Chapter 2 Multiple Testing Methodology

Wenge Guo

March 27, 2012

Motivating example (I)

- Clinical trials in patients with acute lung injury (see Zeiher et al., 2004)
  - Experimental drug is compared to placebo
  - Two primary endpoints:
    - number of days patients are off mechanical ventilation (vent-free days)
    - 28-day all-cause mortality rate
  - Trial is declared successful if the drug is superior to placebo with respect to either primary endpoint
  - Two important secondary endpoints:
    - number of days patients are out of ICU (ICU-free days)
    - overall quality of life at the end of the study
  - Can the secondary findings be included in the product label?

Motivating example (II)

Consider a dose-finding study with m doses tested versus placebo. Let μ_0 be the mean improvement in the placebo arm and μ_i be the mean improvement in the ith dose group, $i = 1, 2, \ldots, m$. The testing problem is formulated in terms of the difference in the mean responses. The null hypothesis is that the treatment effect is not greater than δ and the alternative is that the treatment effect is greater than δ. That is, for $i = 1, 2, \ldots, m$,

$$H_i : \mu_i - \mu_0 \leq \delta \quad \text{versus} \quad K_i : \mu_i - \mu_0 > \delta.$$

Sources of multiplicity

- Multiple endpoints
- Multiple doses
- Subgroup analysis
- Comparisons of multiple treatments with placebo
- Repeated tests of the same hypothesis (interim analysis)

The multiplicity problem

Multiple testing: "p-value ≤ 0.05" rule

- Suppose you carry out 10 hypothesis tests, each at the 5% level (assume all null hypotheses are true and all tests are independent).
  - The probability of declaring a particular test significant under its null hypothesis is 0.05;
  - But the probability of declaring at least 1 of the 10 tests significant is $1 - 0.95^{10} = 0.401$.
  - If you perform 20 hypothesis tests, this probability increases to 0.642; if you perform 100 hypothesis tests, this probability increases to $0.994\ldots$
- An example of microarray experiments
- How to overcome the multiplicity problem? The principle of rejecting a null hypothesis when its p-value ≤ 0.05 is not acceptable.

The multiplicity problem (II)

