

Outline:

- P values,
- Multiple testing.

Reading:

- Chapter 10 in Wasserman.

A Reminder: Size of a Hypothesis Test

Recall the definition of the *size of a hypothesis test*, introduced in handout # 5c.

Definition 1. The *size of a hypothesis test* described by

$$\text{Rule } \phi: \mathcal{X}_0 = \{x : \phi(x) = 0\}, \quad \mathcal{X}_1 = \{x : \phi(x) = 1\}.$$

is defined as follows:

$$\alpha = \max_{\theta \in \text{sp}_{\Theta}(0)} \Pr_{X|\Theta}\{x \in \mathcal{X}_1 | \theta\} = \underbrace{\max_{\theta \in \text{sp}_{\Theta}(0)} P_{\text{FA}}(\phi(\mathbf{X}), \theta)}_{\text{max possible } P_{\text{FA}}} . \quad (1)$$

A hypothesis test is *said to have level α* if its size is less than or equal to α . Therefore, a level- α test is *guaranteed* to have a false-alarm probability less than or equal to α .

P Values

Our general approach in handout # 5c has been to set α in advance and to make a hard binary decision (“accept \mathcal{H}_0 ” or “accept \mathcal{H}_1 ”) depending on α . Such hard decisions do not convey information about how close we were to the opposite decision.

Generally, if \mathcal{H}_1 is accepted for a certain specified α , it will be accepted for $\alpha' > \alpha$. Therefore, there exists a smallest α at which \mathcal{H}_1 is accepted. This motivates the introduction of the p -value.

Definition 2. Suppose that, for every α , we have a *size- α rule* ϕ_α :

$$\text{rule } \phi_\alpha: \quad \mathcal{X}_{0,\alpha} = \{\mathbf{x} : \phi(\mathbf{x}) = 0\}, \quad \mathcal{X}_{1,\alpha} = \{\mathbf{x} : \phi(\mathbf{x}) = 1\} \quad (2)$$

meaning that,

$$\alpha = \max_{\theta \in \text{sp}_\Theta(0)} \Pr_{X|\theta} \{x \in \mathcal{X}_{1,\alpha} | \theta\}.$$

Then, the p -value for this test is the smallest size α for which we can declare \mathcal{H}_1 :

$$p\text{-value} = \inf \{\alpha : \mathbf{x} \in \mathcal{X}_{1,\alpha}\}.$$

Informally, the p -value is a measure of evidence for supporting \mathcal{H}_1 . For example, p -values less than 0.01 are considered *very strong evidence* supporting \mathcal{H}_1 .

There are many misconceptions and warnings regarding p -values. Here are the most important ones.

Warning: A large p -value is *not* strong evidence in favor of \mathcal{H}_0 ; a large p -value can occur for two reasons:

- (i) \mathcal{H}_0 is true or
- (ii) \mathcal{H}_0 is false but the test has low detection probability (power).

Warning: *Do not* confuse the p -value with

$$\Pr_{\Theta|X}\{\theta \in \Theta_0 | x\}$$

used in Bayesian inference. **The p -value is *not* the probability that \mathcal{H}_0 is true.**

Theorem 1. *Suppose that the size- α test is of the form:*

$$\text{declare } \mathcal{H}_1 \text{ if and only if } T(x) \geq c_\alpha.$$

Then, the p -value for this test is

$$p\text{-value} = \max_{\theta \in \text{sp}_\Theta(0)} \Pr_{X|\Theta}\{T(X) \geq T(x) | \theta\} \quad (3)$$

where x is the observed value of X . For $\Theta_0 = \{\theta_0\}$:

$$p\text{-value} = \Pr_{X|\Theta} \{T(X) \geq T(x) | \theta_0\}.$$

In words, Theorem 1 states that

The p -value is the probability that, under \mathcal{H}_0 , a random measurement realization \mathbf{X} is observed yielding a value of the test statistic $T(\mathbf{X})$ that is greater than or equal to what has actually been observed, which is $T(\mathbf{x})$.

Note: This interpretation requires that we allow the experiment to be repeated many times. This is what Bayesians criticize by saying that “data that have never been observed are used for inference.”

