고급바이오정보학

8강. Gene Discover I: Multiple Testing with FWER Control

출처 : 방송통신대학교 바이오정보·통계학과

목차

1. Genome Data
2. Multiple Testing
3. FWER
4. FDR

1. Genome Data

유전체 데이터의 특징

1) 연속형 데이터 : 유전자 발현 (m ≈ 500 ~ 10,000), CNV (m ≈ 1M)

2) 이분법적인 데이터 : mutation (m ≈ 100,000)

3) 3 레벨 데이터 : GWAS(m ≈ 1/2 ~ 1M), WGS(m >> 1M)

=> 변수의 개수가 많음.

1. Genome Data

임상결과 값

1) 연속형 데이터 : 혈압, BMI 등

2) 이분법적인 데이터 : 질병 여부, 반응 여부, 질병 타입 A/B

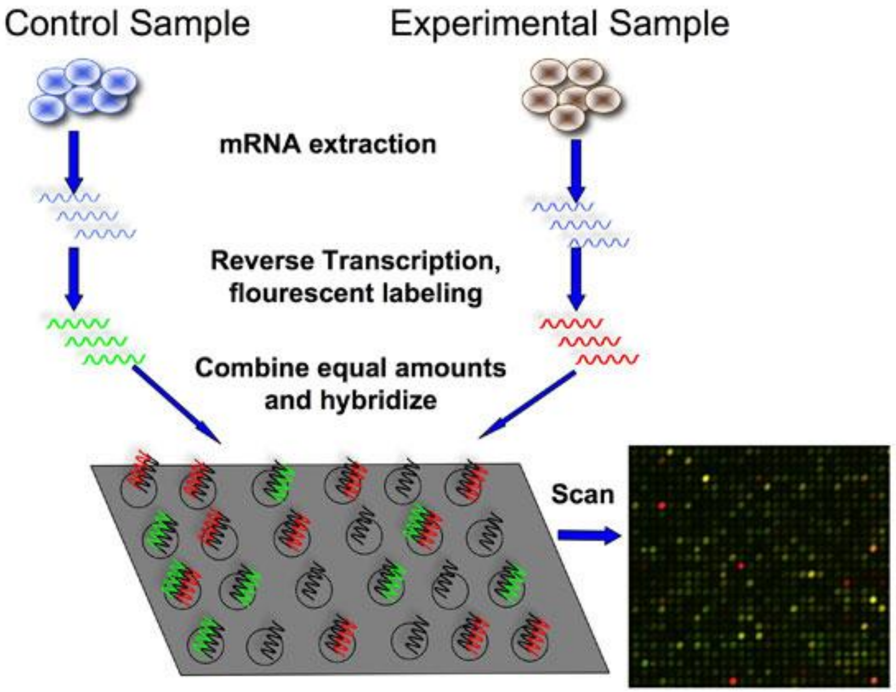
3) 생존 결과 : disease free survival, overall survial

1. Genome Data

Example

- Golub et al. (1999) leukemia study

1. Microarray data for $m = 6810$ genes
2. From $n_{ALL} = 27$ and $n_{AML} = 11$ patients
3. To discover the genes differentially expressed in the two disease groups
4. Discovery procedure involves 6810 t-tests, so that the type I error will be enlarged without a multiple testing adjustment

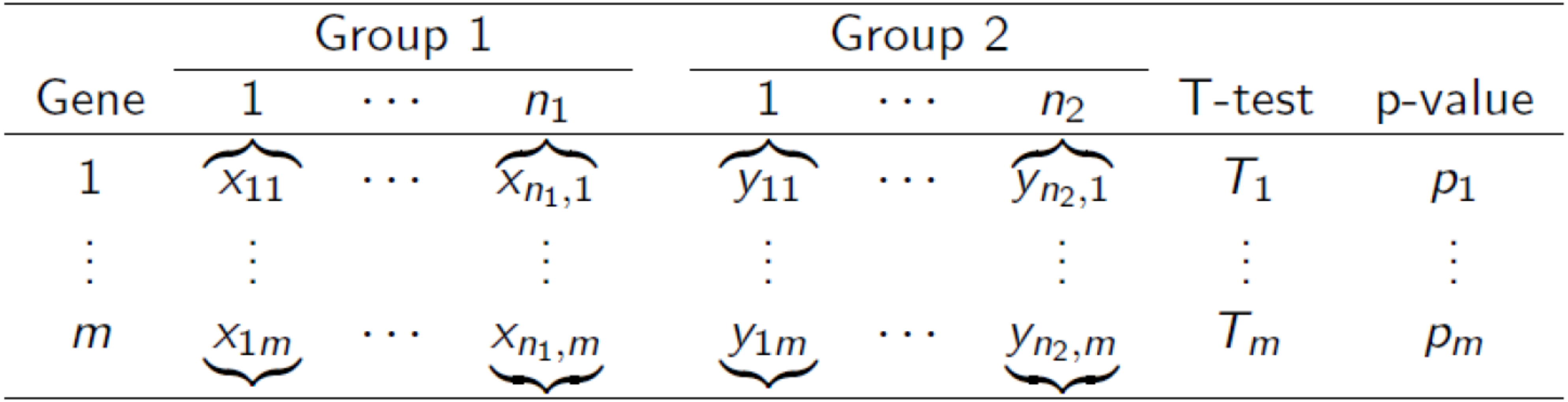

Control Sample
Experimental Sample
mRNA extraction
Reverse Transcription, flourescent labeling
Combine equal amounts and hybridize
Scan