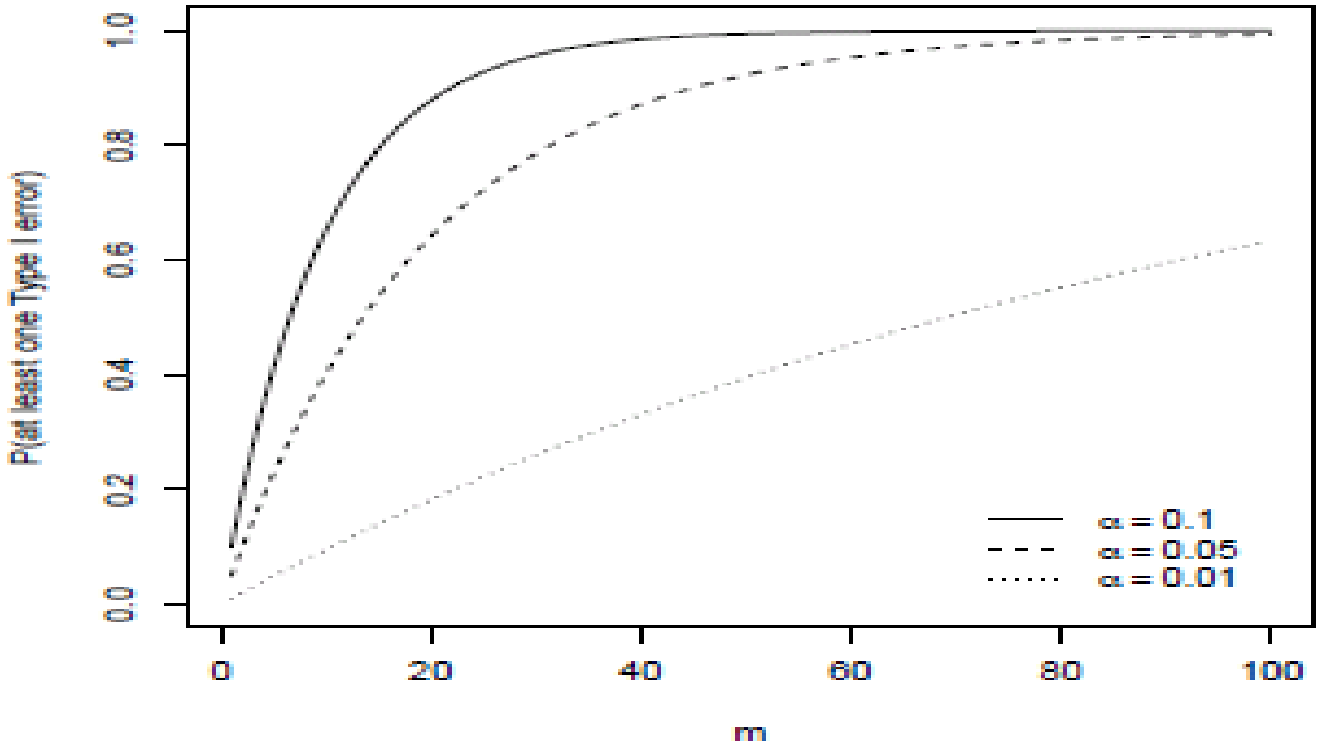


Figure 1.1 Probability of committing at least one Type I error for different significance levels α and number of hypotheses m.

One property of the p-value

- Suppose $X \sim N(\mu, 1)$. For testing $H_0 : \mu \leq 0$ vs. $H_a : \mu > 0$.
- The p-value $P = 1 - \Phi(X)$.
- The null distribution of the p-value is uniform or stochastically larger than the uniform:

$$\begin{aligned} &Pr_{H_0}(P \leq u) = Pr_{H_0}(1 - \Phi(X) \leq u) \\ &= Pr_{H_0}(X \geq \Phi^{-1}(1-u)) \leq 1 - \Phi(\Phi^{-1}(1-u)) = u. \end{aligned}$$

Family of hypotheses

- The term "family" refers to the collection of hypotheses $H_1, \ldots, H_m$ that is being considered for joint testing.
  - Hochberg and Tamhane (1987) define a family to be "any collection of inferences for which it is meaningful to take into account some combined measure of errors".
  - Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999) consider a family to be one in which the questions of interest "form a natural and coherent unit and are considered simultaneously in the decision-making process."
- What tests are to be treated jointly as a family depends on the situation.
- For a family of hypotheses, it is meaningful to take into account some combined measure of error.

Error rate definition

The following table summarizes the results of simultaneously testing m null hypotheses $H_1, \ldots, H_m$ for which m_0 of them are true and m_1 of them are false.

	claimed non-significant	claimed significant	Total
Null	N_{00}	N_{10}	m_0
Non-Null	N_{01}	N_{11}	m_1
Total	U	R	m

Error rate definition (II)

(a) Comparisonwise error rate: the proportion of incorrectly rejected hypotheses among all tested hypotheses.

$$\text{CWER} = E\left\{\frac{N_{10}}{m}\right\} = \frac{E\{N_{10}\}}{m}.$$

(b) Familywise error rate:

$$\text{FWER} = Pr(N_{10} > 0).$$

(c) Generalized familywise error rate

$$k\text{-FWER} = Pr(N_{10} \geq k).$$

Error rate definition (III)

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

(d) False discovery rate:

$$\text{FDR} = E\left(\text{FDP}\right) = E\left(\frac{N_{10}}{R}\right).$$

(e) Positive False discovery rate:

$$\text{pFDR} = E\left(\text{FDP}|R>0\right) = E\left(\frac{N_{10}}{R}|R>0\right).$$

(f) Proportion of False Positives: $P\left(\text{FDP} > \gamma\right) \leq \alpha$.

In clinical trials, we control the FWER.

Relationship between FWER and FDR

- CWER $\leq$ FDR $\leq$ FWER, with equality when all null hypotheses are true.
- FDR $\leq$ pFDR and k-FWER $\leq$ FWER
- FDP $\leq$ FWER
- Good scientific practice requires the specification of the Type I error rate control to be done prior to the data analysis.