Theorem 2. *If the test statistic has a continuous distribution, then, under the simple null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ ($\text{sp}_{\Theta}(0) = \{\theta_0\}$), the p -value has the uniform $U(0, 1)$ distribution. Therefore, if we reject \mathcal{H}_0 when the p -value is less than or equal to α , the probability of false alarm is α .*

In other words, if \mathcal{H}_0 is true and if the conditions of Theorem 2 hold, the p -value is like a random draw from the uniform $U(0, 1)$ distribution. If \mathcal{H}_1 is true and if we repeat the experiment many times, the random p -values will concentrate closer to zero.

Example 1: Detecting a DC Level

Consider the following composite hypothesis-testing problem:

$$\mathcal{H}_0 : \quad \theta = \theta_0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(0) = \{\theta_0\} \quad \text{versus}$$

$$\mathcal{H}_1 : \quad \theta > \theta_0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(1) = (\theta_0, +\infty)$$

where the measurements $X[0], X[1], \dots, X[N - 1]$ are conditionally independent, identically distributed (i.i.d.) given $\Theta = \theta$, modeled as

$$\{X[n] \mid \Theta = \theta\} = \theta + W[n] \quad n = 0, 1, \dots, N - 1$$

with $W[n]$ a zero-mean white Gaussian noise with known variance σ^2 , i.e.

$$W[n] \sim \mathcal{N}(0, \sigma^2)$$

implying

$$f_{\mathbf{X} \mid \Theta}(\mathbf{x} \mid \theta) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \theta)^2 \right] \quad (4)$$

where $\mathbf{x} = [x[0], x[1], \dots, x[N - 1]]^T$. A uniformly most powerful (UMP) test can be easily derived:

$$\phi(\mathbf{x}) : \quad T(\mathbf{x}) = \frac{\bar{x} - \theta_0}{\sigma/\sqrt{N}} \stackrel{\mathcal{H}_1}{\geq} \tau$$

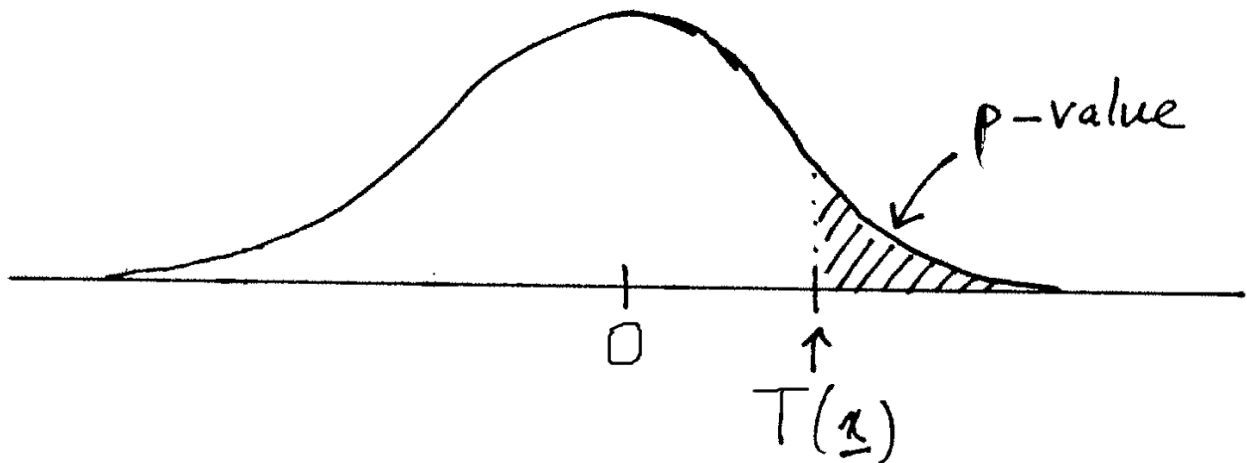
where

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x[n].$$

Under \mathcal{H}_0 , $T(\mathbf{X}) | \{\Theta = \theta_0\}$ is a standard normal random variable.

Under \mathcal{H}_0 , the probability of $T(\mathbf{X})$ being more *extreme* than $T(\mathbf{x})$ is

p -value =



Since the conditions of Theorem 2 hold, we also know that the p -values are uniform $U(0, 1)$ under \mathcal{H}_0 .

