MATH 189
Multiple Testing: FWER and FDR

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

Outline

- Last week we considered hypothesis testing problems on univariate population mean.
  - Variance
  - Distribution: normal, non-normal, CLT
  - Hypothesis testing: error types, decision rules

- This week we extend our study to the problem of testing many means simultaneously.
  - Multiple testing
  - Type I error control: FWER control, FDR control
  - Hotelling's T^2

Hypothesis Testing for Multivariate Mean

- Suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are independently sampled from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
- We would like to test whether the unknown population mean vector $\boldsymbol{\mu}$ equals to a specific vector, say $\boldsymbol{\mu}_0$.

- It is natural to consider the following null and alternative hypotheses:

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad \text{versus} \qquad H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0.$$

Other Expressions of Hypotheses

- We can also express the null and alternative hypotheses as

$$H_0: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_1^0 \\ \mu_2^0 \\ \vdots \\ \mu_p^0 \end{pmatrix} \quad \text{versus} \quad H_1: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_1^0 \\ \mu_2^0 \\ \vdots \\ \mu_p^0 \end{pmatrix},$$

or, equivalently,

$$H_0: \mu_1 = \mu_1^0, \dots, \mu_p = \mu_p^0 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_1^0, \dots, \mu_p \neq \mu_p^0.$$

Testing a mean vector is equivalent to testing multiple means together!

A Naive Scheme to Test Multivariate Mean

- Following the univariate case, a naive scheme for testing a multivariate hypothesis is to compute the t-statistic for each univariate mean.
- Define T_j, the t-statistic for the j-th variable, as

$$T_j = \frac{\bar{x}_j - \mu_j^0}{s_j/\sqrt{n}} \sim t_{n-1},\ j = 1, \dots, p.$$

- Given a significance level α, we reject $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ if $|T_j| > t_{n-1,\alpha/2}$ for at least one $j \in \{1, \dots, p\}$.

Problem with this Naive Scheme

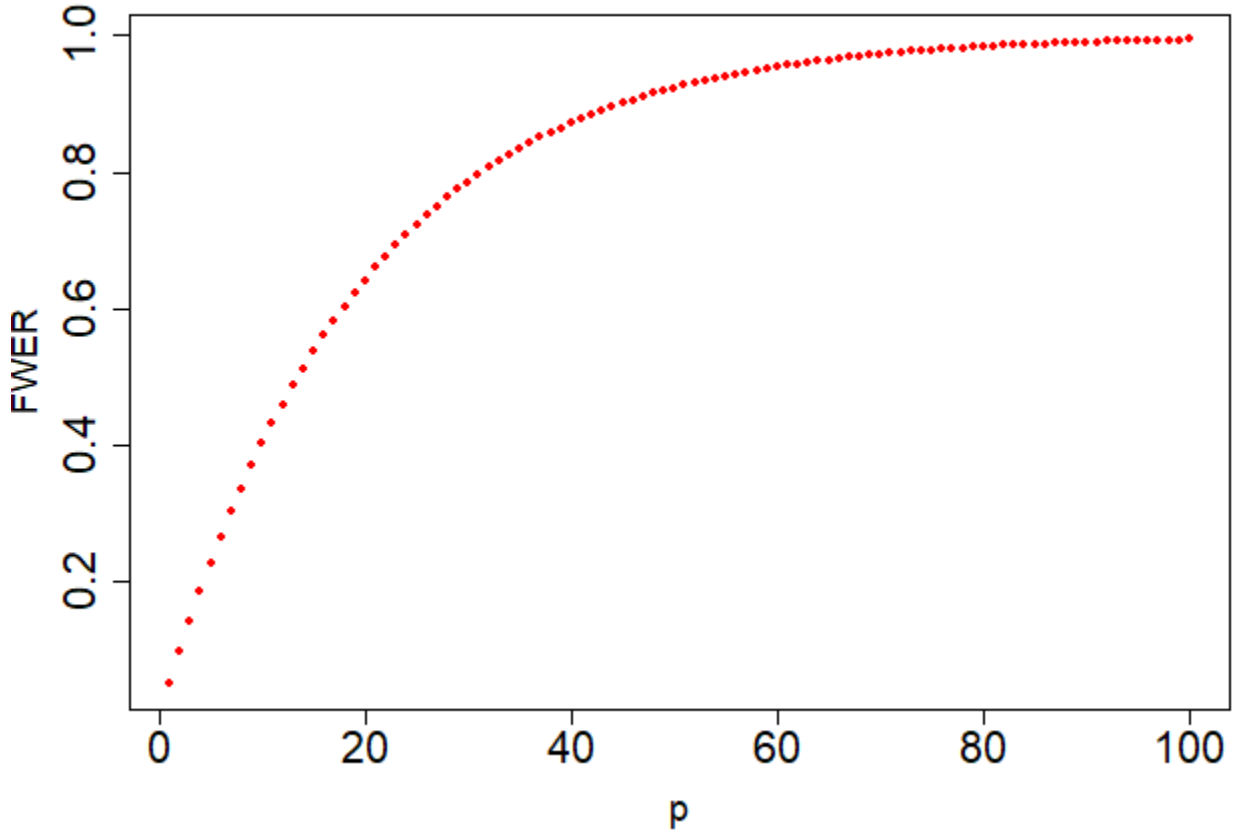
- The biggest issue with this naive scheme is that it does not control the *family-wise error rate*.
- The *family-wise error rate* is the probability of rejecting at least one of the null hypotheses $H_0^{(j)}: \mu_j = \mu_j^0$ when all nulls are true.
- If the p test statistics are independent, rejecting each null hypothesis at significance level α will result in a *family-wise error rate*

