

Multiple Testing Procedures with Applications to Genomics

Part I. Motivation and Overview

Sandrine Dudoit

PB HLTH 240D – Spring 2007

©Copyright 2007, all rights reserved

Outline

These lecture notes are based on the forthcoming book by Dudoit and van der Laan (2007).

Related articles and tech reports may be downloaded from Sandrine Dudoit's website

`www.stat.berkeley.edu/~sandrine`

and Mark van der Laan's website

`www.stat.berkeley.edu/~laan.`

Outline: Part I. Motivation and Overview

- Multiple Hypothesis Testing Problems in Genomics.
- Multiple Hypothesis Testing Framework.
- Data Generating Distribution.
- Parameters.
- Null and Alternative Hypotheses.
- Test Statistics.
- Multiple Testing Procedures.
- Rejection Regions.

Outline: Part I. Motivation and Overview

- Errors in Multiple Hypothesis Testing: Type I, Type II, and Type III errors.
- Type I Error Rates.
- Power.
- Unadjusted and Adjusted p -Values.
- Examples of Multiple Testing Procedures.

Outline: Part II. Methodology

- Test Statistics Null Distribution.
- Single-Step Multiple Testing Procedures for Controlling General Type I Error Rates, $\Theta(F_{V_n})$.
- Step-Down Multiple Testing Procedures for Controlling the Family-Wise Error Rate.
- Augmentation Multiple Testing Procedures for Controlling Generalized Tail Probability Error Rates.
- Resampling-Based Empirical Bayes Multiple Testing Procedures for Controlling Generalized Tail Probability Error Rates.
- Appendix: Miscellaneous Mathematical and Statistical Results.

Outline: Part III. Applications to Genomics and Software Implementation

- Identification of Differentially Expressed and Co-Expressed Genes in High-Throughput Gene Expression Experiments.
- Multiple Tests of Association with Biological Annotation Metadata.
- HIV-1 Sequence Variation and Viral Replication Capacity.
- Genetic Mapping of Complex Human Traits Using Single Nucleotide Polymorphisms: The ObeLinks Project.
- Software Implementation.

Multiple Hypothesis Testing Problems in Genomics

- High-throughput microarray gene expression experiments.
 - Identification of differentially expressed genes by testing for associations between gene expression measures and possibly censored biological and clinical covariates and outcomes.
 - Identification of co-expressed genes by testing for associations in the expression measures of sets of genes across biological samples.
- Biological annotation metadata analysis. Tests of association between gene expression measures and biological annotation metadata.
E.g. Gene Ontology (GO, www.geneontology.org) annotation.

Multiple Hypothesis Testing Problems in Genomics

- **ChIP-Chip experiments.** Identification of **transcription factor (TF) binding sites** in ChIP-Chip experiments, where chromatin immunoprecipitation (ChIP) of transcription factor-bound DNA is followed by microarray (Chip) hybridization of the IP-enriched DNA.
Tests of association between **probe intensity measures** and **target sample** (TF ChIP vs. control sample).
- **Protein sequence analysis.** **Tests of association** between **phenotypes** and **codon/amino acid mutations**.
E.g. Association between viral replication capacity and HIV-1 sequence variation.

Multiple Hypothesis Testing Problems in Genomics

- Genetic mapping of complex traits. Tests of association between phenotypes and genotypes.
E.g. Single nucleotide polymorphisms (SNP), SNP haplotypes, microsatellite marker genotypes, identity by descent (IBD) status.
- Mass-spectroscopy. Tests of association between phenotypes and protein mass-spectroscopy measures.
E.g. Association between leukemia class (ALL vs. AML) and mass-to-charge ratios.

Multiple Hypothesis Testing Problems in Genomics

- Inference for **high-dimensional multivariate distributions**, with complex and unknown dependence structures among variables.
- **Broad range of parameters** of interest.
E.g. Regression coefficients in non-linear models relating patient survival data to genome-wide transcript (i.e., mRNA) levels, DNA copy numbers, or SNP genotypes;
measures of association between GO annotation and parameters of the distribution of microarray expression measures;
pairwise correlation coefficients between transcript levels.
- **Many null hypotheses**, in the thousands or even millions.
- **Complex and unknown dependence structures among test statistics**.
E.g. Directed acyclic graph (DAG) structure of GO terms;
Galois lattice for multilocus composite SNP genotypes.

Multiple Hypothesis Testing Framework

- **Multiplicity problem.** When multiple hypotheses (here, thousands!) are tested simultaneously, there is an **increased chance of committing at least one false positive**, i.e., **Type I error**.

Small (unadjusted) p -values, that would lead to the rejection of a single hypothesis (e.g., 0.001), no longer correspond to significant findings.

- **E.g.** The chance that at least one p -value is less than α for M independent test statistics is $1 - (1 - \alpha)^M$ and converges to one as M increases.

For $M = 1,000$ and $\alpha = 0.01$, this chance is 0.9999568!

- One needs to **adjust for multiple hypothesis testing** (MHT).

Multiple Hypothesis Testing Framework

Hypothesis testing is concerned with using **observed data** to **make decisions** regarding properties of (i.e., hypotheses for) the **unknown data generating distribution**.

A **null hypothesis** states that the data generating distribution belongs to a particular **submodel**, i.e., a set of possibly non-parametric distributions.

Null hypotheses are often expressed in terms of **parameters**, defined as functions of the data generating distribution.

E.g. **Mean vector** or **correlation matrix** of a multivariate distribution of microarray expression measures.

Regression coefficients in linear or non-linear models relating phenotypes to multilocus SNP genotypes.

Multiple Hypothesis Testing Framework

A testing procedure is a data-driven rule for deciding which null hypotheses should be rejected, i.e., declared false.

The decisions to reject or not the null hypotheses are based on test statistics, defined as functions of the empirical distribution, i.e., of the data.

E.g. t -statistics, χ^2 -statistics, F -statistics, and likelihood ratio statistics.

A multiple testing procedure (MTP), for the simultaneous test of $M \geq 1$ null hypotheses, provides rejection regions for each of the M hypotheses, i.e., sets of values for each of M test statistics that lead to the decision to reject the corresponding null hypotheses.

In other words, a MTP produces a random (i.e., data-dependent) set of rejected hypotheses that estimates the set of false null hypotheses.

Multiple Hypothesis Testing Framework

In any testing problem, two types of errors can be committed.

A **Type I error**, or **false positive**, is committed by rejecting a true null hypothesis.

A **Type II error**, or **false negative**, is committed by failing to reject a false null hypothesis.

Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors. Unfortunately, this is not feasible and one seeks a **trade-off** between the two types of errors.

This trade-off typically involves the **minimization of Type II errors**, i.e., the **maximization of power**, subject to a **Type I error constraint**.

Multiple Hypothesis Testing Framework

Whether testing single or multiple hypotheses, one needs the (joint) **distribution of the test statistics** in order to derive rejection regions for procedures that **probabilistically control Type I errors**.

In practice, however, the **true distribution of the test statistics** is **unknown and replaced** by a **null distribution**.

The choice of a proper null distribution is crucial in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the assumed null distribution does indeed provide the desired control under the true distribution.