Definition of power

- The minimal power is

$$\text{Power1} = P(N_{11} > 0),$$

  which is the probability of rejecting at least one false null hypothesis.
- The complete power is

$$\text{Power2} = P(\text{reject all false null hypotheses}).$$

- The average power is

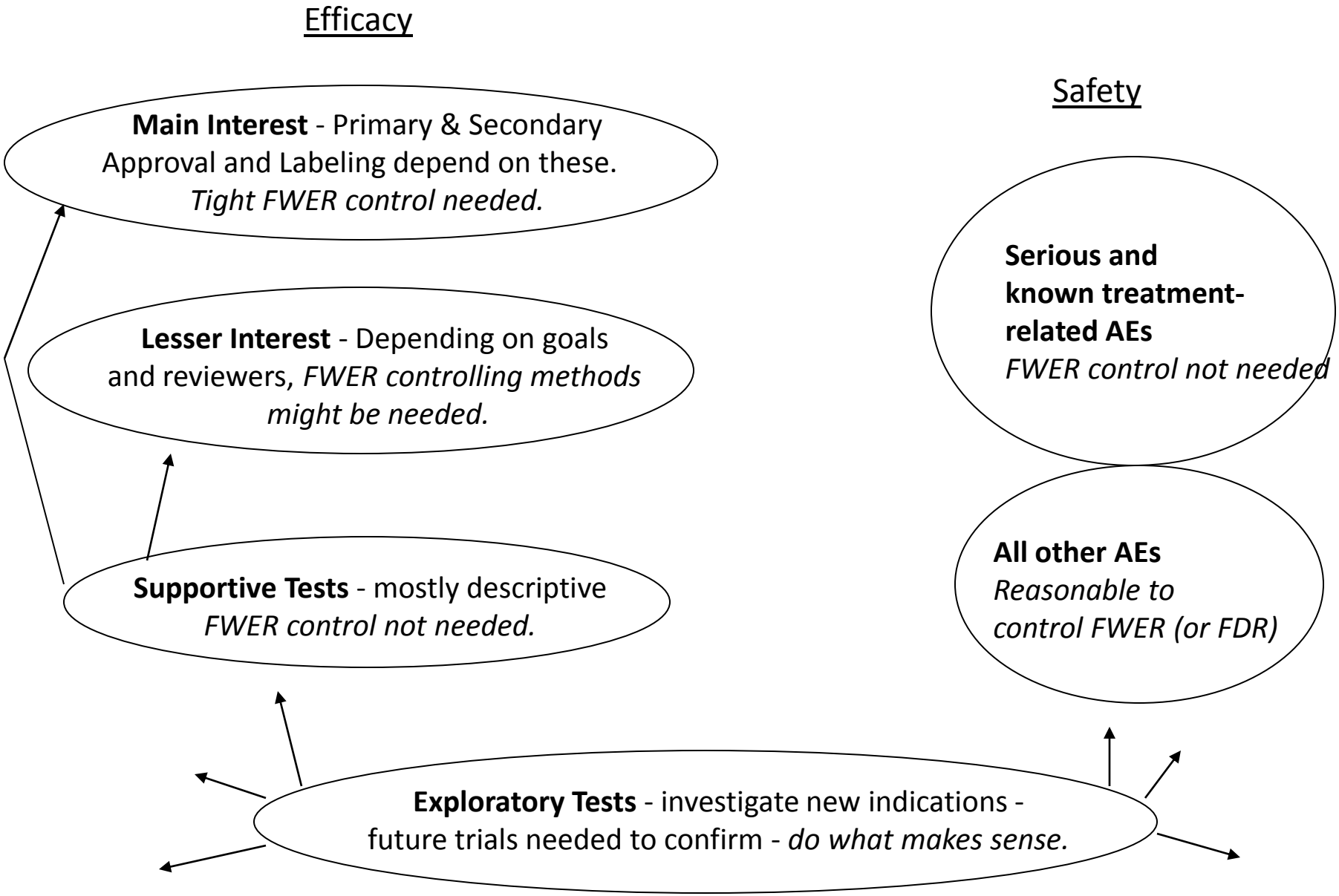
$$\text{Power3} = E\left\{\frac{N_{11}}{m_1}\right\},$$

  which is the average proportion of rejected false hypotheses among all false null hypotheses.