If we wish a hard decision for size α , we can make it based on the p -value instead of $T(\mathbf{x})$:

Example 1': Detecting a DC Level

Consider now the following hypothesis-testing problem under the same measurement model as in Example 1:

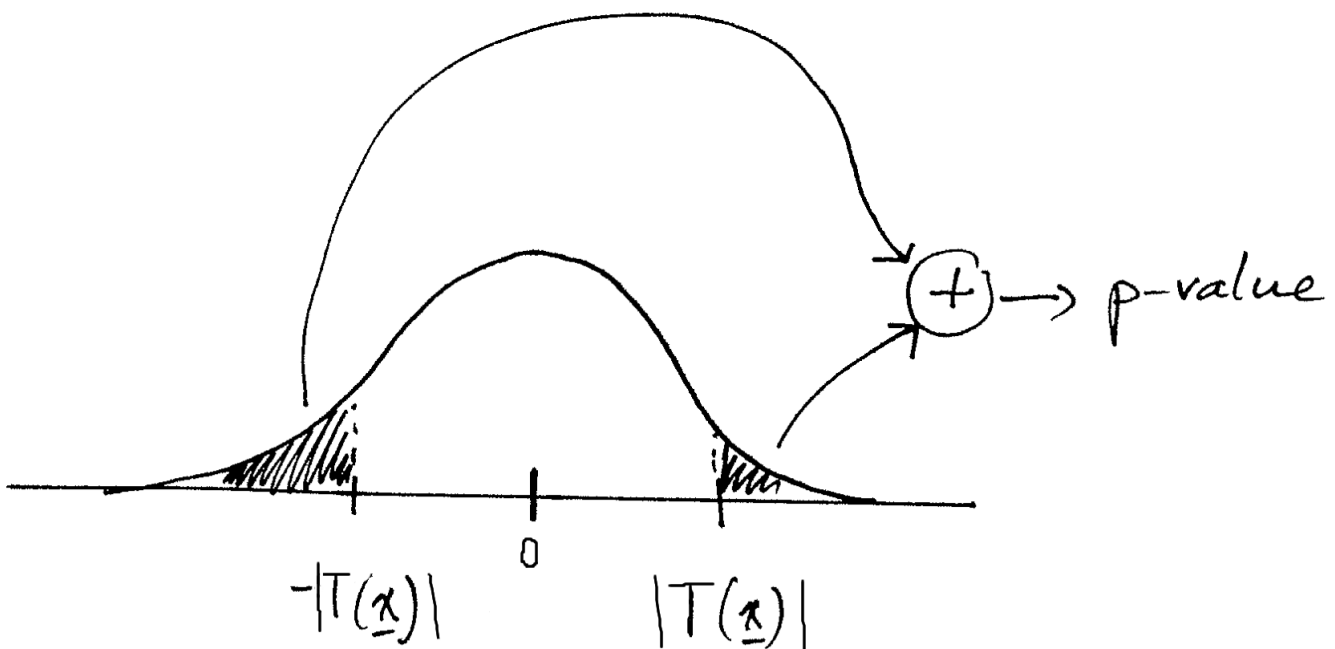
$$\mathcal{H}_0 : \quad \theta = \theta_0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(0) = \{\theta_0\} \quad \text{versus}$$

$$\mathcal{H}_1 : \quad \theta \neq \theta_0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(1) = (-\infty, +\infty) \setminus \{\theta_0\}.$$

The generalized likelihood ratio (GLR) test for this problem can be easily derived:

$$\phi_{\text{GLR}}(\mathbf{x}) : \quad |T(\mathbf{x})| = \left| \frac{\bar{x} - \theta_0}{\sigma/\sqrt{N}} \right| \stackrel{\mathcal{H}_1}{\geq} \tau.$$

Under \mathcal{H}_0 , $T(\mathbf{X}) | \{\Theta = \theta_0\}$ is a standard normal random variable, but now *extreme* has a new meaning:



Under \mathcal{H}_0 , the probability of $|T(\mathbf{X})|$ being more *extreme* than $|T(\mathbf{x})|$ is

p -value =

Since the conditions of Theorem 2 hold, we also know that, under \mathcal{H}_0 , the p -values are uniform $U(0, 1)$.