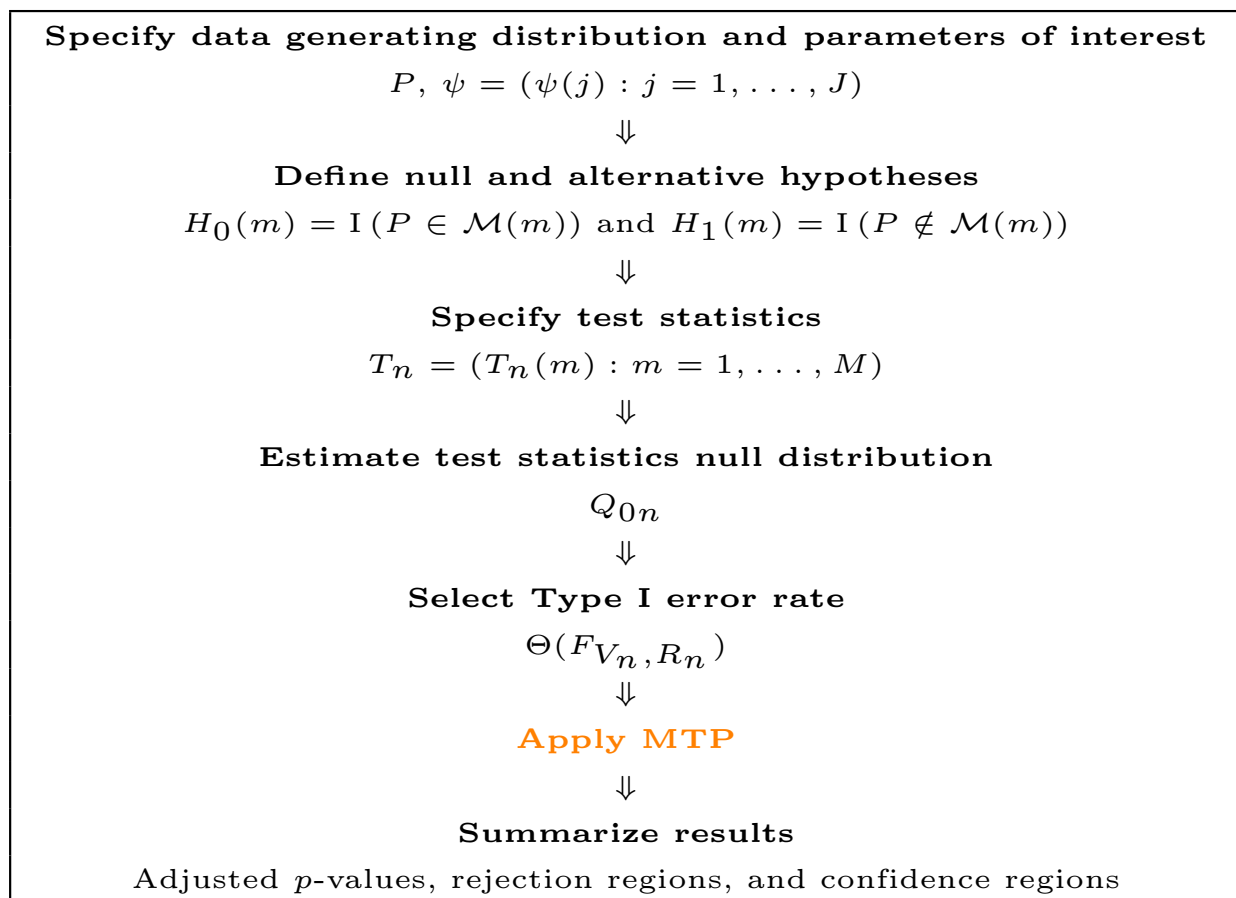
Resampling procedures (e.g., bootstrap and permutation) are particularly useful in this context.

Multiple Hypothesis Testing Framework

- **Data generating distribution:** $\mathcal{X}_n = \{X_i : i = 1, \dots, n\} \stackrel{IID}{\sim} P \in \mathcal{M}$.
- **Parameters:** $\psi = (\psi(j) : j = 1, \dots, J)$, where $\psi(j) = \Psi(P)(j)$.
- **Null and alternative hypotheses:** $H_0(m) = \mathbb{I}(P \in \mathcal{M}(m))$ and $H_1(m) = \mathbb{I}(P \notin \mathcal{M}(m))$, where $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \dots, M$.
- **Test statistics:** $T_n = (T_n(m) : m = 1, \dots, M)$, where $T_n(m) = T(m; \mathcal{X}_n) = T(m; P_n)$.
- **Test statistics null distribution:** Q_0 (or estimator thereof, Q_{0n}).
- **Multiple testing procedures:**
 $\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) = \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m) \text{ is rejected}\}$.
- **Type I error rate:** $\theta_n = \Theta(F_{V_n, R_n})$, where $V_n = \#$ Type I errors and $R_n = \#$ rejected hypotheses.
- **Type II error rate/power:** $\vartheta_n = \Theta(F_{U_n, R_n})$, where $U_n = \#$ Type II errors.
- **Summaries of results:** adjusted p -values, test statistic rejection regions, parameter confidence regions.

Multiple Hypothesis Testing Framework

Table 1: *Multiple hypothesis testing flowchart.*



Multiple Hypothesis Testing Framework

Apply MTP

FWER	$\Pr(V_n > 0)$	Single-step common-cut-off maxT Single-step common-quantile minP Step-down common-cut-off maxT Step-down common-quantile minP Resampling-based empirical Bayes
gFWER	$\Pr(V_n > k)$	Single-step common-cut-off $T(k + 1)$ Single-step common-quantile $P(k + 1)$ Augmentation Resampling-based empirical Bayes
General	$\Theta(F_{V_n})$	Single-step common-cut-off Single-step common-quantile Resampling-based empirical Bayes
TPPFP	$\Pr(V_n / R_n > q)$	Augmentation Resampling-based empirical Bayes
gTP	$\Pr(g(V_n, R_n) > q)$	Augmentation Resampling-based empirical Bayes
FDR	$E[V_n / R_n]$	TPPFP-based Resampling-based empirical Bayes
gEV	$E[g(V_n, R_n)]$	gTP-based Resampling-based empirical Bayes
General	$\Theta(F_{g(V_n, R_n)})$	Resampling-based empirical Bayes

Data Generating Distribution

Data generating distribution. Hypothesis testing is concerned with using **observed data** to **make decisions** regarding properties of (i.e., hypotheses for) the **unknown data generating distribution**.

Let $\mathcal{X}_n \equiv \{X_i : i = 1, \dots, n\}$ denote a **random sample** of n independent and identically distributed (IID) random variables from a **data generating distribution** P .

Suppose that the data generating distribution P is an element of a particular **statistical model** \mathcal{M} , i.e., a set of possibly non-parametric distributions.

$$X_1, \dots, X_n \stackrel{IID}{\sim} P \in \mathcal{M}.$$

Let P_n denote the **empirical distribution**, which places probability $1/n$ on each realization of X in \mathcal{X}_n .