*Acute lymphoblastic leukemia(ALL) :
급성 림프성 백혈병

*Acute myeloid leukemia(AML) :
급성 골수성 백혈병

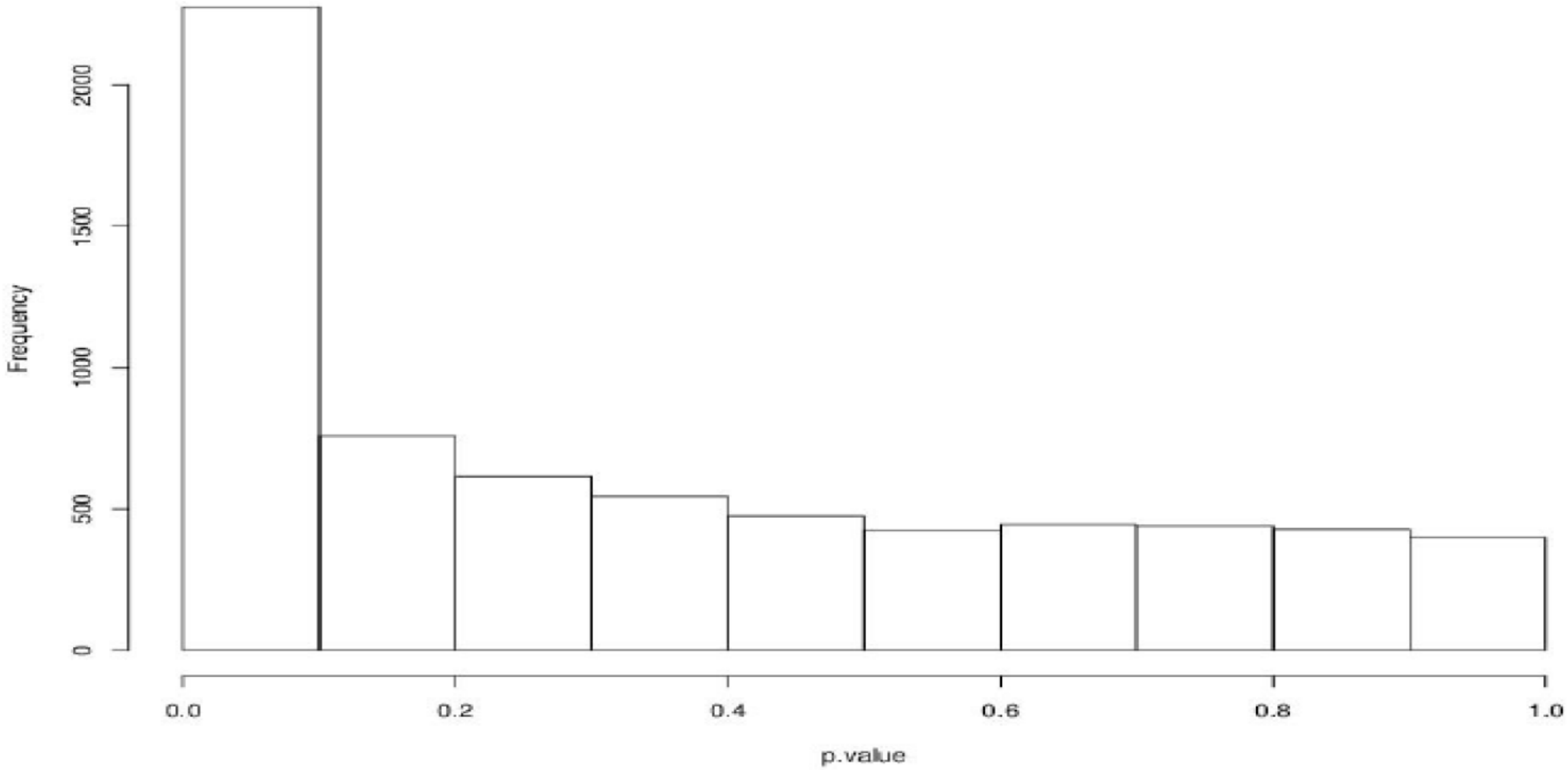
1. Genome Data

- **Microarray Data**



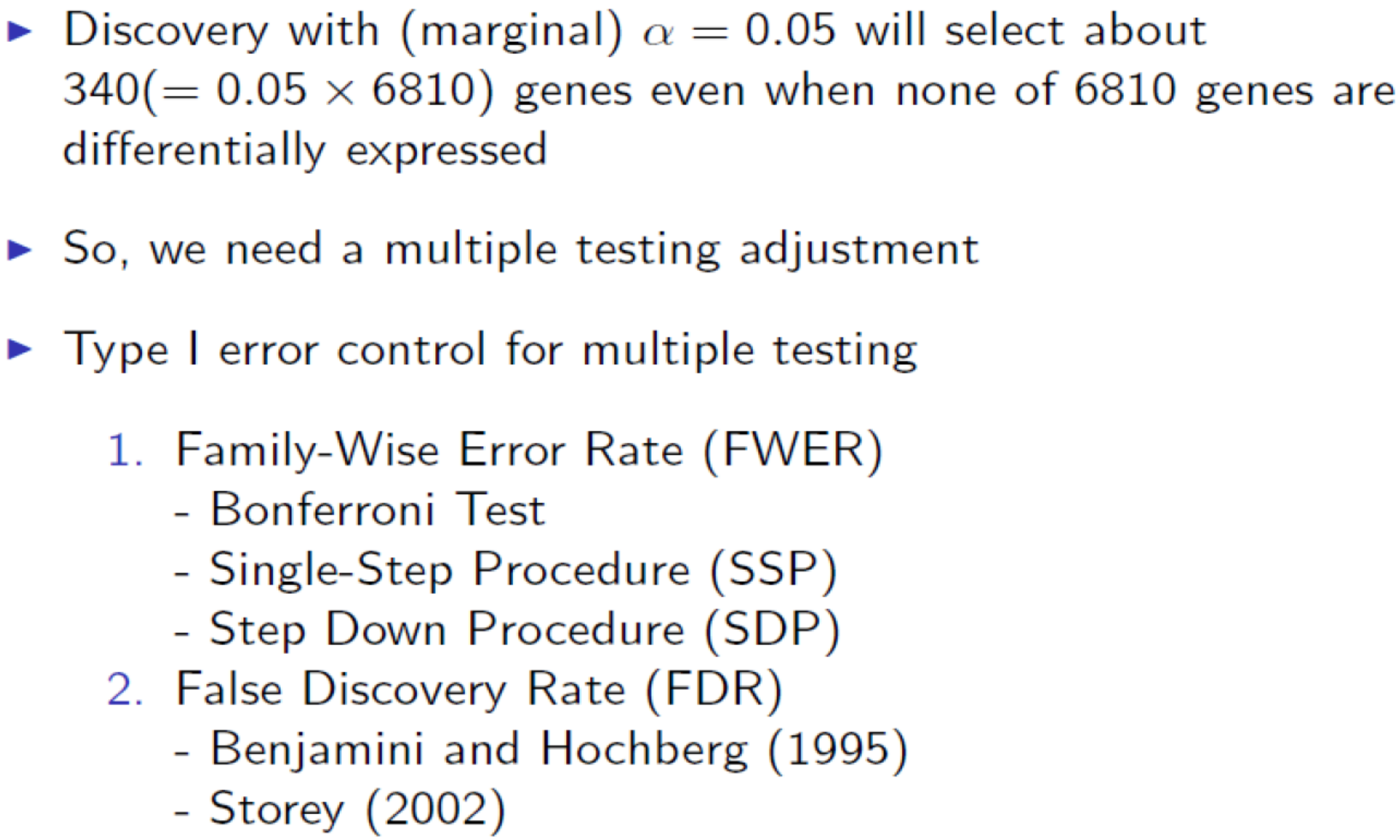
	Group 1			Group 2				
Gene	1	$\cdots$	n_1	1	$\cdots$	n_2	T-test	p-value
1	$\overbrace{x_{11}}$	$\cdots$	$\overbrace{x_{n_1,1}}$	$\overbrace{y_{11}}$	$\cdots$	$\overbrace{y_{n_2,1}}$	T_1	p_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
m	$\underbrace{x_{1m}}$	$\cdots$	$\underbrace{x_{n_1,m}}$	$\underbrace{y_{1m}}$	$\cdots$	$\underbrace{y_{n_2,m}}$	T_m	p_m