$$\text{FWER} = 1 - (1 - \alpha)^p \geq \alpha.$$

- The probability of falsely rejecting at least one null hypothesis is much higher than α if the number of variables is large.

Family-Wise Error Rate versus Dimension

p	FWER
1	0.0500
2	0.0975
3	0.1426
4	0.1854
5	0.2262
10	0.4012
20	0.6415
50	0.9230
100	0.9940



$$\text{FWER} = 1 - (1 - \alpha)^p \geq \alpha.$$

p independent tests, each with significant level $\alpha = 0.05$

FWER Correction

- The naive scheme fails to control the *family-wise error rate* and hence yields a liberal test. That is, we tend to reject null hypotheses more often than we should.
- High *family-wise error rate* means we have a high chance to commit a Type I error: reject at least one null hypothesis when they are all true.
- To control the *family-wise error rate*, we need to correct the naive scheme. If we want to control the *family-wise error rate* at level α, we should set the significance level for each individual test smaller than α.

FWER Correction: Two Approaches

- Many procedures have been developed to control the *family-wise error rate* which is the probability of at least one type I error occurs.

- Two general types of FWER corrections:
  1. **Single step**: equivalent adjustment made to each p-value.

     Bonferroni Correction
  1. **Sequential**: adaptive adjustment made to each p-value.

     Holm's Method

Bonferroni Correction

- A simple yet popular FWER correction is the Bonferroni Correction which is named after Italian mathematician Carlo Emilio Bonferroni for its use of Bonferroni inequalities.
- Suppose we have m null hypotheses (e.g. a mean vector of m variables). Denote $p_1, \ldots, p_m$ the corresponding p-values.
- To control the FWER at level α, the Bonferroni Correction sets the significance level for each individual test at α/m instead of α.
- The intuition of Bonferroni Correction is as follows:

$$\text{FWER} = \mathbb{P}\left\{\bigcup_{i=1}^{m}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m}\left\{\mathbb{P}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m\frac{\alpha}{m} = \alpha.$$