Parameters

Parameters. Define **parameters** as arbitrary functions of the data generating distribution P :

$$\Psi(P) = \psi = (\psi(j) : j = 1, \dots, J) \in \mathbb{R}^J, \text{ where} \\ \psi(j) = \Psi(P)(j) \in \mathbb{R}.$$

Parameters of interest include (functions of) means, quantiles, covariances, correlation coefficients, and regression coefficients.

Parameters

E.g. For a J -dimensional random vector $X = (X(j) : j = 1, \dots, J) \in \mathbb{R}^J$, with $X \sim P$, one may define the following parameters.

Expected values

$$\psi(j) = \mathbb{E}[X(j)] = \int x(j) dP(x(j)).$$

Covariances

$$\psi(j, j') = \text{Cov}[X(j), X(j')] = \int x(j)x(j') dP(x(j), x(j')) - \mathbb{E}[X(j)] \mathbb{E}[X(j')].$$

Null and Alternative Hypotheses

General submodel hypotheses. Define M pairs of null and alternative hypotheses in terms of a collection of M **submodels**, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \dots, M$, for the data generating distribution P . Specifically, the M **null hypotheses** and corresponding **alternative hypotheses** are defined, respectively, as

$$H_0(m) \equiv \mathbb{I}(P \in \mathcal{M}(m)) \quad \text{and} \quad H_1(m) \equiv \mathbb{I}(P \notin \mathcal{M}(m)). \quad (1)$$

Thus, $H_0(m)$ is true (i.e., $H_0(m) = 1$) if the data generating distribution P belongs to submodel $\mathcal{M}(m)$;

$H_0(m)$ is false otherwise (i.e., $H_0(m) = 0$).

This general submodel representation covers tests of means, quantiles, covariances, correlation coefficients, and regression coefficients in linear and non-linear models (e.g., logistic, survival, time-series models).

Null and Alternative Hypotheses

Parametric hypotheses. In many testing problems, the **submodels concern parameters**, i.e., functions

$\Psi(P) = \psi = (\psi(m) : m = 1, \dots, M) \in \mathbb{R}^M$ of the data generating distribution P , and each null hypothesis $H_0(m)$ refers to a single parameter, $\psi(m) = \Psi(P)(m) \in \mathbb{R}$.

One-sided tests $H_0(m) = \mathbb{I}(\psi(m) \leq \psi_0(m))$ (2)

vs. $H_1(m) = \mathbb{I}(\psi(m) > \psi_0(m)), \quad m = 1, \dots, M.$

Two-sided tests $H_0(m) = \mathbb{I}(\psi(m) = \psi_0(m))$ (3)

vs. $H_1(m) = \mathbb{I}(\psi(m) \neq \psi_0(m)), \quad m = 1, \dots, M.$

The hypothesized **null values**, $\psi_0(m)$, are frequently zero.

Null and Alternative Hypotheses

Sets of true and false null hypotheses. Let

$$\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathcal{M}(m)\} \quad (4)$$

denote the set of $h_0 \equiv |\mathcal{H}_0|$ true null hypotheses.

Likewise, let

$$\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \{m : H_1(m) = 1\} = \{m : P \notin \mathcal{M}(m)\} = \mathcal{H}_0^c(P) \quad (5)$$

be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ false null hypotheses.

The goal of a multiple testing procedure is to accurately estimate, i.e., reject, the set \mathcal{H}_1 , while probabilistically controlling false positives.

Null and Alternative Hypotheses

Complete null hypothesis. The complete null hypothesis H_0^C is defined as

$$H_0^C \equiv \prod_{m=1}^M H_0(m) = \prod_{m=1}^M \mathbf{I}(P \in \mathcal{M}(m)) = \mathbf{I}(P \in \cap_{m=1}^M \mathcal{M}(m)). \quad (6)$$

The complete null hypothesis is true if and only if all M individual null hypotheses $H_0(m)$ are true, i.e., if and only if the data generating distribution P belongs to the intersection $\cap_{m=1}^M \mathcal{M}(m)$ of the M submodels.

Test Statistics

Test statistics. A testing procedure is a random or data-driven rule for deciding which null hypotheses should be rejected, i.e., for estimating the set of false null hypotheses

$$\mathcal{H}_1 = \{m : H_0(m) = 0\} = \{m : P \notin \mathcal{M}(m)\}.$$

The decisions to reject or not the null hypotheses are based on an M -vector of test statistics, $T_n = (T_n(m) : m = 1, \dots, M)$, that are functions $T_n(m) = T(m; \mathcal{X}_n) = T(m; P_n)$ of the data \mathcal{X}_n , i.e., of the empirical distribution P_n .

Denote the typically unknown (finite sample) joint distribution of the test statistics T_n by $Q_n = Q_n(P)$.

Test Statistics

For the test of single-parameter null hypotheses of the form $H_0(m) = \mathbf{I}(\psi(m) \leq \psi_0(m))$ or $H_0(m) = \mathbf{I}(\psi(m) = \psi_0(m))$, $m = 1, \dots, M$, consider two main types of test statistics, **difference statistics**,

$$T_n(m) \equiv \text{Estimator} - \text{Null value} = \sqrt{n}(\psi_n(m) - \psi_0(m)), \quad (7)$$

and **t-statistics** (i.e., standardized differences),

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n} \frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \quad (8)$$

Here, $\hat{\Psi}(P_n) = \psi_n = (\psi_n(m) : m = 1, \dots, M)$ denotes an **estimator** for the parameter $\Psi(P) = \psi = (\psi(m) : m = 1, \dots, M)$ and $(\sigma_n(m)/\sqrt{n} : m = 1, \dots, M)$ denote the estimated **standard errors** for elements $\psi_n(m)$ of ψ_n .

Test Statistics

This general representation for the test statistics covers standard one-sample and two-sample t -statistics for testing hypotheses concerning mean parameters, but also test statistics for correlation coefficients and regression coefficients in linear and non-linear models.

Test statistics for other types of null hypotheses include F -statistics, χ^2 -statistics, and likelihood ratio statistics.

Multiple Testing Procedures

Multiple testing procedures. A **multiple testing procedure** (MTP) provides **rejection regions** $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the corresponding null hypothesis $H_0(m)$ and declare that $P \notin \mathcal{M}(m)$, $m = 1, \dots, M$.

In other words, a MTP produces a random (i.e., data-dependent) **set of rejected hypotheses** \mathcal{R}_n that **estimates the set of false null hypotheses** \mathcal{H}_1 ,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : H_0(m) \text{ is rejected}\}, \quad (9)$$

where $\mathcal{C}_n(m) = \mathcal{C}(m; T_n, Q_{0n}, \alpha)$, $m = 1, \dots, M$, denote possibly random rejection regions.

Multiple Testing Procedures

The long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ and $\mathcal{C}(m; T_n, Q_{0n}, \alpha)$ emphasizes that the MTP depends on the following three ingredients:

1. the **data**, $\mathcal{X}_n = \{X_i : i = 1, \dots, n\}$, through the M -vector of **test statistics**, $T_n = (T_n(m) : m = 1, \dots, M)$;
2. an (estimated) M -variate **test statistics null distribution**, Q_{0n} , for deriving rejection regions, confidence regions, and adjusted p -values.
3. the **nominal Type I error level** α , i.e., a user-supplied upper bound for a suitably defined Type I error rate.