1. Genome Data



- Golub data: P-values by two-sample t-tests

1. Genome Data



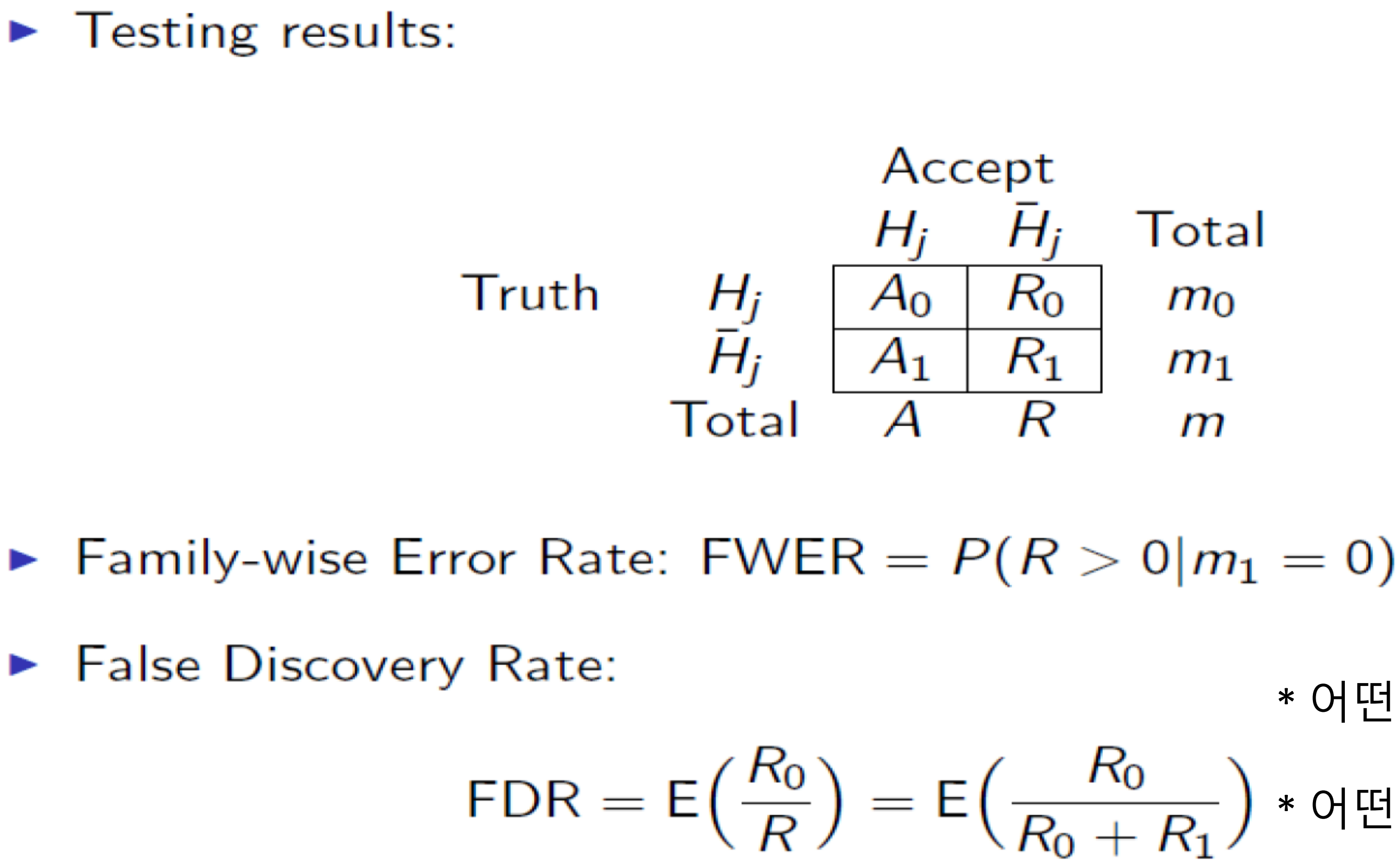
- Discovery with (marginal) $\alpha = 0.05$ will select about $340 (= 0.05 \times 6810)$ genes even when none of 6810 genes are differentially expressed

- So, we need a multiple testing adjustment

- Type I error control for multiple testing

  1. Family-Wise Error Rate (FWER)
     - Bonferroni Test
     - Single-Step Procedure (SSP)
     - Step Down Procedure (SDP)
  2. False Discovery Rate (FDR)
     - Benjamini and Hochberg (1995)
     - Storey (2002)

2. Multiple Testing



- Testing results:

		Accept		
		H_j	$\bar{H}_j$	Total
Truth	H_j	A_0	R_0	m_0
	$\bar{H}_j$	A_1	R_1	m_1
	Total	A	R	m

- Family-wise Error Rate: $\text{FWER} = P(R > 0 | m_1 = 0)$
- False Discovery Rate:

$$\text{FDR} = \text{E}\left(\frac{R_0}{R}\right) = \text{E}\left(\frac{R_0}{R_0 + R_1}\right)$$

*대립가설이 True가 하나도 없는데, 귀무가설을 하나이상 Reject할 확률 즉, 모든 변수가 유의하지 없는데 하나라도 유의하다고 나올 확률

* 귀무가설을 Reject한 모든 변수중에서 False-Positive의 비율

* 어떤 통계량이 더 보수적인 통계량인가 ??

* 어떤 통계량을 주로 사용되는가 ??

3. FWER : Bonferroni Test

Bonferroni Test은 다중 비교에서 생길 수 있는 오류를 보정하는 방법

통계추론에서 관측 데이터에 대해 귀무가설이 성립할 확률이 낮을 경우 귀무가설을 기각함으로써 대립가설을 채택한다.