Example 2: Detecting a Positive DC Level in AWGN (versus nonnegative DC level)

Consider the following composite hypothesis-testing problem:

$$\mathcal{H}_0 : \quad \theta \leq 0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(0) = (-\infty, 0] \quad \text{versus}$$

$$\mathcal{H}_1 : \quad \theta > 0 \quad \text{i.e. } \theta \in \text{sp}_{\Theta}(1) = (0, +\infty)$$

where the measurements $X[0], X[1], \dots, X[N - 1]$ are conditionally i.i.d. given $\Theta = \theta$, modeled as

$$\{X[n] \mid \Theta = \theta\} = \theta + W[n] \quad n = 0, 1, \dots, N - 1$$

with $W[n]$ a zero-mean white Gaussian noise with known variance σ^2 , i.e.

$$W[n] \sim \mathcal{N}(0, \sigma^2)$$

implying

$$f_{\mathbf{X} \mid \Theta}(\mathbf{x} \mid \theta) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \theta)^2 \right] \quad (5)$$

where $\mathbf{x} = [x[0], x[1], \dots, x[N - 1]]^T$. A sufficient statistic for θ is

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x[n].$$

and

$$f_{\bar{X}|\Theta}(\bar{x}|\theta) = \mathcal{N}(\bar{x}|\theta, \sigma^2/N). \quad (6)$$

We start by writing the classical Neyman-Pearson test for the simple hypotheses with $\text{sp}_{\Theta}^{\text{simple}}(0) = \{\theta_0\}$ and $\text{sp}_{\Theta}^{\text{simple}}(1) = \{\theta_1\}$, where $\theta_0 \in (-\infty, 0]$ and $\theta_1 \in (0, +\infty)$, implying

$$\frac{f_{\bar{X}|\Theta}(\bar{x}|\theta_1)}{f_{\bar{X}|\Theta}(\bar{x}|\theta_0)} = \frac{(2\pi\sigma^2/N)^{-1/2} \cdot \exp[-\frac{1}{2\sigma^2/N}(\bar{x} - \theta_1)^2]}{(2\pi\sigma^2/N)^{-1/2} \cdot \exp[-\frac{1}{2\sigma^2/N}(\bar{x} - \theta_0)^2]} \stackrel{\mathcal{H}_1}{\geq} \lambda$$

and

$$\theta_0 < \theta_1.$$

Taking log etc. leads to

$$(\theta_1 - \theta_0)\bar{x} \stackrel{\mathcal{H}_1}{\geq} \eta$$

and, since $\theta_0 < \theta_1$, to

$$\phi(\mathbf{x}) : \quad \bar{x} \stackrel{\mathcal{H}_1}{\geq} \tau.$$

Hence, we transformed our likelihood ratio in such a way that θ_0 and θ_1 disappear from the test statistic, i.e. we accomplished **(1)** above.

The power function of this test is

$$\Pr_{\mathbf{X}|\Theta}\{\bar{X} > \tau | \theta\} = \Pr_{\mathbf{X}|\Theta}\left\{\frac{\bar{X} - \theta}{\sigma/\sqrt{N}} > \frac{\tau - \theta}{\sigma/\sqrt{N}} \mid \theta\right\} = Q\left(\frac{\tau - \theta}{\sigma/\sqrt{N}}\right)$$

which is an increasing function of θ . Recall the definition (1) of test size:

$$\begin{aligned} \max_{\theta \in \text{sp}_{\Theta}(0)} P_{\text{FA}}(\phi(\mathbf{X}), \theta) &= \max_{\theta \in \text{sp}_{\Theta}(0)} \Pr_{X|\Theta} \{\bar{X} > \tau \mid \theta\} \\ &= \max_{\theta \in (-\infty, 0]} Q\left(\frac{\tau - \theta}{\sigma/\sqrt{N}}\right) = Q\left(\frac{\tau}{\sigma/\sqrt{N}}\right). \end{aligned}$$

The most powerful test is achieved if the test size α is reached by equality:

$$\tau = \frac{\sigma}{\sqrt{N}} Q^{-1}(\alpha).$$

Hence, we have accomplished (2), since this τ yields exactly the size α for our test $\phi(\mathbf{X})$.

What is the p -value for the observed x and test statistic \bar{x} ? Recall (3):

$$p\text{-value} = \max_{\theta \in \text{sp}_{\Theta}(0)} \Pr_{X|\Theta} \{\bar{X} \geq \bar{x} \mid \theta\} = Q\left(\frac{\bar{x}}{\sigma/\sqrt{N}}\right).$$

In this case, the conditions of Theorem 2 do not hold and the p -values are not necessarily uniform $U(0, 1)$ under \mathcal{H}_0 .