Rejection Regions

Rejection regions. Given a proper test statistics null distribution Q_0 (or estimator thereof, Q_{0n}), the main task is to specify **rejection regions** for each null hypothesis, so that the resulting procedure probabilistically controls Type I errors.

We consider MTPs based on **nested rejection regions**, so that

$$\mathcal{C}(m; T_n, Q_{0n}, \alpha_1) \subseteq \mathcal{C}(m; T_n, Q_{0n}, \alpha_2), \quad \text{whenever } \alpha_1 \leq \alpha_2. \quad (10)$$

Rejection Regions

Rejection regions are typically defined in terms of **intervals**, such as,

$$\mathcal{C}_n(m) = (u_n(m), +\infty), \quad (11)$$

$$\mathcal{C}_n(m) = (-\infty, l_n(m)),$$

$$\mathcal{C}_n(m) = (-\infty, l_n(m)) \cup (u_n(m), +\infty),$$

where $l_n(m) = l(m; T_n, Q_{0n}, \alpha)$ and $u_n(m) = u(m; T_n, Q_{0n}, \alpha)$ are to-be-determined lower and upper **critical values**, or **cut-offs**, computed under the null distribution Q_{0n} for the test statistics T_n .

Rejection Regions

Two-sided rejection regions of the form

$\mathcal{C}_n(m) = (-\infty, l_n(m)) \cup (u_n(m), +\infty)$ allow the use of asymmetric cut-offs for two-sided tests.

Unless specified otherwise, we assume that large values of the test statistic $T_n(m)$ provide evidence against the corresponding null hypothesis $H_0(m)$, that is, we consider one-sided rejection regions of the form $\mathcal{C}_n(m) = (c_n(m), +\infty)$, where $c_n(m) = c(m; T_n, Q_{0n}, \alpha)$.

For two-sided tests of single-parameter null hypotheses using difference or t -statistics, as in Equations (7) and (8), one could take absolute values of the test statistics.

Errors in MHT: Type I, Type II, and Type III errors

In any testing problem, two types of errors can be committed.

- A **Type I error**, or **false positive**, is committed by rejecting a true null hypothesis ($\mathcal{R}_n \cap \mathcal{H}_0$).
E.g. Report an association between phenotype and genotype when in truth there is no such association.
- A **Type II error**, or **false negative**, is committed by failing to reject a false null hypothesis ($\mathcal{R}_n^c \cap \mathcal{H}_1$).
E.g. Fail to identify a true association between phenotype and genotype.

Errors in MHT: Type I, Type II, and Type III errors

Table 2: *Type I and Type II errors in multiple hypothesis testing.*

		Null hypotheses		
		Non-rejected, \mathcal{R}_n^c	Rejected, \mathcal{R}_n	
True, \mathcal{H}_0	$W_n = \mathcal{R}_n^c \cap \mathcal{H}_0 $	$V_n = \mathcal{R}_n \cap \mathcal{H}_0 $		h_0
False, \mathcal{H}_1	$U_n = \mathcal{R}_n^c \cap \mathcal{H}_1 $	$S_n = \mathcal{R}_n \cap \mathcal{H}_1 $		h_1
		$M - R_n$	R_n	M

Errors in MHT: Type I, Type II, and Type III errors

The number of **rejected null hypotheses** is

$$R_n \equiv |\mathcal{R}_n| = \sum_{m=1}^M \mathbf{I}(T_n(m) \in \mathcal{C}_n(m)), \quad (12)$$

the number of **Type I errors** or **false positives** is

$$V_n \equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathbf{I}(T_n(m) \in \mathcal{C}_n(m)), \quad (13)$$

the number of **Type II errors** or **false negatives** is

$$U_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathbf{I}(T_n(m) \notin \mathcal{C}_n(m)), \quad (14)$$

the number of **true negatives** is

$$W_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathbf{I}(T_n(m) \notin \mathcal{C}_n(m)) = M - R_n - U_n = h_0 - V_n, \quad (15)$$

and the number of **true positives** is

$$S_n \equiv |\mathcal{R}_n \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathbf{I}(T_n(m) \in \mathcal{C}_n(m)) = R_n - V_n = h_1 - U_n. \quad (16)$$

Errors in MHT: Type I, Type II, and Type III errors

Note that S_n , U_n , V_n , and W_n each depend on the unknown data generating distribution P through the unknown set of true null hypotheses $\mathcal{H}_0 = \mathcal{H}_0(P)$.

Therefore,

- the numbers $h_0 = |\mathcal{H}_0|$ and $h_1 = |\mathcal{H}_1| = M - h_0$ of true and false null hypotheses are **unknown parameters** (row margins of Table 2),
- the number of rejected hypotheses R_n is an **observable random variable** (column margins of Table 2), and
- S_n , U_n , V_n , and W_n are **unobservable random variables** (cells of Table 2).

Errors in MHT: Type I, Type II, and Type III errors

Ideally, one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors.

Unfortunately, this is not feasible and one seeks a **trade-off** between the two types of errors.

A standard approach is to specify an acceptable level α for a suitably defined Type I error rate and derive testing procedures (i.e., rejection regions) that aim to **minimize a Type II error rate**, i.e., **maximize power**, within the class of tests with **Type I error level at most α** .

Errors in MHT: Type I, Type II, and Type III errors

For two-sided tests concerning single-parameter null hypotheses, one is often interested in determining the **direction of rejection** for the null hypotheses.

For instance, in microarray experiments, one may wish to know whether genes are **over-** or **under-expressed** in, say, treated cells compared to untreated cells.

In this setting, one can commit a **Type III error** by correctly rejecting a false null hypothesis $H_0(m) = \mathbb{I}(\psi(m) = \psi_0(m))$, but incorrectly concluding that $\psi(m) < \psi_0(m)$ when in truth $\psi(m) > \psi_0(m)$ (or vice versa).

Control of Type III errors, as well as Type I errors, brings in additional complexities and is not considered here.

Type I Error Rates

Type I error rates. When testing multiple hypotheses, there are many possible definitions for the Type I error rate and power of a testing procedure.

Accordingly, we define a **Type I error rate** as a **parameter**

$\theta_n = \Theta(F_{V_n, R_n})$ of the joint distribution F_{V_n, R_n} of the numbers of Type I errors $V_n = |\mathcal{R}_n \cap \mathcal{H}_0|$ and rejected hypotheses $R_n = |\mathcal{R}_n|$.

We focus primarily on Type I error rates such that

$$\Theta(F_{V_n, R_n}) \in [0, 1].$$

Type I Error Rates

Such a representation covers a broad class of Type I error rates, defined as **generalized tail probability** (gTP) error rates,

$$gTP(q, g) \equiv \Pr(g(V_n, R_n) > q), \quad (17)$$

and **generalized expected value** (gEV) error rates,

$$gEV(g) \equiv E[g(V_n, R_n)], \quad (18)$$

for arbitrary functions $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n .