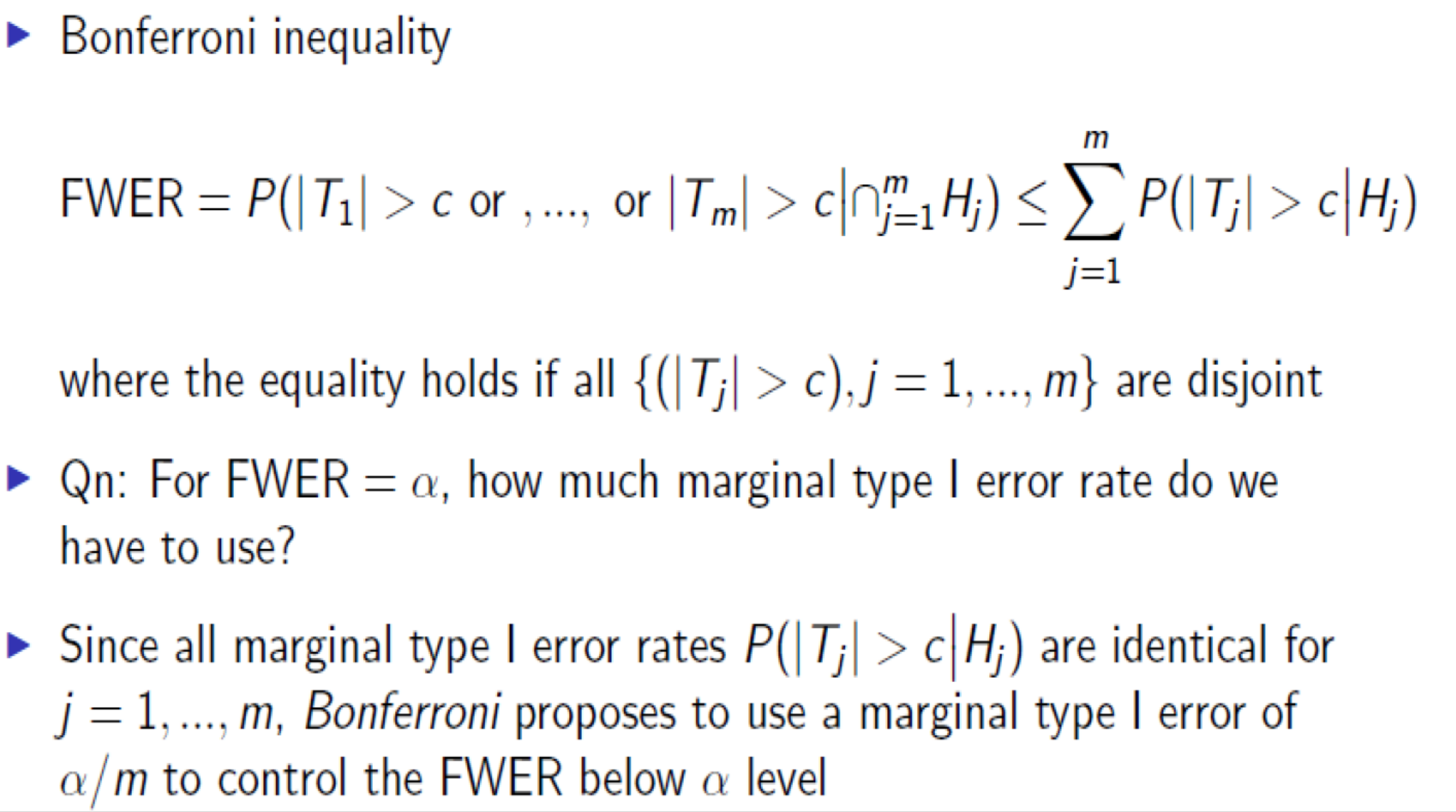
이때, 검정하는 가설의 숫자가 늘어나면 귀무가설이 기각될 확률이 낮더라도 기각될 가능성 더욱 늘어나게 된다.

이와 같은 경우, 귀무가설이 참임에도 불구하고 기각하는 제1종 오류가 발생한다.

이와 같은 오류를 보정하기 위해서 여러개의 가설들에 대해서 최소한 하나의 제1종오류가 발생할 가능성(familywise error rate; FWER)을 계산해 보정할 수 있다.

예를 들어 n개의 독립 혹은 비독립의 가설을 검정할 경우, 유의확률을 1/n로 낮추어 검정하는 것이다.

3. FWER : Bonferroni Test



- Bonferroni inequality

$$\text{FWER} = P(|T_1| > c \text{ or } , \ldots, \text{ or } |T_m| > c \big| \cap_{j=1}^{m} H_j) \leq \sum_{j=1}^{m} P(|T_j| > c \big| H_j)$$

where the equality holds if all $\{(|T_j| > c), j = 1, \ldots, m\}$ are disjoint

- Qn: For FWER $= \alpha$, how much marginal type I error rate do we have to use?
- Since all marginal type I error rates $P(|T_j| > c \big| H_j)$ are identical for $j = 1, \ldots, m$, *Bonferroni* proposes to use a marginal type I error of α/m to control the FWER below α level