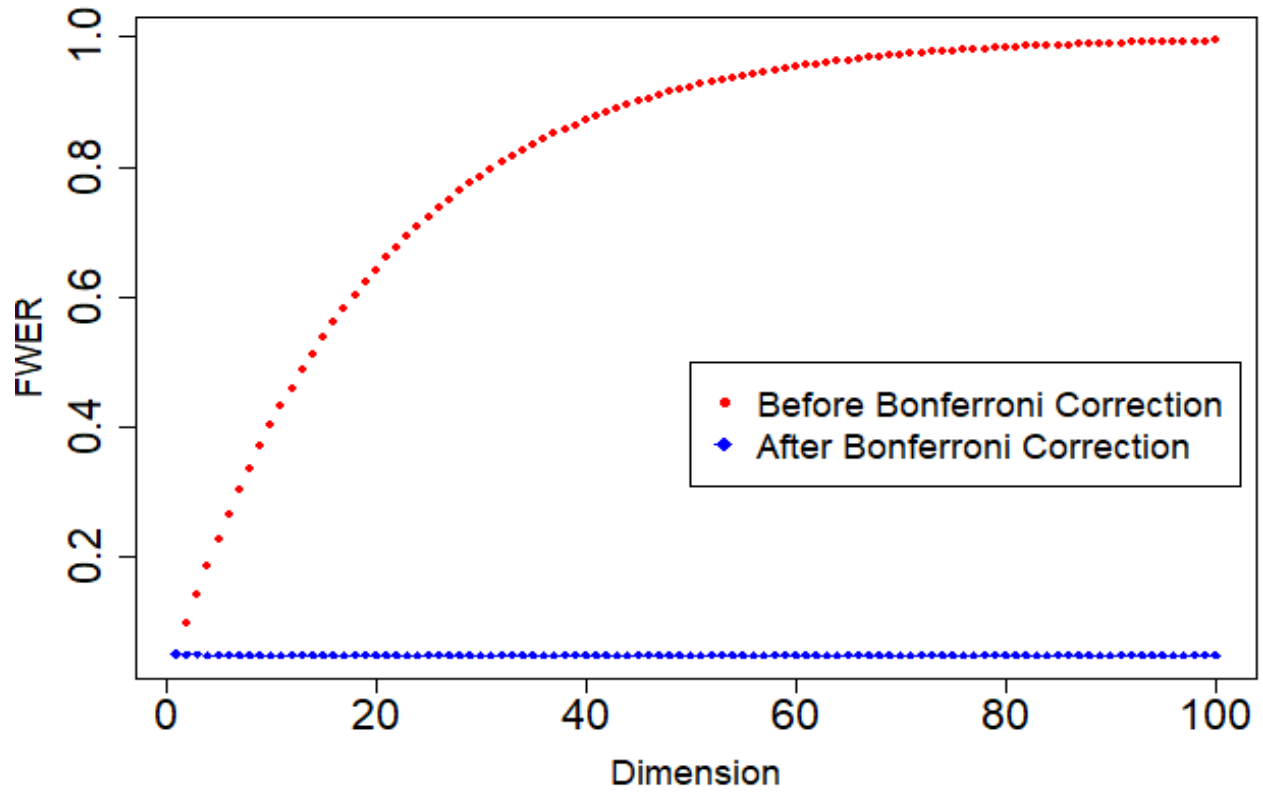
Bonferroni Correction (cont.)

- For independent tests, the Bonferroni Correction yields a family-wise error rate of approximately α.

m	FWER (Naive Scheme)	FWER (Bonferroni Correction)
1	0.0500	0.0500
2	0.0975	0.0494
3	0.1426	0.0492
4	0.1854	0.0491
5	0.2262	0.0490
10	0.4012	0.0489
20	0.6415	0.0488
50	0.9230	0.0488
100	0.9940	0.0488

Bonferroni Correction (cont.)

- For independent tests, the Bonferroni Correction yields a family-wise error rate of approximately α.