The special case $g(v, r) = v$ corresponds to the gFWER and PFER.

The special case $g(v, r) = v/r$ corresponds to the TPPFP and FDR.

Type I Error Rates

Type I error rates based on the distribution of the number of Type I errors: $\Theta(F_{V_n})$.

- The **family-wise error rate** (FWER) is the probability of at least one Type I error,

$$FWER \equiv \Pr(V_n > 0) = 1 - F_{V_n}(0). \quad (19)$$

- The **generalized family-wise error rate** (gFWER), for a user-supplied integer $k \in \{0, \dots, M\}$, is the probability of at least $(k + 1)$ Type I errors. That is,

$$gFWER(k) \equiv \Pr(V_n > k) = 1 - F_{V_n}(k). \quad (20)$$

When $k = 0$, the gFWER reduces to the usual family-wise error rate, FWER.

Type I Error Rates

- The **per-comparison error rate** (PCER) is the expected value of the proportion of Type I errors among the M tests,

$$PCER \equiv \frac{1}{M} \mathbb{E}[V_n] = \frac{1}{M} \int v dF_{V_n}(v). \quad (21)$$

- The **per-family error rate** (PFER) is the expected value of the number of Type I errors,

$$PFER \equiv \mathbb{E}[V_n] = \int v dF_{V_n}(v). \quad (22)$$

Type I Error Rates

Type I error rates based on the distribution of the proportion of Type I errors among the rejected hypotheses: $\Theta(F_{V_n/R_n})$.

- The **tail probability for the proportion of false positives (TPFP)** among the rejected hypotheses, for a user-supplied constant $q \in (0, 1)$, is defined as

$$TPFP(q) \equiv \Pr\left(\frac{V_n}{R_n} > q\right) = 1 - F_{V_n/R_n}(q). \quad (23)$$

- The **false discovery rate (FDR)** is the expected value of the proportion of Type I errors among the rejected hypotheses,

$$FDR \equiv \mathbb{E}\left[\frac{V_n}{R_n}\right] = \int q dF_{V_n/R_n}(q). \quad (24)$$

Error rates of the form $\Theta(F_{V_n/R_n})$ are defined with the convention that $V_n/R_n \equiv 0$ if $R_n = 0$.

Type I Error Rates

The FDR may be rewritten as

$$\begin{aligned}
 FDR &= \mathbb{E} \left[\frac{V_n}{\max \{R_n, 1\}} \right] & (25) \\
 &= \mathbb{E} \left[\frac{V_n}{R_n} \mid V_n > 0 \right] \Pr(V_n > 0) \\
 &= \mathbb{E} \left[\frac{V_n}{R_n} \mid R_n > 0 \right] \Pr(R_n > 0).
 \end{aligned}$$

Under the **complete null hypothesis** $H_0^C = \mathbb{I} \left(P \in \bigcap_{m=1}^M \mathcal{M}(m) \right)$, all R_n rejected hypotheses are Type I errors, hence $V_n/R_n = 1$ and $FDR = FWER = \Pr(V_n > 0)$.

FDR-controlling procedures therefore also control the FWER in the weak sense.

In general, because $V_n/R_n \leq 1$, the FDR is less than or equal to the FWER for any given MTP, $FDR \leq FWER$.

Type I Error Rates

Error rates $\Theta(F_{V_n/R_n})$, based on the **proportion** of false positives (e.g., TPPFP and FDR), are especially **appealing for the large-scale testing** problems encountered in genomics, compared to error rates $\Theta(F_{V_n})$, based on the **number** of false positives (e.g., gFWER and PFER), as they do not increase exponentially with the number M of tested hypotheses.

However, error rates $\Theta(F_{V_n/R_n})$ tend to be **more difficult to control** than error rates $\Theta(F_{V_n})$, as they are based on the **joint** distribution of V_n and R_n , rather than only the marginal distribution of V_n . In particular, error rates $\Theta(F_{V_n/R_n})$ involve the distribution of test statistics for the **false null hypotheses** \mathcal{H}_1 , via the number $S_n = |\mathcal{R}_n \cap \mathcal{H}_1| = R_n - V_n$ of true positives.

Type I Error Rates

For controlling general Type I error rates $\Theta(F_{V_n})$, our proposed multiple testing procedures rely on the following two assumptions concerning the mapping $\Theta : F \rightarrow \Theta(F)$, that defines the Type I error rate as a parameter of the distribution F_{V_n} of the number of Type I errors V_n .

Given two cumulative distribution functions (CDF) F_1 and F_2 on $\{0, \dots, M\}$, define a **distance measure** d by

$$d(F_1, F_2) \equiv \max_{x=0, \dots, M} |F_1(x) - F_2(x)|. \quad (26)$$

Type I Error Rates

Assumption $\mathbf{M}\Theta$. [Monotonicity of Θ] The mapping Θ is **non-decreasing**. That is, given two CDFs F_1 and F_2 on $\{0, \dots, M\}$,

$$F_1 \geq F_2 \quad \implies \quad \Theta(F_1) \leq \Theta(F_2). \quad (27)$$

Assumption $\mathbf{C}\Theta$. [Continuity of Θ] The mapping Θ is **uniformly continuous**. That is, given two sequences $\{F_{1n}\}$ and $\{F_{2n}\}$ of CDFs on $\{0, \dots, M\}$,

$$\lim_{n \rightarrow \infty} d(F_{1n}, F_{2n}) = 0 \quad \implies \quad \lim_{n \rightarrow \infty} (\Theta(F_{2n}) - \Theta(F_{1n})) = 0. \quad (28)$$

In most cases, we only need continuity at a fixed CDF F_1 , i.e., the special case $F_{1n} = F_1$.

Power

Power. Within a class of multiple testing procedures that control a given Type I error rate $\theta_n = \Theta(F_{V_n, R_n})$ at an acceptable level α , one seeks procedures that maximize power, that is, minimize a suitably defined Type II error rate.

As with Type I error rates, the concepts of Type II error rate and power can be extended in various ways when moving from single to multiple hypothesis testing.

Accordingly, we define **power** as a parameter $\vartheta_n = \Theta(F_{U_n, R_n})$ of the joint distribution F_{U_n, R_n} of the numbers of Type II errors $U_n = |\mathcal{R}_n^c \cap \mathcal{H}_1|$ and rejected hypotheses $R_n = |\mathcal{R}_n|$.

Recall that the numbers of true positives S_n and Type II errors U_n satisfy $S_n + U_n = h_1$ (Table 2).