3. FWER : Bonferroni Test

- Bonferroni Test는 매우 보수적인 방법

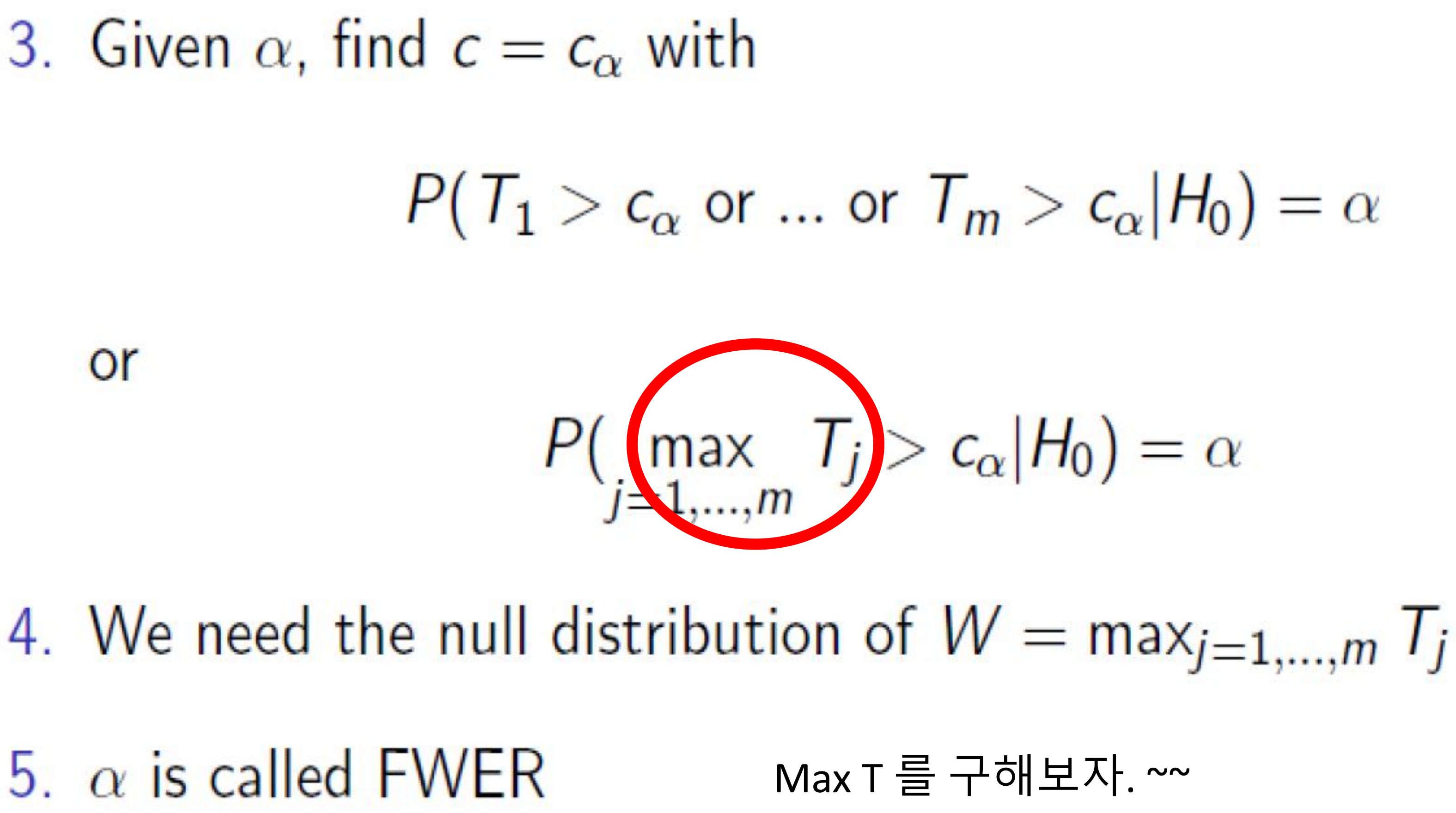
  - 어떤 유전자들은 coexpressed 됨.

  - m 이 매우 큼

- 모든 유전자가 완벽히 coexpressed된다면(T_1 = ... = T_m),
  FWER = P(| T_1 | > C | H_1)과 제 1종의 오류는 α / m 가 아니고, α 임.

- 제 1종의 오류는 α / m 와 α 사이의 어떤 값임.

3. FWER : Single-Step Multiple Test



3. Given α, find $c = c_\alpha$ with

$$P(T_1 > c_\alpha \text{ or } \dots \text{ or } T_m > c_\alpha | H_0) = \alpha$$

or

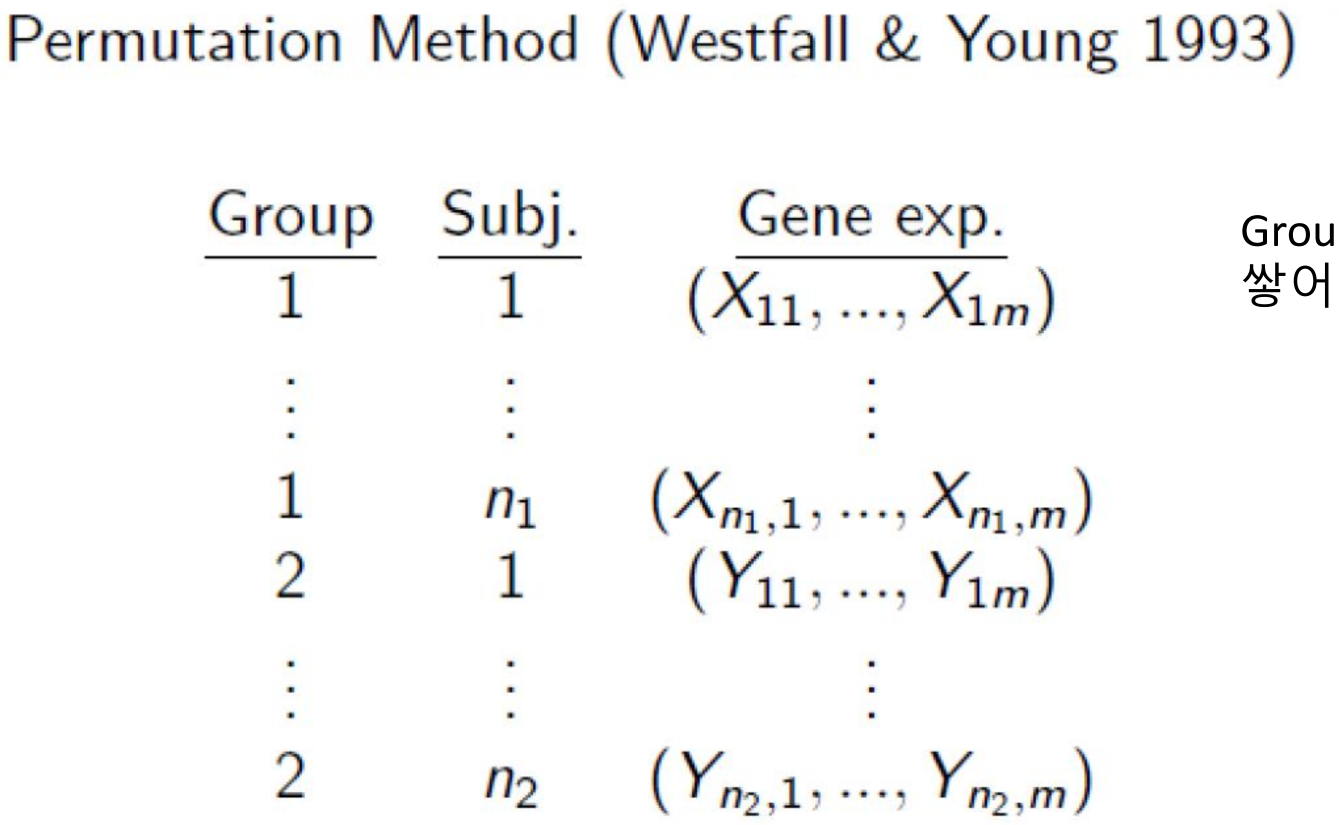
$$P(\boxed{\max_{j=1,\dots,m} T_j} > c_\alpha | H_0) = \alpha$$

