473–1488.

Edwards, D. G. and Hsu, J. C. (1983). Multiple comparisons with the best treatment. *Journal of the American Statistical Association*, 78:965–971.

Fabian, V. (1962). On multiple decision methods for ranking population means. *Annals of Mathematical Statistics*, 33:248–254.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.

Finner, H. (1994). Two-sided tests and one-sided confidence bounds. *Annals of Statistics*, 22:1502–1516.

Finner, H. (1999). Stepwise multiple test procedures and control of directional errors. *The Annals of Statistics*, 27:274–289.

Finner, H. and Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, 30:1194 –1213.

Finner, H. and Strassburger, K. (2007). Step-up related simultaneous confidence intervals for MCC and MCB. *Biometrical Journal*, 49(1):40–51.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*. Wiley, New York.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo Series 150, Institute of Statistics, University of North Carolina, Chapel Hill, NC.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, 7:225–245.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures – Theory and Methodology of Selecting and Ranking Populations*. John Wiley, New York.

Han, Y., Tang, S.-Y., Lin, H.-M., and Hsu, J. C. (2020). Exact simultaneous confidence intervals for logical selection of a biomarker cut-point. Unpublished.

Hayter, A. J. and Hsu, J. C. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association*, 89:128–136.

Holmes, M. V., Perel, P., Shah, T., Hingorani, A. D., and Casas, J. P. (2011). Cyp2c19 genotype, clopidogrelmetabolism, platelet function, and cardiovascular events: A systematic review and meta-analysis. *Journal of the American Medical Association*, 306:2704–2714.

Hoskins, J. M., Carey, L. A., and McLeod, H. L. (2009). CYP2D6 and tamoxifen: DNA matters in breast cancer. *Nature Reviews: Cancer*, 9:576–586.

Hsu, J. C. (1981). Simultaneous confidence intervals for all distances from the 'best'. *Annals of Statistics*, 9:1026–1034.

Hsu, J. C. (1982). Simultaneous inference with respect to the best treatment in block designs. *Journal of the American Statistical Association*, 77:461–467.

Hsu, J. C. (1984). Constrained two-sided simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics*, 12:1136–1144.

Hsu, J. C. and Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association*, 94:468–482.

Huang, Y. and Hsu, J. C. (2007). Hochberg's step-up method: Cutting corners off Holm's step-down method. *Biometrika*, 22:2244–2248.

Kil, S., Kaizar, E., Tang, S.-Y., and Hsu, J. C. (2020). *Principles and Practice of Clinical Trials*, chapter Confident Statistical Inference with Multiple Outcomes, Subgroups, and Other Issues of Multiplicity. Springer International Publishing, Cham.

Lawrence, J. (2019). Familywise and per-family error rates of multiple comparison procedures. *Statistics in Medicine*, 38(19):3586–3598.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. John Wiley, New York, second edition.

Lin, H.-M., Xu, H., Ding, Y., and Hsu, J. C. (2019). Correct and logical inference on efficacy in subgroups and their mixture for binary outcomes. *Biometrical Journal*, 61:8–26.

Liu, Y. and Hsu, J. C. (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. *Journal of the American Statistical Association*, 104:1661–1670.

Mega, J. L., Close, S. L., Wiviott, S. D., Shen, L., Walker, J. R., Simon, T., Antman, E. M., Braunwald, E., and Sabatine, M. S. (2010). Genetic variants in ABCB1 and CYP2C19 and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON-TIMI 38 trial: a pharmacogenetic analysis. *The Lancet*, 376:1312–1319.

Mega, J. L., Hochholzer, W., III, A. L. F., Kluk, M. J., Angiolillo, D. J., Kereiakes, D. J., Isserman, S., Rogers, W. J., Ruff, C. T., Contant, C., Pencina, M. J., Scirica, B. M., Longtine, J. A., Michelson, A. D., and Sabatine, M. S. (2011). Dosing clopidogrel based on cyp2c19 genotype and the effect on platelet reactivity in patients with stable cardiovascular disease. *Journal of the American Medical Association*, 306:2221–2228.

Nebert, D. and Russell, D. (2002). Clinical importance of the cytochromes P450. *The Lancet*, 360:1155–1162.

Paré, G., Mehta, S. R., Yusuf, S., Anand, S. S., Connolly, S. J., Hirsh, J., Simonsen, K., Bhatt, D. L., Fox, K. A., and Eikelboom, J. W. (2010). Effects of CYP2C19 genotype on outcomes of clopidogrel treatment. *New England Journal of Medicine*, 363:1704–1714.

Schnell, P., Tang, Q., Muller, P., and Carlin, B. P. (2017). Subgroup inference for multiple treatments and multiple endpoints in an alzheimers disease treatment trial. *Ann. Appl. Stat.*, 11:949–966.

Schroth, W. (2009). Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *Journal of the American Medical Association*, 302:1429–1436.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Annals of Statistics*, 8:1342–1348.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81:826–831.

Stefansson, G., Kim, W., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. In Gupta, S. S. and Berger, J. O., editors, *Statistical Decision Theory and Related Topics IV*, volume 2, pages 89–104. Springer-Verlag, New York.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927.

Takeuchi, K. (1973). *Studies in Some Aspects of Theoretical Foundations of Statistical Data Analysis (in Japanese)*. Toyo Keizai Shinposha, Tokyo.

Takeuchi, K. (2010). Basic ideas and concepts for multiple comparison procedures. *Biometrical Journal*, 52:722–734.

Tukey, J. W. (1953). The Problem of Multiple Comparisons. Dittoed manuscript of 396 pages, Department of Statistics, Princeton University.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6:100–116.

Tukey, J. W. (1992). Where should multiple comparisons go next? In Hoppe, F. M., editor, *Multiple Comparisons, Selection, and Applications in Biometry: A Festschrift in Honor of Charles W. Dunnett*, chapter 12, pages 187–208. Marcel Dekker, New York.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92(437):299–306.

Westfall, P. H., Bretz, F., and Tobias, R. D. (2013). Directional error rates of closed testing procedures. *Statistics in Biopharmaceutical Research*, 5:345–355.

Westfall, P. H. and Tobias, R. D. (2007). Multiple testing of general contrasts: Truncated closure and the extended shafferroyen method. *Journal of the American Statistical Association*, 102:487–494.

Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. SAS Publishing, 2nd edition.

Xu, H., Nuamah, I., Liu, J., Lim, P., and Sampson, A. (2009). A Dunnett-Bonferroni-based parallel gatekeeping procedure for doseresponse clinical trials with multiple endpoints. *Pharmaceutical Statistics*, 8(4):301–316.

Part II

Applications in Medicine

Index