Power

- The probability of rejecting **at least one** false null hypothesis, i.e., of at least one true positive,

$$AnyPwr \equiv \Pr(S_n \geq 1) = \Pr(U_n \leq h_1 - 1) = F_{U_n}(h_1 - 1). \quad (29)$$

- The probability of rejecting **all** false null hypotheses, i.e., of no Type II errors,

$$AllPwr \equiv \Pr(S_n = h_1) = \Pr(U_n = 0) = F_{U_n}(0). \quad (30)$$

- The **average power**, i.e., the expected value of the proportion of rejected hypotheses among the false null hypotheses,

$$AvgPwr \equiv \frac{1}{h_1} \mathbb{E}[S_n] = \frac{1}{h_1} \mathbb{E}[h_1 - U_n] = 1 - \frac{1}{h_1} \int u dF_{U_n}(u). \quad (31)$$

Power

- The **true discovery rate** (TDR), i.e., the expected value of the proportion of true positives among the rejected hypotheses,

$$TDR \equiv \mathbb{E} \left[\frac{S_n}{R_n} \right] = \mathbb{E} \left[\frac{R_n - V_n}{R_n} \right], \quad (32)$$

with the convention that $(R_n - V_n)/R_n \equiv 0$ if $R_n = 0$. The TDR may be rewritten as

$$TDR = \mathbb{E} \left[\frac{R_n - V_n}{R_n} \mid R_n > 0 \right] \Pr(R_n > 0) = \Pr(R_n > 0) - FDR.$$

One can think of the TDR as a power analogue of the FDR. If all null hypotheses are false (i.e., $h_1 = M$), then *TDR* reduces to *AnyPwr*.

Unadjusted and Adjusted p -Values

As in the case of single hypothesis testing, one can report the results of a multiple testing procedure in terms of the following quantities.

- **Rejection regions** for the test statistics.
- **Confidence regions** for the parameters of interest.
- **Adjusted p -values**. The adjusted p -value for a particular null hypothesis is the smallest nominal Type I error level (for the multiple test of all M hypotheses) at which one would reject this null hypothesis.

The smaller the adjusted p -value, the stronger the evidence against the corresponding null hypothesis.

Unadjusted and Adjusted p -Values

Unadjusted p -values. Consider testing M null hypotheses $H_0(m)$, $m = 1, \dots, M$, individually at level α , based on test statistics $T_n = (T_n(m) : m = 1, \dots, M)$, with unknown true distribution $Q_n = Q_n(P)$ and assumed null distribution Q_0 (or estimator thereof, Q_{0n}).

Null hypothesis $H_0(m)$ is rejected at **single test nominal Type I error level α** if $T_n(m) \in \mathcal{C}_n(m; \alpha)$.

The **rejection regions** $\mathcal{C}_n(m; \alpha) = \mathcal{C}(Q_{0,m}, \alpha)$ are based solely on the **marginal** null distributions $Q_{0,m}$ (or estimators thereof, $Q_{0n,m}$) and are chosen such that the **chance of a Type I error is at most α for each test**,

$$\Pr_{Q_{0,m}} (T_n(m) \in \mathcal{C}_n(m; \alpha)) \leq \alpha, \quad (33)$$

and the nestedness assumption of Equation (10) is satisfied.

Unadjusted and Adjusted p -Values

The **unadjusted p -value** $P_{0n}(m) = P(T_n(m), Q_{0,m})$, for the **single test** of null hypothesis $H_0(m)$, is defined as

$$\begin{aligned} P_{0n}(m) &\equiv \inf \{ \alpha \in [0, 1] : \text{Reject } H_0(m) \text{ at single test nominal level } \alpha \} \\ &= \inf \{ \alpha \in [0, 1] : T_n(m) \in \mathcal{C}_n(m; \alpha) \}, \quad m = 1, \dots, M. \end{aligned} \quad (34)$$

That is, the unadjusted p -value $P_{0n}(m)$, for null hypothesis $H_0(m)$, is the **smallest nominal Type I error level** of the **single hypothesis testing procedure** at which one would reject $H_0(m)$, given $T_n(m)$.

The **smaller** the unadjusted p -value $P_{0n}(m)$, the **stronger the evidence** against the corresponding null hypothesis $H_0(m)$.

Unadjusted p -values may also be referred to as **marginal** or **raw p -values**.

Unadjusted and Adjusted p -Values

For **one-sided rejection regions** of the form

$\mathcal{C}_n(m; \alpha) = (c_n(m; \alpha), +\infty)$ and continuous and strictly increasing marginal null distributions $Q_{0,m}$, unadjusted p -values are given by

$$P_{0n}(m) = 1 - Q_{0,m}(T_n(m)), \quad m = 1, \dots, M. \quad (35)$$

Unadjusted and Adjusted p -Values

Adjusted p -values. Adjusted p -values, for the test of multiple hypotheses, are defined as straightforward extensions of unadjusted p -values, for the test of individual hypotheses.

Consider any multiple testing procedure $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_0, \alpha)$, with rejection regions $\mathcal{C}_n(m; \alpha) = \mathcal{C}(m; T_n, Q_0, \alpha)$. Then, one can define an M -vector of **adjusted p -values**, $\tilde{P}_{0n} = (\tilde{P}_{0n}(m) : m = 1, \dots, M) = \tilde{P}(T_n, Q_0) = \tilde{P}(\mathcal{R}(T_n, Q_0, \alpha) : \alpha \in [0, 1])$, as

$$\begin{aligned} \tilde{P}_{0n}(m) &\equiv \inf \{ \alpha \in [0, 1] : \text{Reject } H_0(m) \text{ at nominal MTP level } \alpha \} \\ &= \inf \{ \alpha \in [0, 1] : m \in \mathcal{R}_n(\alpha) \} \\ &= \inf \{ \alpha \in [0, 1] : T_n(m) \in \mathcal{C}_n(m; \alpha) \}, \quad m = 1, \dots, M. \end{aligned}$$

Unadjusted and Adjusted p -Values

That is, the adjusted p -value $\tilde{P}_{0n}(m)$, for null hypothesis $H_0(m)$, is the **smallest nominal Type I error level** (e.g., gFWER, TPPFP, or FDR) of the **multiple hypothesis testing procedure** at which one would reject $H_0(m)$, given T_n .

As in single hypothesis tests, the **smaller** the adjusted p -value $\tilde{P}_{0n}(m)$, the **stronger the evidence** against the corresponding null hypothesis $H_0(m)$.

Thus, one rejects $H_0(m)$ for small adjusted p -values $\tilde{P}_{0n}(m)$.

For instance, the adjusted p -values for classical FWER-controlling **Bonferroni procedure** are

$$\tilde{P}_{0n}(m) = \min \{MP_{0n}(m), 1\}.$$

Unadjusted and Adjusted p -Values

Under the nestedness assumption of Equation (10), one has two **equivalent representations** for a MTP, in terms of **rejection regions** for the test statistics and in terms of **adjusted p -values**.

Specifically, the set of rejected null hypotheses at multiple test nominal Type I error level α is

$$\mathcal{R}_n(\alpha) = \{m : T_n(m) \in \mathcal{C}_n(m; \alpha)\} = \left\{m : \tilde{P}_{0n}(m) \leq \alpha\right\}. \quad (37)$$

Unadjusted and Adjusted p -Values

As in the single hypothesis case, reporting the results of a MTP in terms of **adjusted p -values**, as opposed to only rejection or not of the null hypotheses, offers several **advantages**.

- Adjusted p -values can be defined for **any Type I error rate** (e.g., gFWER, TPPFP, or FDR).
- They reflect the strength of the evidence against each null hypothesis in terms of the **Type I error rate for the entire MTP**.
- They are **flexible summaries** of a MTP, in the sense that results are supplied for **all Type I error levels α** , i.e., the level α need not be chosen ahead of time.