Multiple Testing

We conduct many hypothesis tests in some applications, e.g.

- bioinformatics and
- sensor networks.

Here, we perform *many* typically binary *tests*, say one test per node in a sensor network. This is different from *testing multiple hypotheses* that we considered in handout # 5b, where we performed *a single test of multiple hypotheses*.

An Example

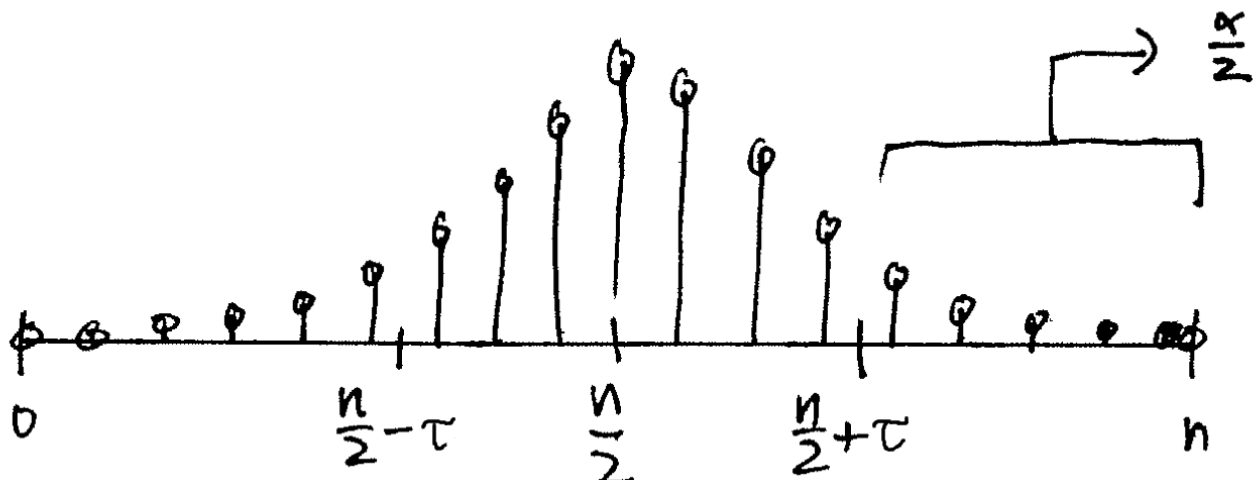
Suppose that we wish to decide whether a certain coin is fair:

$$\mathcal{H}_0 : \quad \theta = \frac{1}{2} \quad \text{i.e. } \theta \in \text{sp}_\theta(0) = \{\theta_0\} \quad \text{versus}$$

$$\mathcal{H}_1 : \quad \theta \neq \frac{1}{2} \quad \text{i.e. } \theta \in \text{sp}_\theta(1) = \text{sp}_\theta(1) = (-\infty, +\infty) \setminus \{\frac{1}{2}\}$$

where the parameters θ is the probability of heads. You flip the coin n times and observe the number x of heads. A natural test is

$$\left| x - \frac{n}{2} \right| \stackrel{\mathcal{H}_1}{\geq} \tau.$$



To ensure size α , choose the threshold τ such that

$$2 \sum_{k=\lceil \frac{n}{2} + \eta \rceil}^n \binom{n}{k} \left(\frac{1}{2}\right)^n \approx \alpha.$$

Now, suppose that you have M different coins that you wish to test.

If you perform that test that we just discussed, you will 'discover' about $M \cdot \alpha$ unfair coins even if all the coins are fair.

Question: If the number of coins is $M = 1000$, $\alpha = 0.05$, and we discovered 50 unfair coins, would you really believe these coins were unfair?

Clearly, we need an alternative notion of 'size.' A common choice is family-wise error rate (FWER):

$$\text{FWER} = \Pr\{ \geq 1 \mathcal{H}_1 \text{ accepted} \mid \text{all } \mathcal{H}_0 \text{ true} \}.$$

Sidak Correction

Assume

- simple null hypotheses:

$$\mathcal{H}_{0,i} : \theta_i = \theta_{0,i} \quad i = 1, 2, \dots, M$$

- each member i (of the 'family') conducts a test with size α based on its measurement X_i ,
- the measurements $X_i, i = 1, 2, \dots, M$ are conditionally independent given θ_i .