4. We need the null distribution of $W = \max_{j=1,\dots,m} T_j$

5. α is called FWER

Max T 를 구해보자. ~~

3. FWER : Single-Step Multiple Test

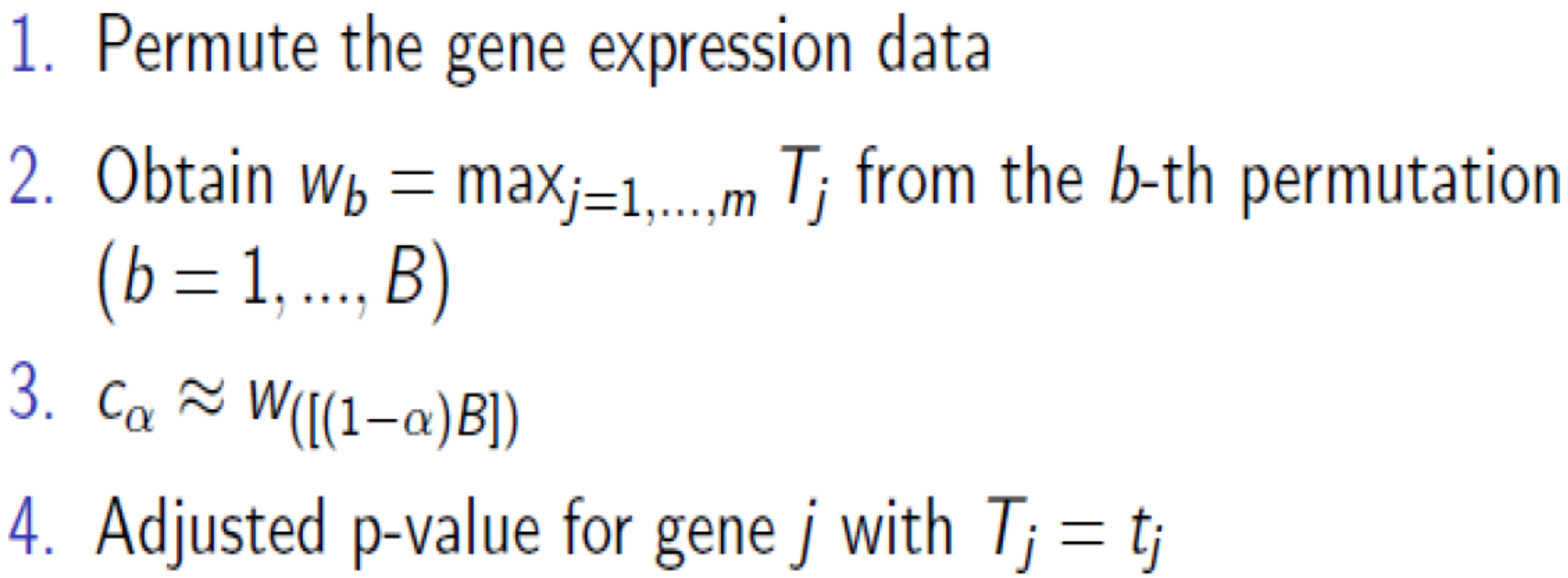


Permutation Method (Westfall & Young 1993)

Group	Subj.	Gene exp.
1	1	$(X_{11}, \ldots, X_{1m})$
⋮	⋮	⋮
1	n_1	$(X_{n_1,1}, \ldots, X_{n_1,m})$
2	1	$(Y_{11}, \ldots, Y_{1m})$
⋮	⋮	⋮
2	n_2	$(Y_{n_2,1}, \ldots, Y_{n_2,m})$

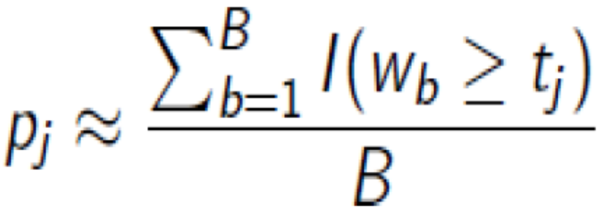
Group 1과 Group2 의 데이터를 무작위로 쌓어서 반복적으로 추출함.

3. FWER : Single-Step Multiple Test



1. Permute the gene expression data
2. Obtain $w_b = \max_{j=1,\ldots,m} T_j$ from the b-th permutation $(b = 1, \ldots, B)$
3. $c_\alpha \approx w_{([(1-\alpha)B])}$
4. Adjusted p-value for gene j with $T_j = t_j$

B번째 무작위로 추출한 데이터에서 얻은 최대 T 값을 W라고 함.