Unadjusted and Adjusted p -Values

- They provide convenient **benchmarks to compare different MTPs**, whereby smaller adjusted p -values indicate a less conservative procedure.
- **Plots of sorted adjusted p -values** allow investigators to examine sets of rejected hypotheses associated with various Type I error rates (e.g., gFWER, TPPFP, or FDR) and nominal levels α . Such plots provide tools to decide on an appropriate combination of the number of rejected hypotheses and tolerable false positive rate for a particular experiment and available resources.

Examples of MTPs

Given a suitable test statistics null distribution Q_0 (or estimator thereof, Q_{0n}), the main task is to specify **rejection regions** for each null hypothesis, i.e., **cut-offs** for each test statistic.

Among the different approaches for defining rejection regions, we distinguish the following.

- **Marginal vs. joint** multiple testing procedures.
- **Single-step vs. stepwise** multiple testing procedures.
- **Common-cut-off vs. common-quantile** multiple testing procedures.

Examples of MTPs: Marginal vs. Joint

- **Marginal** procedures are based solely on the **marginal distributions of the test statistics** (e.g., FWER-controlling single-step Bonferroni procedure).
- **Joint** procedures take into account the **dependence structure of the test statistics** (e.g., FWER-controlling single-step maxT procedure).

Joint MTPs tend to be **more powerful** than marginal MTPs.

N.B. While a procedure may be marginal, proof of Type I error control by this MTP may require certain assumptions on the dependence structure of the test statistics (e.g., FWER-controlling step-up Hochberg procedure).

Examples of MTPs: Single-Step vs. Stepwise

- In **single-step** procedures, each null hypothesis $H_0(m)$ is tested using a rejection region that is independent of the results of the tests of other hypotheses and is not a function of the data \mathcal{X}_n (unless these data are used to estimate the null distribution).
- In **stepwise** procedures, the decision to reject a particular null hypothesis depends on the outcome of the tests of other hypotheses. That is, the (single-step) testing procedure is applied to a **sequence of successively smaller nested random (i.e., data-dependent) subsets of null hypotheses**, defined by the **ordering** of the test statistics (common-cut-off MTPs) or unadjusted p -values (common-quantile MTPs). The rejection regions are therefore allowed to depend on the data \mathcal{X}_n via the test statistics T_n .

Stepwise MTPs tend to be **more powerful** than single-step MTPs.

Examples of MTPs: Single-Step vs. Stepwise

Stepwise procedures are of two main types, depending on the **order in which the null hypotheses are tested**.

- In **step-down** procedures, the **most significant** null hypotheses (i.e., the null hypotheses with the largest test statistics for common-cut-off MTPs or smallest unadjusted p -values for common-quantile MTPs) are considered successively, with further tests depending on the outcome of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected.
- In contrast, for **step-up** procedures, the **least significant** null hypotheses are considered successively, again with further tests depending on the outcome of earlier ones. As soon as one null hypothesis is rejected, all remaining more significant hypotheses are rejected.

Examples of MTPs: Common-Cut-Off vs. Common-Quantile

- In **common-cut-off** procedures, the **same cut-off c_0** is used for each test statistic (e.g., FWER-controlling single-step and step-down maxT procedures, based on maxima of test statistics).
- In **common-quantile** procedures, the cut-offs are the **δ_0 -quantiles** of the marginal null distributions of the test statistics (e.g., FWER-controlling single-step and step-down minP procedures, based on minima of unadjusted p -values).

The latter p -value-based procedures place the null hypotheses on an “equal footing”, i.e., are **more balanced** than their common-cut-off counterparts, and may therefore be preferable.

However, this comes at the expense of increased computational complexity.

Examples of MTPs

Consider the test of M null hypotheses, $H_0(m)$, $m = 1, \dots, M$, based on test statistics $(T_n(m) : m = 1, \dots, M)$, with null distribution Q_0 .

Let $(P_{0n}(m) : m = 1, \dots, M)$ and $(\tilde{P}_{0n}(m) : m = 1, \dots, M)$ denote, respectively, the corresponding unadjusted and adjusted p -values.

Order the M null hypotheses according to their unadjusted p -values, from smallest to largest, that is, define indices $O_n(m)$, so that $P_{0n}(O_n(1)) \leq \dots \leq P_{0n}(O_n(M))$.

Examples of MTPs

FWER-controlling single-step Bonferroni (1936) MTP [Marginal, common-quantile]

$$\tilde{p}_{0n}(m) = \min \{M p_{0n}(m), 1\}. \quad (38)$$

FWER-controlling single-step maxT MTP [Joint, common-cut-off]

$$\tilde{p}_{0n}(m) = \Pr_{Q_0} \left(\max_{m=1, \dots, M} Z(m) \geq t_n(m) \right), \quad (39)$$

where $Z = (Z(m) : m = 1, \dots, M) \sim Q_0$.

FWER-controlling single-step minP MTP [Joint, common-quantile]

$$\tilde{p}_{0n}(m) = \Pr_{Q_0} \left(\min_{m=1, \dots, M} P_0(m) \leq p_{0n}(m) \right), \quad (40)$$

where $P_0(m) \equiv \bar{Q}_{0,m}(Z(m)) = 1 - Q_{0,m}(Z(m))$ and $Z \sim Q_0$.

Examples of MTPs

FWER-controlling step-down Holm (1979) MTP [Marginal, common-quantile]

$$\tilde{p}_{0n}(o_n(m)) = \max_{h=1, \dots, m} \{ \min \{ (M - h + 1) p_{0n}(o_n(h)), 1 \} \}. \quad (41)$$

FWER-controlling step-up Hochberg (1988) MTP* [Marginal, common-quantile]

$$\tilde{p}_{0n}(o_n(m)) = \min_{h=m, \dots, M} \{ \min \{ (M - h + 1) p_{0n}(o_n(h)), 1 \} \}. \quad (42)$$

FDR-controlling step-up Benjamini and Hochberg (1995) MTP* [Marginal, common-quantile]

$$\tilde{p}_{0n}(o_n(m)) = \min_{h=m, \dots, M} \left\{ \min \left\{ \frac{M}{h} p_{0n}(o_n(h)), 1 \right\} \right\}. \quad (43)$$

* Assumptions on the joint distribution of the test statistics.