Weak control and strong control

- Strong control: control type I error rate under any combination of true and false null hypotheses.
- Weak control: control type I error rate only when all null hypotheses are true.
- Generally, we will focus on strong control of type I error rate, since we do not know which combination of true and false hypotheses the setting is.
- Strong control of the FWER for the primary objects is mandated by regulators in all confirmatory clinical trials.

“Families” in clinical trials



Union-intersection testing

Consider testing

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

In the context of union-intersection testing,

- one rejects the global hypothesis of no effect if there is evidence of a possible effect with respect to at least one individual objective.
- carrying out the individual test at an unadjusted α level leads to an inflated probability of rejecting H_I and can compromise the validity of statistical inference — a multiplicity adjustment is needed!

Several test methods

- Bonferroni test: reject $\bigcap_{i=1}^{m} H_i$ if

$$\min(P_1, P_2, \ldots, P_m) \leq \alpha/m.$$

- Fisher combination test for independent p-values: reject $\bigcap_{i=1}^{m} H_i$ if

$$-2\sum_{i=1}^{m} ln(P_i) > \chi^2(1-\alpha, 2m).$$

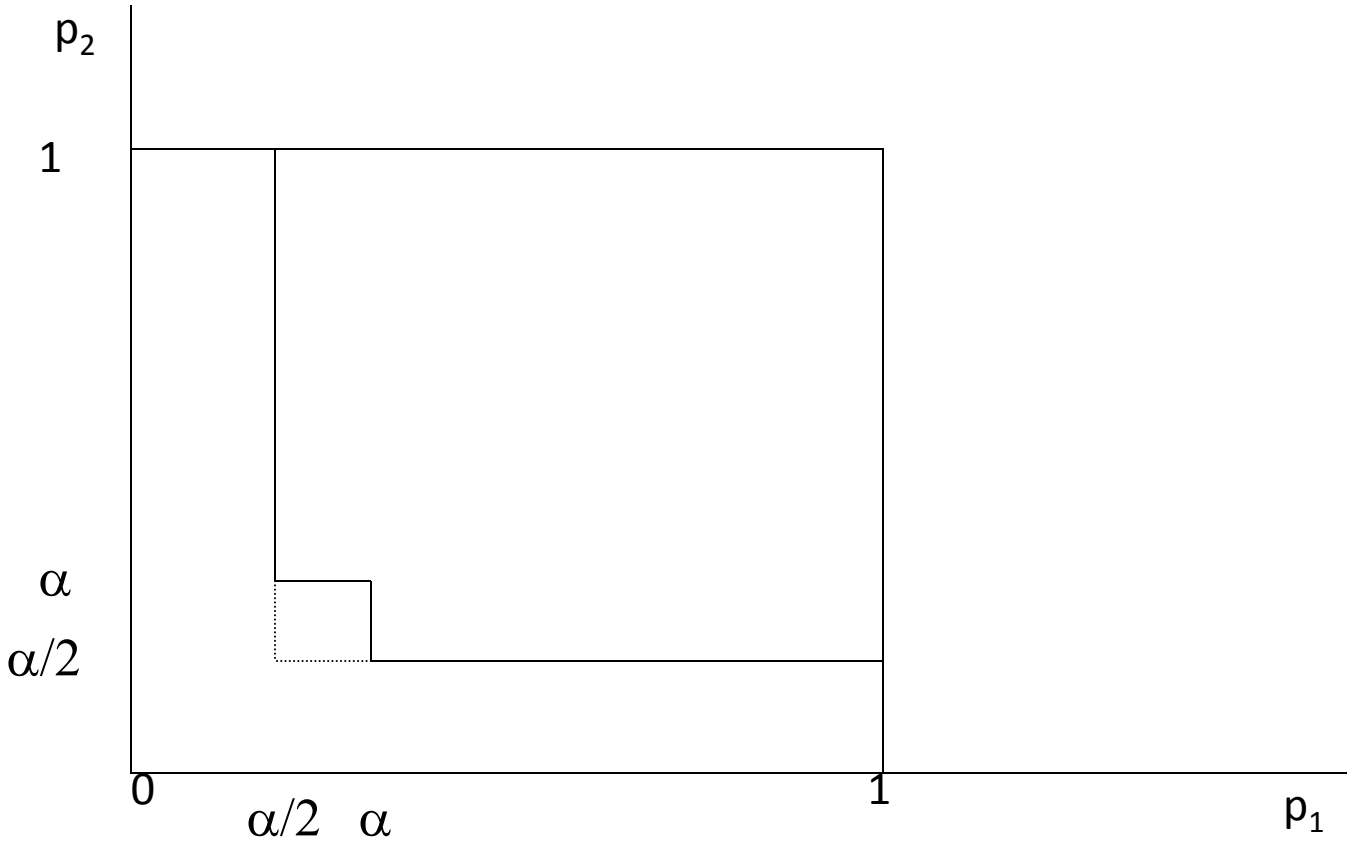
Simon's test

- Use all p-values $P_1, P_2, \ldots, P_m$ not just the MinP
- Simes's test rejects $\bigcap_{i=1}^{m} H_i$ if

$$P_{(j)} \leq j\alpha/m \text{ for at least one j .}$$

- Type I error at most α under independence or positive dependence of p-values.

Rejection Regions



P(Simes Reject) = $1 - (1-\alpha/2)^2 + (\alpha/2)^2 = \alpha$