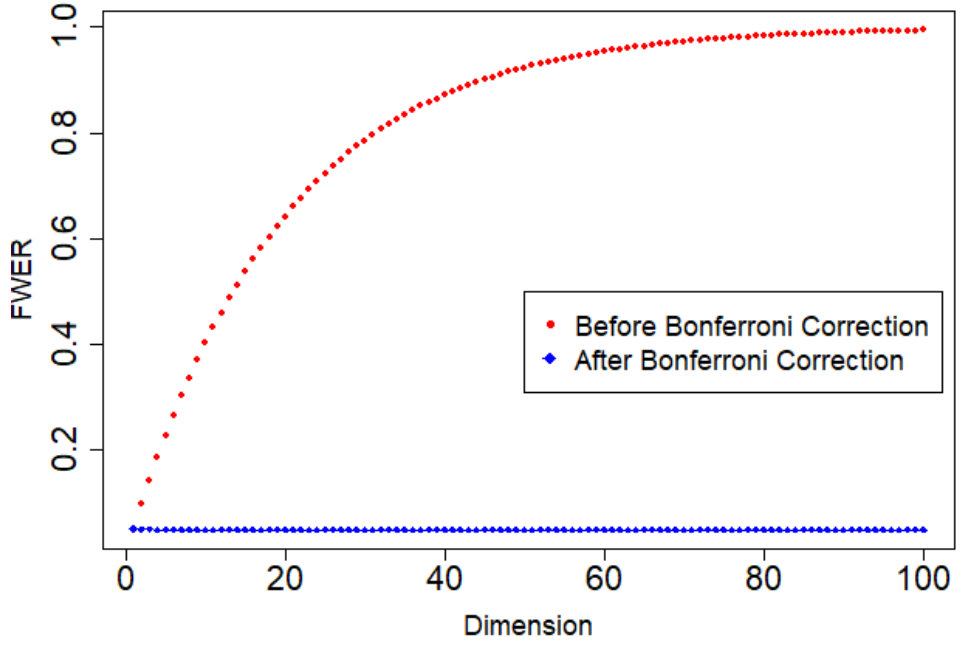
Pros and Cons of Bonferroni Correction

Pros

- Control FWER at a given level α
- Protects Type I errors
- Easy to implement

Cons

- Need the assumption of independence
- Conservative when there is strong dependence among variables
- Vulnerable to Type II errors (failing to reject the null hypothesis when you should)



Holm's Method

- Holm's Method is a simple sequential method to correct the FWER.
- Suppose we have m null hypotheses (e.g. a mean vector of m variables). Denote $p_{(1)}, \ldots, p_{(m)}$ the corresponding p-values in ascending order.

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

- To control the FWER at level α, we adjust the significance level for the hypothesis testing with the j-th smallest p-value by

$$\alpha^* = \alpha/(m - j + 1).$$

- For example, when $m = 100$, we reject the j-th hypothesis if

$$p_{(1)} \leq \frac{\alpha}{100}, p_{(2)} \leq \frac{\alpha}{99}, \ldots, p_{(j)} \leq \frac{\alpha}{m-j+1}, \ldots, p_{(100)} \leq \alpha.$$

Holm's Method (cont.)

- Holm's Method downscales the significance level α (or upscales the p-value) by a factor depending on the order of the p-value.
- The Bonferroni correction rejects null hypotheses with p-value less than α/m, in order to control FWER at α. The cost of this protection against type I errors is an increased risk of accepting one or more false null hypotheses (i.e. of committing one or more type II errors).
- The Holm's method also controls the maximum family-wise error rate at α, but with a lower increase of type II error risk.

Bonferroni vs Holm

Pros

- One-time adjustment
- Uniform significance level

Cons

- Low power (high type II error rate)

Pros

- Sequential adjustment
- Adaptive significance level
- Uniformly higher power (less type II error)

Cons

- Hard to define confidence interval
- No guarantee on type II error control (better than Bonferroni though)

Issues about Controlling FWER

- FWER is appropriate when you want to guard against ANY type I error.
- When we control the FWER, the testing procedure need "very strong evidence" to reject a null hypothesis. As a result, the testing procedure may fail to rejects some hypothesis when the null hypothesis is false.
- We gain a strict control of type I error at a price of loose control of type II error.
- FWER control is important when the cost of type I error is much higher than type II error.

Example: FDA Tests on Drug Side Effects

- Before a new drug is proved by FDA, they need to test if the drug has severe side effects to patients. The test is based on the following null and alternative hypotheses:

$$H_0: \text{The drug has severe side effects.}$$

$$H_1: \text{The drug does not have severe side effects.}$$

In this case, the type I and type II errors stand for:

- Type I error: Claim a drug does not have severe side effects when it has.
- Type II error: Claim a drug has severe side effects when it does not have.

Example: FDA Tests on Drug Side Effects

H_0: The drug has severe side effects.
H_1: The drug does not have severe side effects.

	Type I Error	Type II Error
Decision	Claim a drug does not have severe side effects when it has	Claim a drug has severe side effects when it does not have.
Costs	The drug is approved by FDA. Many patients will suffer from the side effects. People may lose their lives. Financial cost to the pharmaceutical company due to potential lawsuits.	The drug is not approved by FDA. Financial cost to the pharmaceutical company due to more future research.

Type I error is more important in this case!

Example: Genotype and Disease

- In genomic studies, it is often of interest to find the genotypes that have a significant association with a disease of interest.
- Usually, scientists collect a pool of millions of genotypes and data analysts want to screen out irrelevant genotypes before conducting a finer research.
- We need to carry out a large amount of testing problems.
- For each genotype, the null and alternative hypotheses are:

  H_0: The association is zero.

  H_1: The association is not zero.