Examples of MTPs

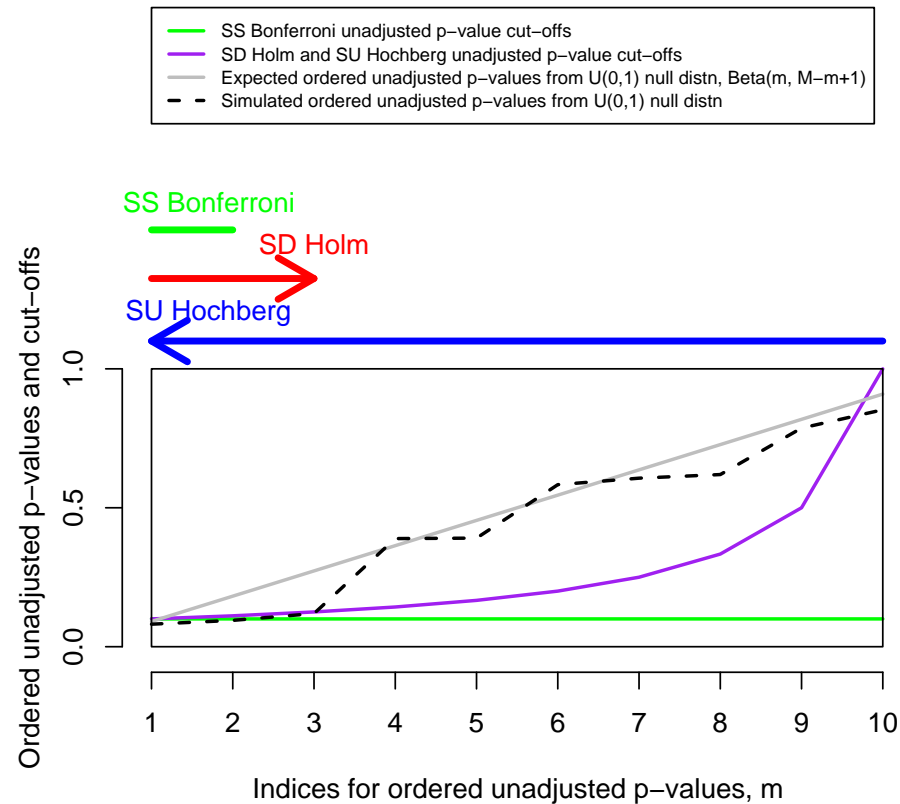


Figure 1: Comparison of single-step, step-down, and step-up procedures. Cut-offs for FWER-controlling marginal Bonferroni, Holm, and Hochberg procedures.

Examples of MTPs

Ordered unadjusted p -values

$$P_{0n}(O_n(1)) \leq P_{0n}(O_n(2)) \leq P_{0n}(O_n(3)) \leq \dots \leq P_{0n}(O_n(M))$$

Single-step Bonferroni adjusted p -values

$$MP_{0n}(O_n(1)) \leq MP_{0n}(O_n(2)) \leq MP_{0n}(O_n(3)) \leq \dots \leq MP_{0n}(O_n(M))$$

$$\tilde{P}_{0n}(O_n(m)) = \min \{MP_{0n}(O_n(m)), 1\}$$

Step-down Holm adjusted p -values

$$MP_{0n}(O_n(1)) \quad ? \quad (M - 1)P_{0n}(O_n(2)) \quad ? \quad (M - 2)P_{0n}(O_n(3)) \quad ? \quad \dots \quad ? \quad 1P_{0n}(O_n(M))$$



$$\tilde{P}_{0n}(O_n(m)) = \max_{h=1, \dots, m} \{ \min \{ (M - h + 1) P_{0n}(O_n(h)), 1 \} \}$$

Step-up Hochberg adjusted p -values

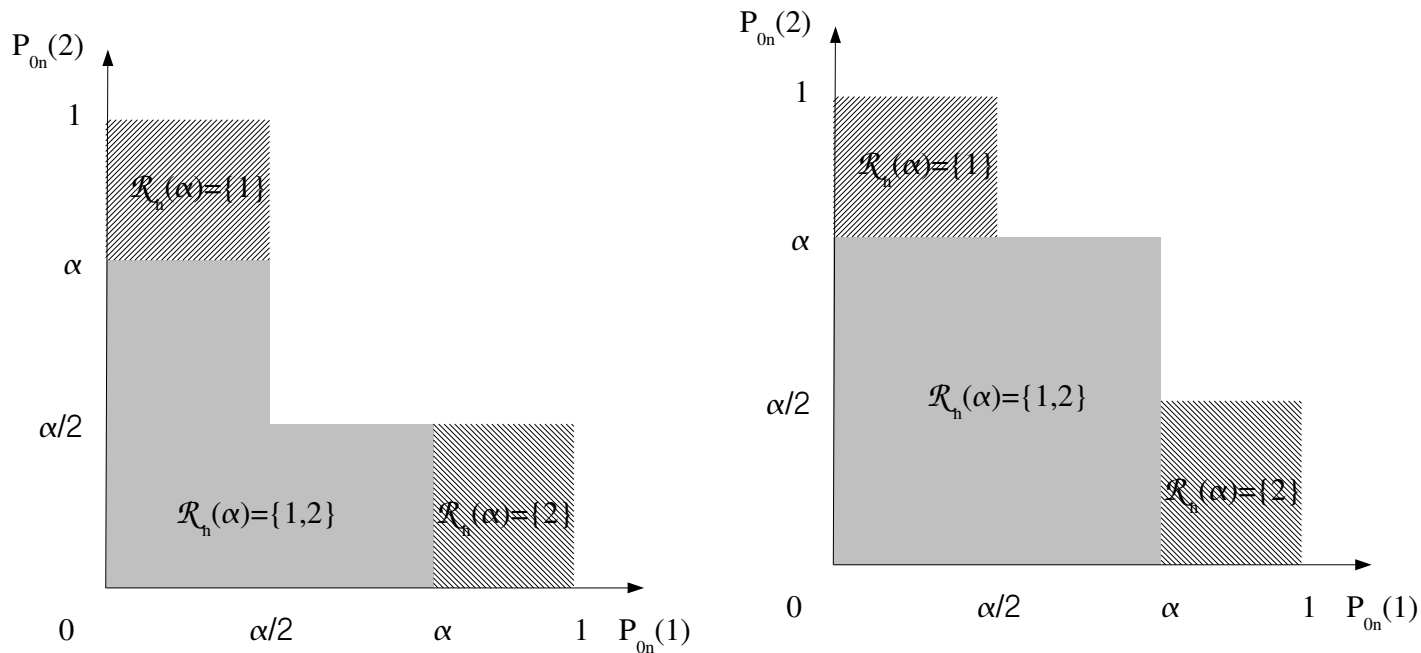
$$MP_{0n}(O_n(1)) \quad ? \quad (M - 1)P_{0n}(O_n(2)) \quad ? \quad (M - 2)P_{0n}(O_n(3)) \quad ? \quad \dots \quad ? \quad 1P_{0n}(O_n(M))$$



$$\tilde{P}_{0n}(O_n(m)) = \min_{h=m, \dots, M} \{ \min \{ (M - h + 1) P_{0n}(O_n(h)), 1 \} \}$$

Figure 2: Comparison of single-step, step-down, and step-up procedures. Adjusted p -values for FWER-controlling marginal Bonferroni, Holm, and Hochberg procedures.

Examples of MTPs



Step-down Holm MTP

Step-up Hochberg MTP

Figure 3: Comparison of step-down and step-up procedures. Rejection regions for FWER-controlling marginal Holm and Hochberg procedures, for the test of $M = 2$ null hypotheses.

References

- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.
- C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, pages 3–62, 1936.
- S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, New York, 2007. (In preparation).
- Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.