P(Bonferroni Reject) = $1 - (1-\alpha/2)^2 = \alpha - (\alpha/2)^2$

Intersection-union testing

Consider testing

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

- Intersection-union testing naturally arises in studied when a significant outcome with respect to two or more objectives is required in order to declare the study successful.
- No multiplicity adjustment is necessary to control the size of a test but the individual hypotheses cannot be tested at levels higher than the nominal significance level either.

Control of Type I Error for IU tests

Consider testing the union of $H_{1:}\ \delta_1=0$ and $H_{2:}\ \delta_2=0$. Suppose $\delta_1=0$ ***or*** $\delta_2=0$. Then

P(Type I error)

= P(Reject H_0) (1)

= P($p_1 \leq .05$ *and* $p_2 \leq .05$) (2)

< min{P($p_1 \leq .05$), P($p_2 \leq .05$)} (3)

=.05. (4)

Note: The inequality at (3) becomes an approximate equality when p_2 is extremely noncentral.

Two examples

- New therapies for the treatment of Alzheimer's disease are required to demonstrate their effect on both cognition and global clinical scores.
- Bioequivalence: The "TOST test:
  - Test 1. $H_{01} : \delta \leq -\delta_0$ vs. $H_{A1} : \delta > -\delta_0$.
  - Test 2. $H_{01} : \delta \geq \delta_0$ vs. $H_{A1} : \delta < \delta_0$.
  - Can test both at $\alpha = 0.05$, but must reject both.

Closure principle

Closed principle is a powerful tool for constructing multiple testing methods.

Closure principle: A hypothesis is rejected in the context of multiple testing if and only if all intersection hypotheses containing this hypothesis are rejected by the local tests in the context of single test.