Example: FDA Tests on Drug Side Effects

H_0: The association is zero.
H_1: The association is not zero.

	Type I Error	Type II Error
Decision	Claim an association when it does not exist.	Ignore an association when it exists.
Costs	We include some non-significant genotypes which may be filtered out in next steps.	We missed an significant genotype which will be ignored in future studies.

Type II error is more important in this case!

What if We Can Live With Some Type I Errors?

- However, in many cases, we could tolerate a certain amount (or percentage) of type I errors (e.g. the genotype example).
- Instead of FWER, a more relevant quantity we may want to control is the false discovery rate (FDR).
- FDR control will lead to procedures which is less conservative and more powerful than FWER control.
- FDR control is still a very active area in statistical research.

False Discovery Rate

- Suppose we have m null hypotheses, among which m_0 are true (we do not know this in practice!)
- The test results can be summarized in the following table:

Decision	H_0 is True	H_0 is False	Total Number
Do not reject $\boldsymbol{H_0}$	U	T (type II)	$m - R$
Reject $\boldsymbol{H_0}$	V (type I)	S	R
Total Number	m_0	$m - m_0$	m

Decision	H_0 is True	H_0 is False	Total Number
Do not reject $\boldsymbol{H_0}$	U	T	$m - R$
Reject $\boldsymbol{H_0}$	V	S	R
Total Number	m_0	$m - m_0$	m

- False discovery propitiation (FDP) is defined as the number of false rejection (V) and total rejection (R).

$$\mathrm{FDP} = \frac{V}{R}, \text{ when } R > 0.$$

- False discovery rate (FDR) is defined as the expected value of FDP

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}) = \mathbb{E}\left(\frac{V}{R}\right), \text{ when } R > 0.$$

- Family-wise error rate (FWER) can be expressed as

$$\mathrm{FWER} = \mathbb{P}(V \geq 1).$$

Why Controlling FDR is Important?

- While the FWER control is appealing, the resulting thresholds often suffer from low power. In practice, this tends to wipe out evidence of the most interesting effects (true nulls can be buried by a large number of non-rejected false nulls).

- FDR control offers a way to increase power while maintaining some principle bound on type I error.

- It is based on the assessment of the chance you observe a type I error:

  4 false discoveries out of 10 rejected null hypotheses is a more serious error than 20 false discoveries out of 100 rejected null hypotheses.

The Benjamini-Hochberg Procedure

- Benjamini and Hochberg (1995) introduced the concept of FDR and proposed a simple procedure to control it. We will call it the BH procedure.
- Suppose we have m null hypotheses: $H_0^{(1)}, \dots, H_0^{(m)}$. Denote $p_{(1)}, \dots, p_{(m)}$ the corresponding p-values in ascending order.

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

- For a given α, find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$, denoted by k^*.
- Reject the null hypotheses $H_0^{(j)}$ with p-value $p_j \leq p_{(k^*)}$.

The Benjamini-Hochberg Procedure (cont.)

- In BH (1995) paper, they proved (for independent tests), the BH procedure controls the FDR at the given level α by showing

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}) = \mathbb{E}\left(\frac{V}{R}\right) \leq \frac{\mathrm{m_0}}{m}\alpha \leq \alpha.$$

- When the tests are dependent, BH procedure (under certain conditions) can control the FDR at level α up to a factor $\ln(m)$:

$$\mathrm{FDR} \leq \ln(m)\alpha.$$

- The BH procedure can be extended to dependent tests cases with some modifications.

Example: BH procedure on a Simulated Example

- To illustrate how the BH procedure works, we consider the following toy example.
- Suppose we have a dataset of 10 variables. We want to test if the population mean vector $\boldsymbol{\mu}$ equals to a given vector $\boldsymbol{\mu}^0$.
- Null and alternative hypothesis:

$$H_0: \mu_1 = \mu_1^0, \ldots, \mu_{10} = \mu_{10}^0 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_1^0, \ldots, \mu_{10} \neq \mu_{10}^0.$$

- For each argument in the null hypothesis, we can calculate a *t*-statistic and hence a p-value.
- Denote the p-value in ascending order as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(10)}$.