Denote by A the event that there is at least one false alarm and by A_i the event that the i th member is falsely alarmed. Then

$$A = \cup_{i=1}^M A_i$$

and A_i are independent given θ_i , implying

$$\begin{aligned}\text{FWER} &= \Pr\{A \mid \text{all } \mathcal{H}_0 \text{ true}\} \\ &= 1 - \Pr\{A^c \mid \text{all } \mathcal{H}_0 \text{ true}\} \\ &= 1 - \Pr\{\cap_{i=1}^M A_i^c \mid \text{all } \mathcal{H}_0 \text{ true}\} \\ &\stackrel{\text{cond. indep.}}{=} 1 - \prod_{i=1}^M \Pr_{X_i \mid \theta_{0,i}}\{A_i^c \mid \theta_{0,i}\} \\ &= 1 - (1 - \alpha)^M.\end{aligned}$$

Thus, to achieve $\text{FWER} = \alpha'$, set

$$\alpha = 1 - (1 - \alpha')^{1/M}$$

in each individual test.

Note: If α is small and M not too large,

$$1 - (1 - \alpha)^M \approx M \alpha.$$

Bonferroni Correction

Assume

- simple null hypotheses:

$$\mathcal{H}_{0,i} : \theta_i = \theta_{0,i} \quad i = 1, 2, \dots, M$$

- each member i (of the 'family') conducts a test with size α based on its measurement X_i .

If A_i are not conditionally independent given θ_i , then the union-bound inequality implies

$$\begin{aligned} \text{FWER} &= \Pr\{A \mid \text{all } \mathcal{H}_0 \text{ true}\} \\ &= \Pr\{\cup_{i=1}^M A_i \mid \text{all } \mathcal{H}_0 \text{ true}\} \\ &\leq \sum_{i=1}^M \underbrace{\Pr_{X_i \mid \theta_{0,i}}\{A_i \mid \theta_{0,i}\}}_{\alpha} = M \alpha. \end{aligned}$$

Therefore,

$$\alpha = \frac{\alpha'}{M}$$

leads to

$$\text{FWER} \leq \alpha'.$$

This adjustment is more conservative than the Sidac correction, but is also more general, since it does not require conditionally independent hypotheses.

False Discovery Rate (FDR)

Assume

- simple null hypotheses:

$$\mathcal{H}_{0,i} : \theta_i = \theta_{0,i} \quad i = 1, 2, \dots, M$$

- each member i (of the 'family') conducts a test with size α based on its measurement X_i .

FWER is too conservative in some applications. Perhaps we can allow a few false alarms if, by doing that, we can significantly increase the number of correct decisions.

This led Benjamini and Hochberg to introduce *false discovery rate (FDR)*:

$$\text{FDR} = \text{E} [\text{FDP}]$$

where FDP is *false discovery proportion*, defined as as

$$\text{FDP} = \begin{cases} \text{FD}/D, & D > 0, \\ 0, & D = 0 \end{cases}$$

where D is the number of 'discoveries' (i.e. cases where $\mathcal{H}_{1,i}$ is accepted) and FD is the number of 'false discoveries' (i.e. cases where $\mathcal{H}_{1,i}$ is incorrectly accepted).

Benjamini and Hochberg showed how to ensure

$$\text{FDR} \leq \alpha''.$$

The Benjamini-Hochberg (BH) Method

(i) Denote the ordered p -values by $p_{(1)} < p_{(2)} < \dots < p_{(m)}$.

(ii) Define

$$l_i = \frac{i \alpha''}{C_m m} \quad \text{and} \quad D = \max\{i : p_{(i)} < l_i\}$$

where C_m is defined to be 1 if the p -values p_i are conditionally independent given θ_i and $C_m = \sum_{i=1}^m (1/i)$ otherwise.

(iii) Define the *BH rejection threshold* $\tau = p_{(D)}$.

(iv) Accept all $\mathcal{H}_{1,i}$ for which $p_i \leq \tau$.

Theorem 3. *(formulated and proved by Benjamini and Hochberg) If the above BH method is applied, then, regardless of how many null hypotheses are true and regardless of the distribution of the p -values when the null hypothesis is false,*

$$FDR = E[FDP] \leq \alpha''.$$

Hence, for conditionally independent p -values given θ_i , we have