An example: Consider simultaneously testing

$$H_1 : \mu_1 \leq \mu_0 \quad \text{and} \quad H_2 : \mu_2 \leq \mu_0.$$

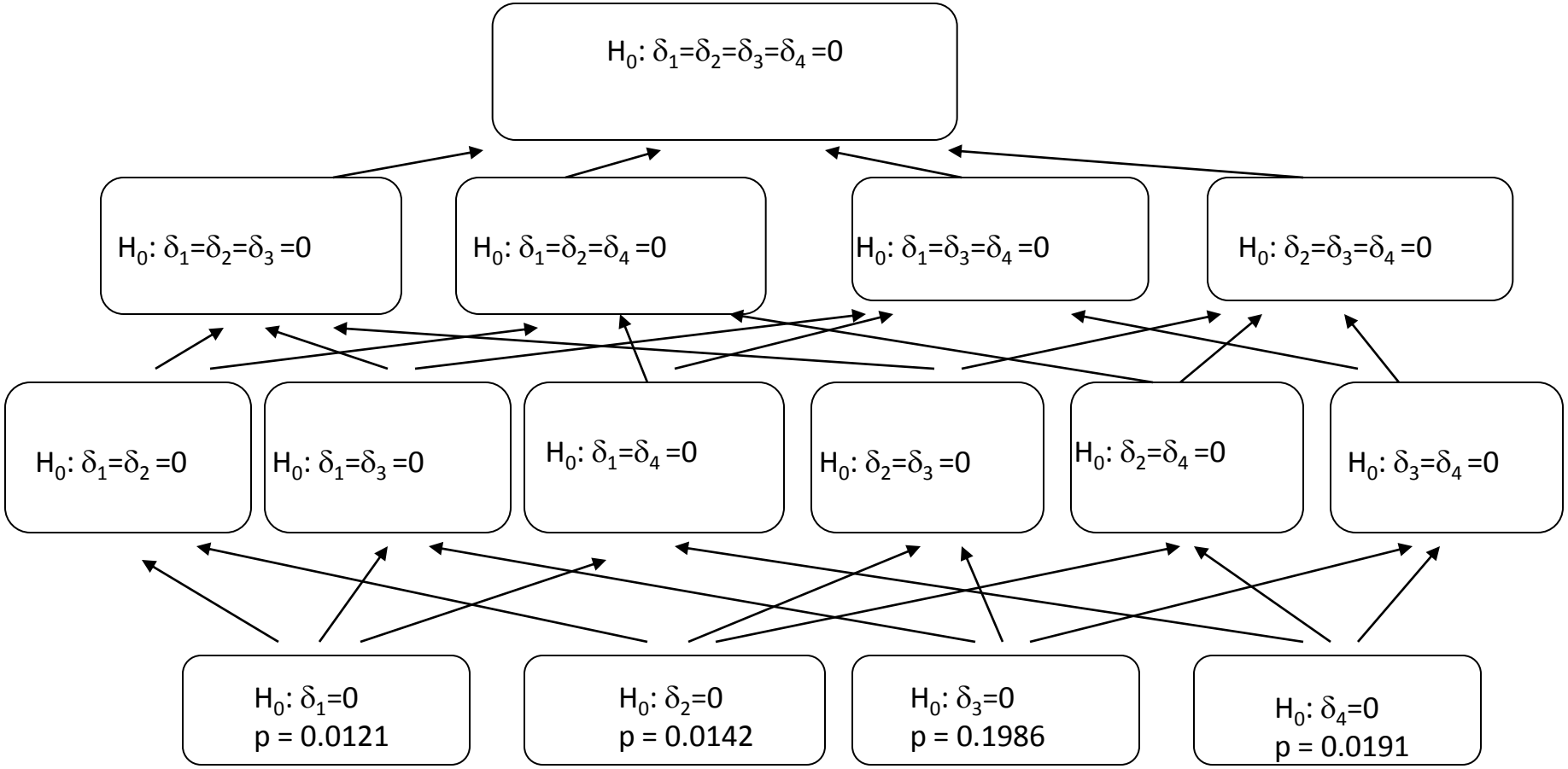
Form the the closure of the family of hypotheses: all possible intersections of these hypotheses.

Use the Bonferroni test to construct a local test: reject H_1 if $P_1 \leq \alpha$, reject H_2 if $P_2 \leq \alpha$, and reject $H_1 \bigcap H_2$ if $P_1 \leq \alpha/2$ or $P_2 \leq \alpha/2$.

Closed Testing Method(s)