Example: BH procedure on Simulated p-values

Rank (j)	p-value $p_{(j)}$	BH Threshold ($\frac{j}{m}\alpha$)	$p_{(j)} > \frac{j}{m}\alpha$	Reject H_0 ?
1	0.0008	0.005	No	Yes
2	0.0090	0.010	No	Yes
3	0.0148	0.015	No	Yes
4	0.0215	0.020	Yes	No
5	0.0301	0.025	Yes	No
6	0.0450	0.030	Yes	No
7	0.0641	0.035	Yes	No
8	0.0689	0.040	Yes	No
9	0.0722	0.045	Yes	No
10	0.0831	0.050	Yes	No

Set $\alpha = 5\%$. We find $k = 3$ by checking $p_{(k)} \leq \frac{k}{m}\alpha$. Hence, we reject the first three hypotheses.

Example: Visualization of BH Procedure

- We can visualize the BH procedure as follows. First, we plot p_j versus j for $j = 1, \ldots, m$. Then we draw a line through the origin with slope α/m.

- To control the FDR at level α, the BH procedure will reject all the hypotheses with p-values below the line.

- We visualize the BH procedure with $\alpha = 20\%$.

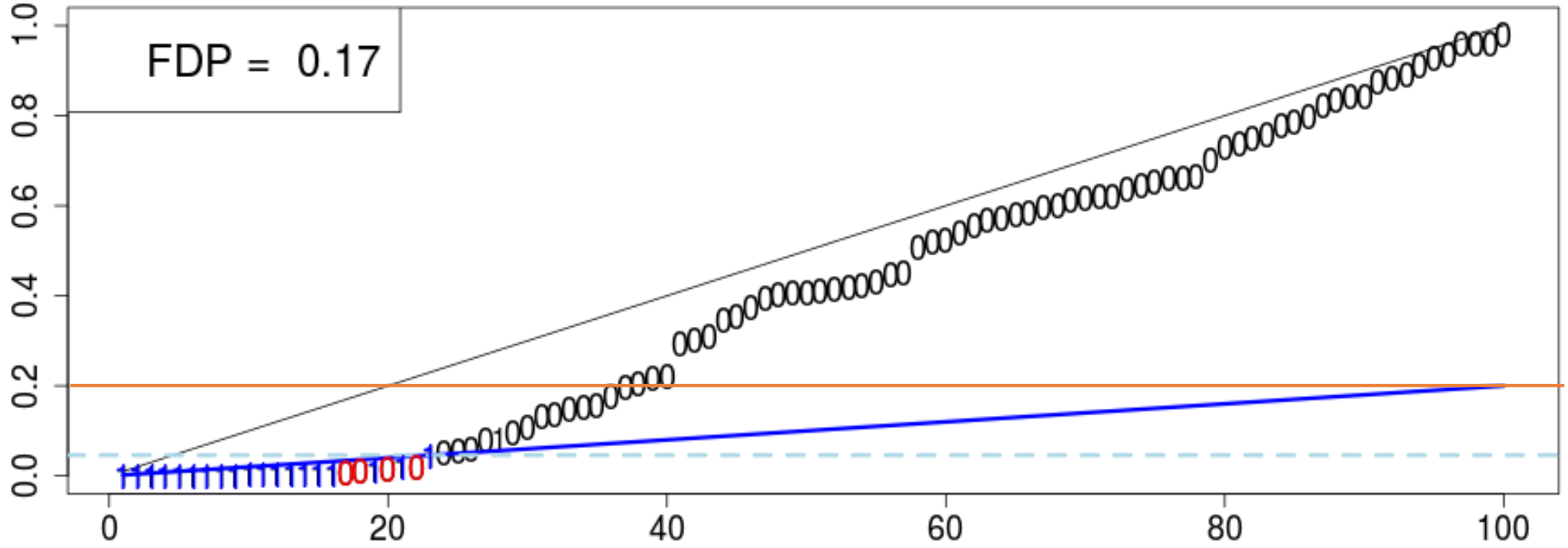
- The data generating process is described in the next slide.