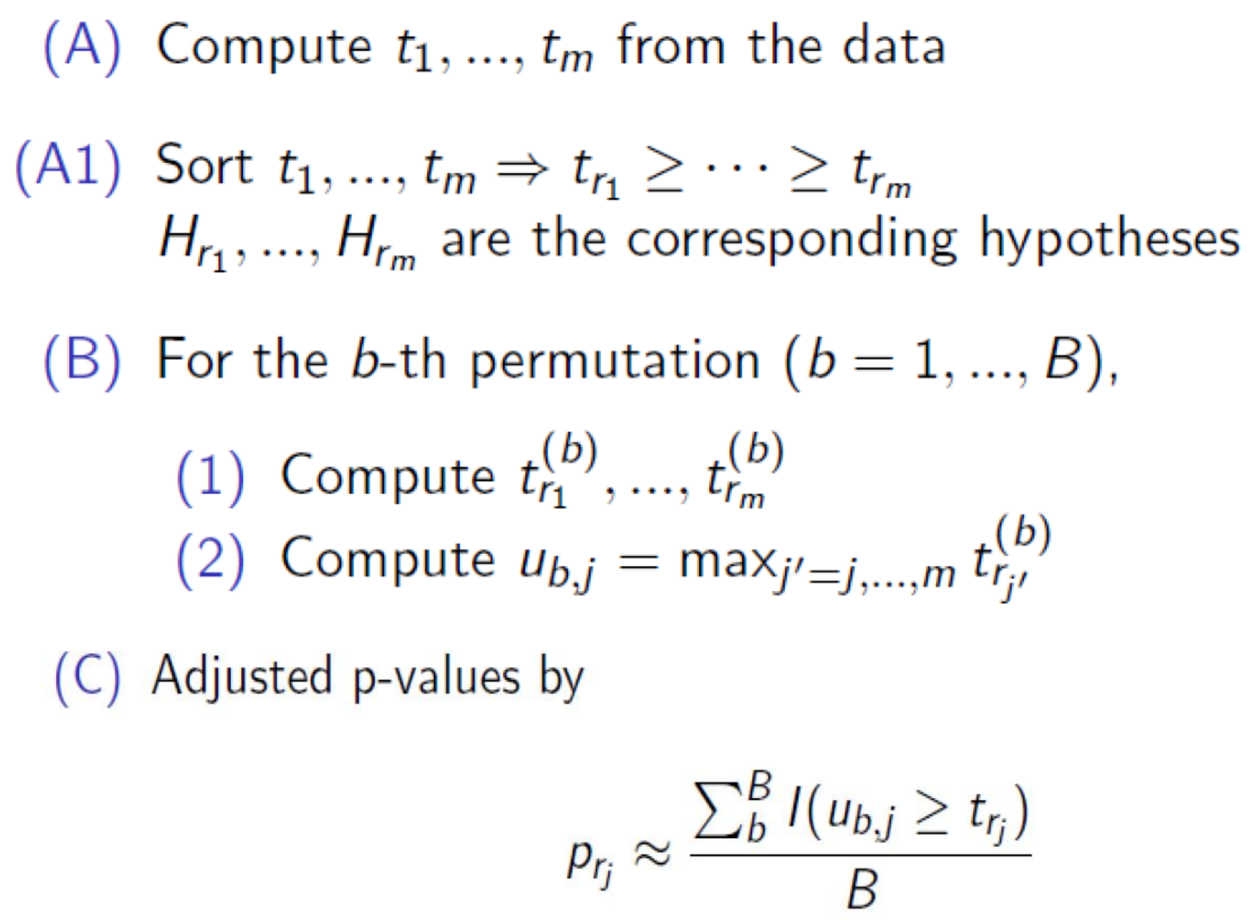


$$p_j \approx \frac{\sum_{b=1}^{B} I(w_b \geq t_j)}{B}$$

유전자 j의 t 값보다 큰 W 값을 카운트해서 B(permutation 횟수)로 나누면 p-value을 얻음.

B : 10,000번 정도

3. FWER : Step-Down Procedure



(A) Compute $t_1, \ldots, t_m$ from the data

(A1) Sort $t_1, \ldots, t_m \Rightarrow t_{r_1} \geq \cdots \geq t_{r_m}$
$H_{r_1}, \ldots, H_{r_m}$ are the corresponding hypotheses

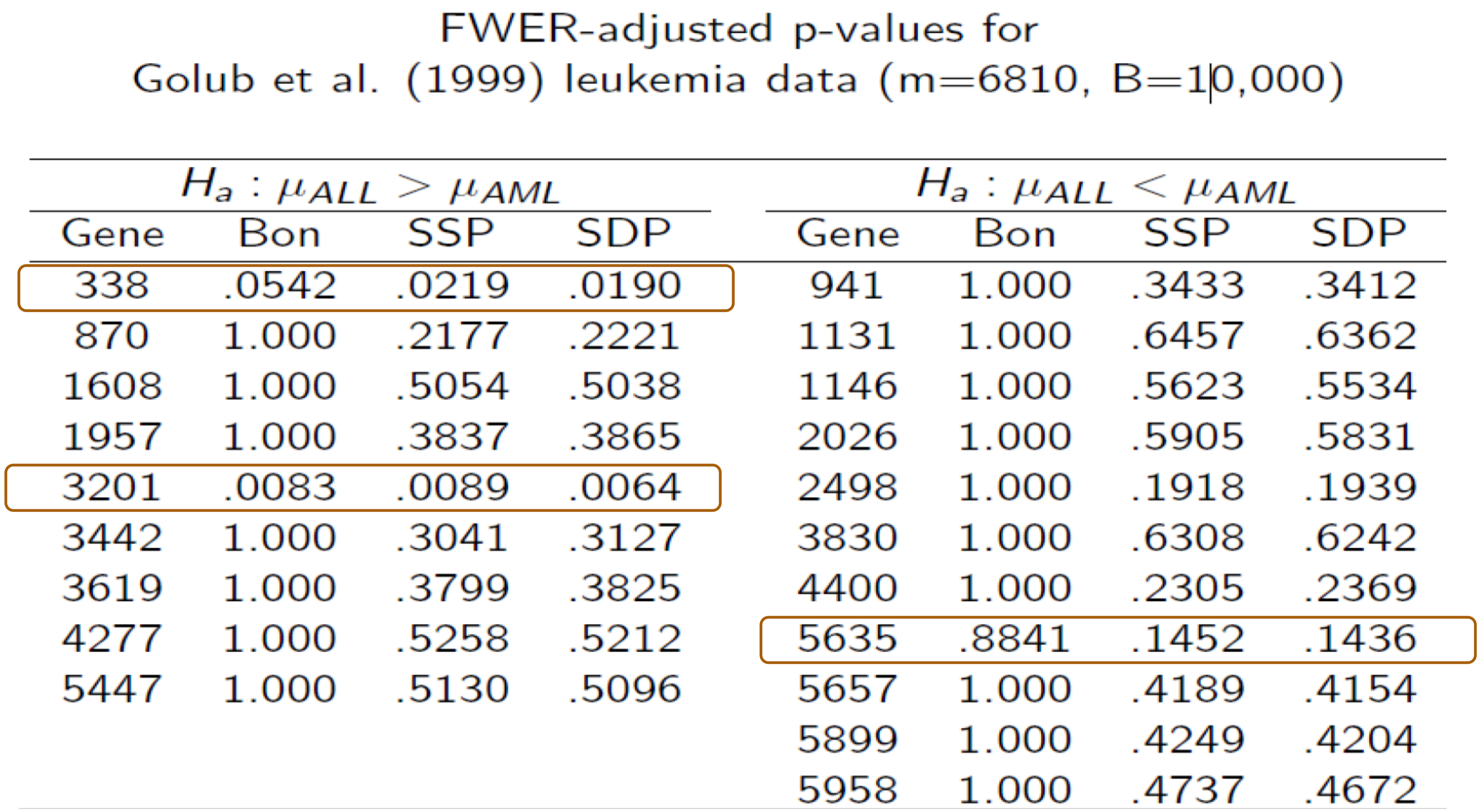
(B) For the b-th permutation $(b = 1, \ldots, B)$,

(1) Compute $t_{r_1}^{(b)}, \ldots, t_{r_m}^{(b)}$
(2) Compute $u_{b,j} = \max_{j'=j,\ldots,m} t_{r_{j'}}^{(b)}$