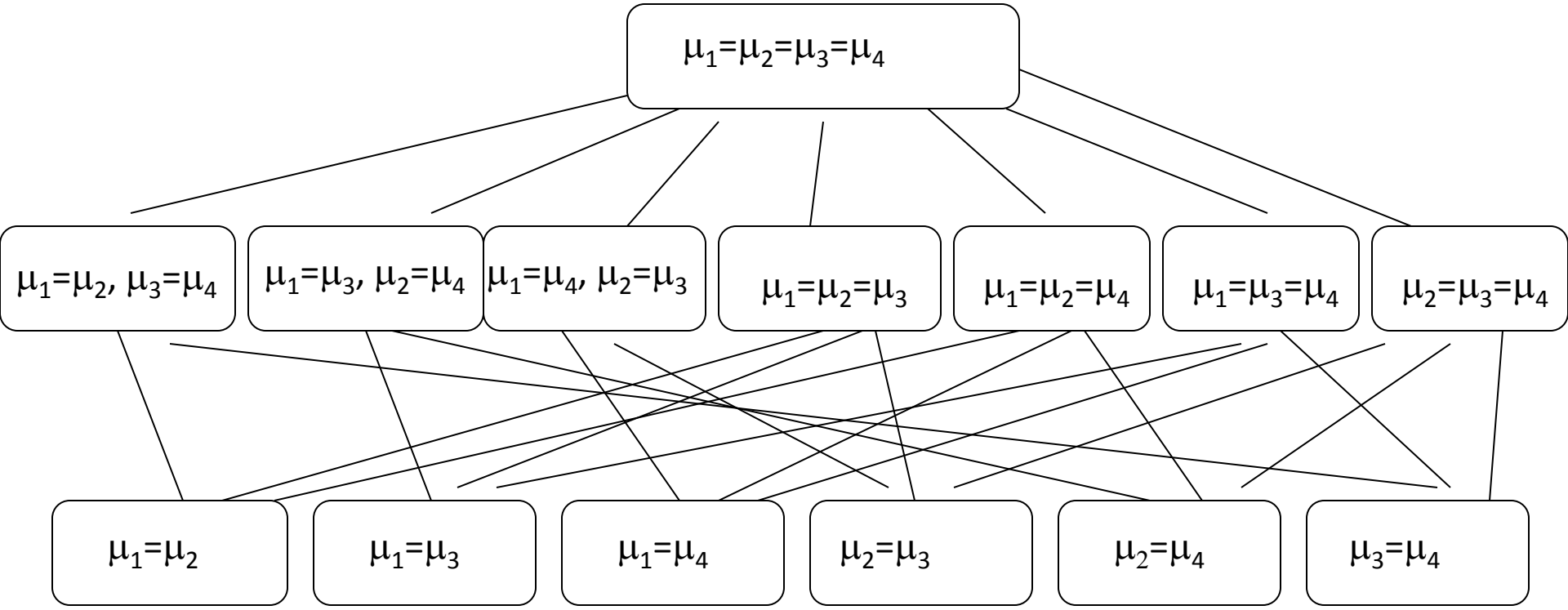
- Form the closure of the family by including all intersection hypotheses.
- Test every member of the closed family by a (suitable) α-level test. (Here, α refers to comparison-wise error rate).
- A hypothesis can be rejected provided that
  - its corresponding test is significant at level α, and
  - every other hypothesis in the family that implies it is rejected by its α–level test.

Closed Testing – Multiple Endpoints



Where δ_j = mean difference, treatment -control, endpoint j.

Closed Testing – Multiple Comparisons



Note: Logical implications imply that there are only 14 nodes, not $2^6 - 1 = 63$ nodes.

Control of FWER with closed tests

Suppose $H_{i_1}, H_{i_2}, \ldots, H_{i_{m_0}}$ all are true (unknown to you which ones).

$$\begin{aligned} &\{\text{reject at least one of } H_{i_1}, \ldots, H_{i_{m_0}} \text{ using CTP}\} \\ &\quad \subset \{\text{reject } H_{i_1} \cap \ldots \cap H_{i_{m_0}}\}. \end{aligned}$$

Thus

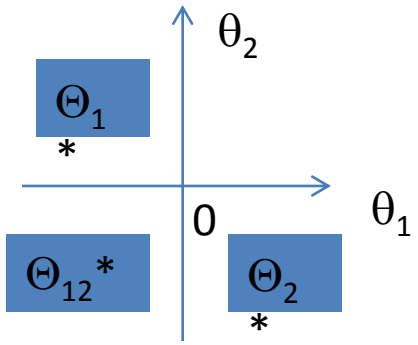
$$\begin{aligned} & P\{\text{reject at least one of } H_{i_1}, \ldots, H_{i_{m_0}} | H_{i_1}, \ldots, H_{i_{m_0}} \text{ all are true}\} \\ \leq \quad & P\{\text{reject } H_{i_1} \cap \ldots \cap H_{i_{m_0}} | H_{i_1}, \ldots, H_{i_{m_0}} \text{ all are true}\} \leq \alpha. \end{aligned}$$