Data Generating Process

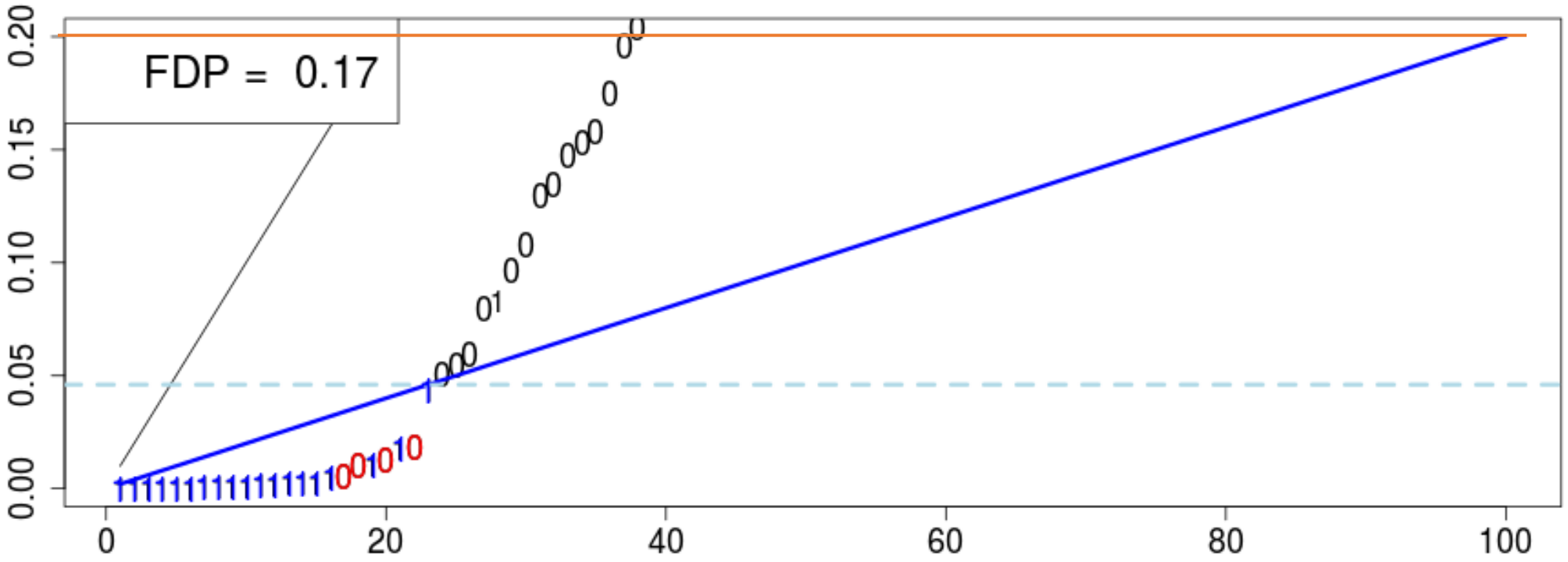
- We generate a simulated dataset of 100 independent normal random variables each with variance 1. We set the mean of first 80 random variables to be 0 and the rest 20 to be 3.
- For each random variable, we sample 100 observations.
- We test the following null and alternative hypothesis:

$$H_0: \mu_1 = 0, \dots, \mu_{100} = 0 \quad \text{vs} \quad H_1: \mu_1 \neq 0, \dots, \mu_{100} \neq 0.$$

- In an oracle scenario, we should accept the first 80 null hypothesis and reject the last 20 null hypotheses.
- To control the FDR at level α (e.g. 20%), we would observe, in average, less than 100α type I errors.



- The blue solid line is the BH threshold with slope $\frac{\alpha}{m}$ and $\alpha = 0.2$.
- The orange solid line is $\alpha = 0.2$ (nominal level)
- Each symbol denote the p-value of a random variable. 0 stands for true null and 1 stands for false null.
- Blue 1 denotes correctly rejected hypothesis, red 0 denotes type I error and black 1 denotes type II error.



- Let's take a closer look!
- The hypotheses with p-values under the blue solid line are rejected.
- We rejected 23 hypotheses: 19 are correct and 4 are type I errors (4 red 0s).
- The false discovery proportion is $\mathrm{FDP} = \frac{4}{23} \approx 0.174$. (This is not FDR!)
- We also have one type II error (the black 1 above the blue solid line).
- Without BH procedure, we will have much more type I errors (0s under the orange line).

FWER control vs FDR control

FWER

- Most strict control of type I error
- Conservative procedure
- Inflated type II errors
- Low power
- Important for applications where no type I error is allowed (e.g. drug tests)

FDR

- Tolerant for some type I error
- Less conservative
- Less type II errors
- Higher power
- Popular in applications where people can live with some type I errors (e.g. genomic studies)

If we consider m tests whose null hypotheses are all true, then FWER = FDR