Examples of Closed Testing Methods

When the Composite Test is…	**Then the Closed Method is …**
• Bonferroni MinP	• Holm's Method
• Resampling-Based MinP	• Westfall-Young method
• Simes	
• Simple or weighted test	• Hommel's method
	• Fixed sequence test (a-priori ordered)
• …	• …

Partitioning Principle

- Powerful tool to construct multiple testing method
- For example, $H_{01} = \{\theta_1 \le 0\}$, $H_{02} = \{\theta_2 \le 0\}$
- The parameter space of $\{\theta_1, \theta_2\}$ can be partitioned into *disjoint* subspaces
  - $\Theta_{12}^* = \{\theta_1 \le 0 \text{ and } \theta_2 \le 0\}$
  - $\Theta_1^* = \{\theta_1 \le 0 \text{ and } \theta_2 > 0\}$
  - $\Theta_2^* = \{\theta_1 > 0 \text{ and } \theta_2 \le 0\}$
  - $\{\theta_1 > 0 \text{ and } \theta_2 > 0\}$



- Under $H_{01} \cup H_{02}$, true value of $\{\theta_1, \theta_2\}$ can only exist in one of the 3 partitions: $\{\Theta_{12}^*, \Theta_1^*, \Theta_2^*\}$. Therefore, we only need to test each of the three partitions at level α, while controlling the overall type I error at level α.

Adjusted p-values

The adjusted p-value for a hypothesis is the smallest FWER or FDR level at which the hypothesis is rejected.

An example: A closed testing procedure rejects a hypothesis, H_i, if all intersection hypotheses containing H_i are rejected. If $P_I, I \subset \{1, \ldots, m\}$, denotes the p-value for testing H_I, the adjusted p-value for H_i is $\widetilde{P}_i = \max_{I:i\in I} P_I$. The H_i is rejected if the adjusted p-value does not exceed the pre-specified α level, i.e., $\widetilde{P}_i \leq \alpha$.

Several types of stepwise procedures

Let $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m$ be a sequence of increasing critical constants, and $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ be ordered p-values. Let $H_{(1)}, \ldots, H_{(m)}$ be the corresponding null hypotheses.

- Stepdown procedure: A stepdown procedure begins with the most significant hypothesis $H_{(1)}$, gradually step down to the least significant hypothesis $H_{(m)}$, and continues rejecting hypotheses as long as their corresponding p-values $P_{(i)}$ are less than the corresponding critical values α_i.
- Stepup procedure: A stepup procedure begins with the least significant hypothesis $H_{(m)}$, gradually step up to the most significant hypothesis $H_{(1)}$, and continues accepting hypotheses as long as their corresponding p-values $P_{(i)}$ are greater than the corresponding critical values α_i.
- Single-step procedure: A stepwise procedure with same critical constant, $\alpha_1 = \alpha_2 = \ldots = \alpha_m$.

Some typical stepwise procedures

- FWER controlling procedures:
  - Bonferroni: A single-step procedure with $\alpha_i = \alpha/m$
  - Sidak: A single-step procedure with $\alpha_i = 1 - (1-\alpha)^{1/m}$
  - Holm: A step-down procedure with $\alpha_i = \alpha/(m-i+1)$
  - Hochberg: A step-up procedure with $\alpha_i = \alpha/(m-i+1)$
  - minP method: A resampling-based single-step procedure with $\alpha_i = c_\alpha$, where c_α be the α quantile of the distribution of the minimum p-value.
- FDR controlling procedure:
  - Benjamini-Hochberg: A step-up procedure with $\alpha_i = i\alpha/m$.

Distributional assumptions

- Procedures don't make any assumption about the joint distribution of the test statistics. It only rely on univariate p-values and thus have a simple form. They are often called p-value based procedures or nonparametric procedures.
- Procedures that make specific distributional assumptions that the test statistics follow a multivariate normal or t-distribution. They are often called parametric procedures.
- Procedures that do not make specific assumptions and attempt to approximate the true joint distribution of the test statistics by using resampling-based methods. They are often called bootstrap or permutation methods or resampling-based procedures.