(C) Adjusted p-values by

$$p_{r_j} \approx \frac{\sum_b^B I(u_{b,j} \geq t_{r_j})}{B}$$

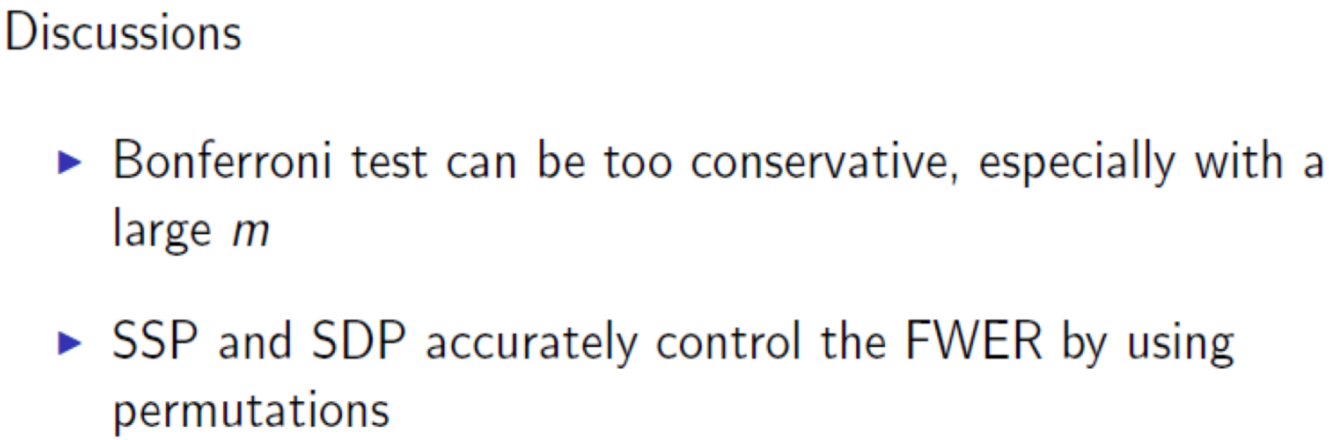
3. FWER



FWER-adjusted p-values for
Golub et al. (1999) leukemia data (m=6810, B=10,000)

$H_a : \mu_{ALL} > \mu_{AML}$				$H_a : \mu_{ALL} < \mu_{AML}$			
Gene	Bon	SSP	SDP	Gene	Bon	SSP	SDP
338	.0542	.0219	.0190	941	1.000	.3433	.3412
870	1.000	.2177	.2221	1131	1.000	.6457	.6362
1608	1.000	.5054	.5038	1146	1.000	.5623	.5534
1957	1.000	.3837	.3865	2026	1.000	.5905	.5831
3201	.0083	.0089	.0064	2498	1.000	.1918	.1939
3442	1.000	.3041	.3127	3830	1.000	.6308	.6242
3619	1.000	.3799	.3825	4400	1.000	.2305	.2369
4277	1.000	.5258	.5212	5635	.8841	.1452	.1436
5447	1.000	.5130	.5096	5657	1.000	.4189	.4154
				5899	1.000	.4249	.4204
				5958	1.000	.4737	.4672

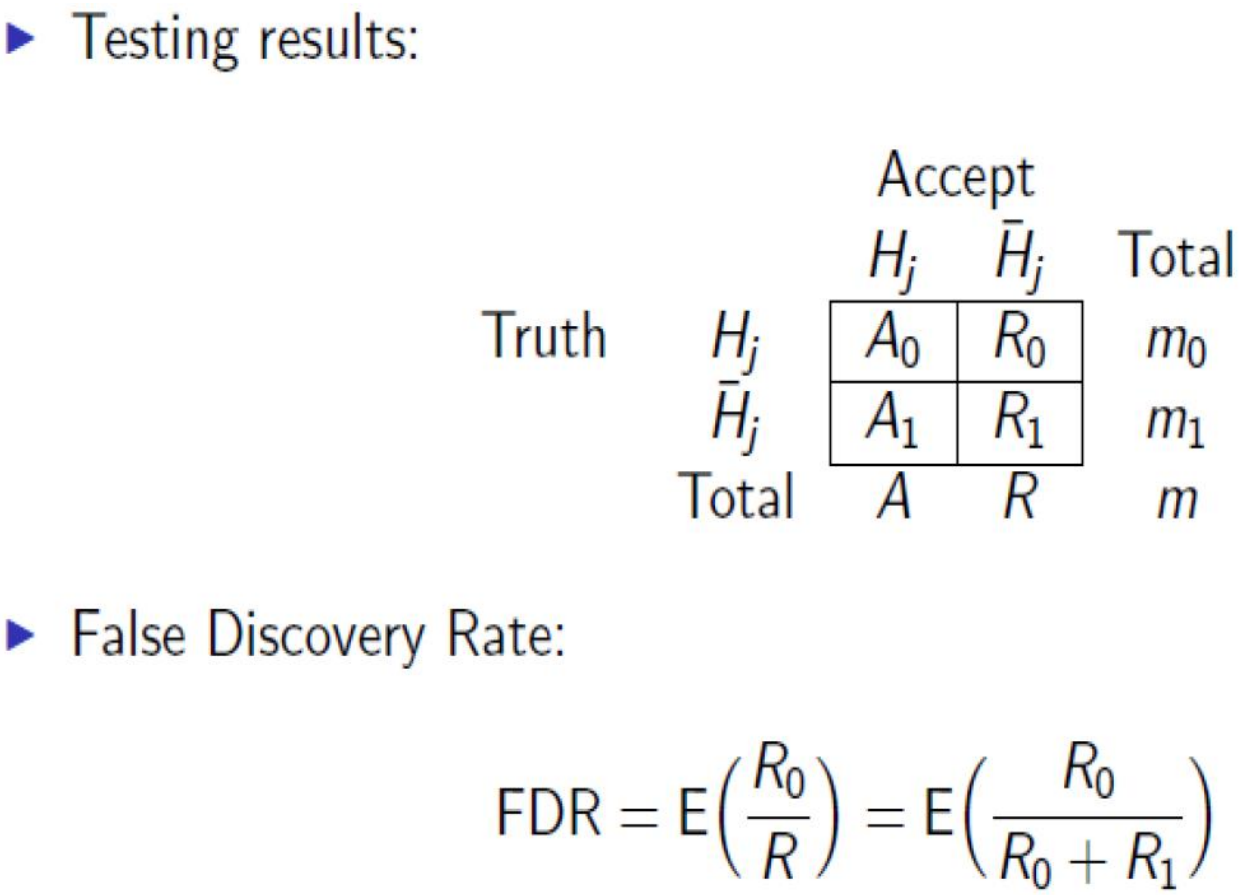
3. FWER



Discussions

- Bonferroni test can be too conservative, especially with a large m
- SSP and SDP accurately control the FWER by using permutations

4. FDR



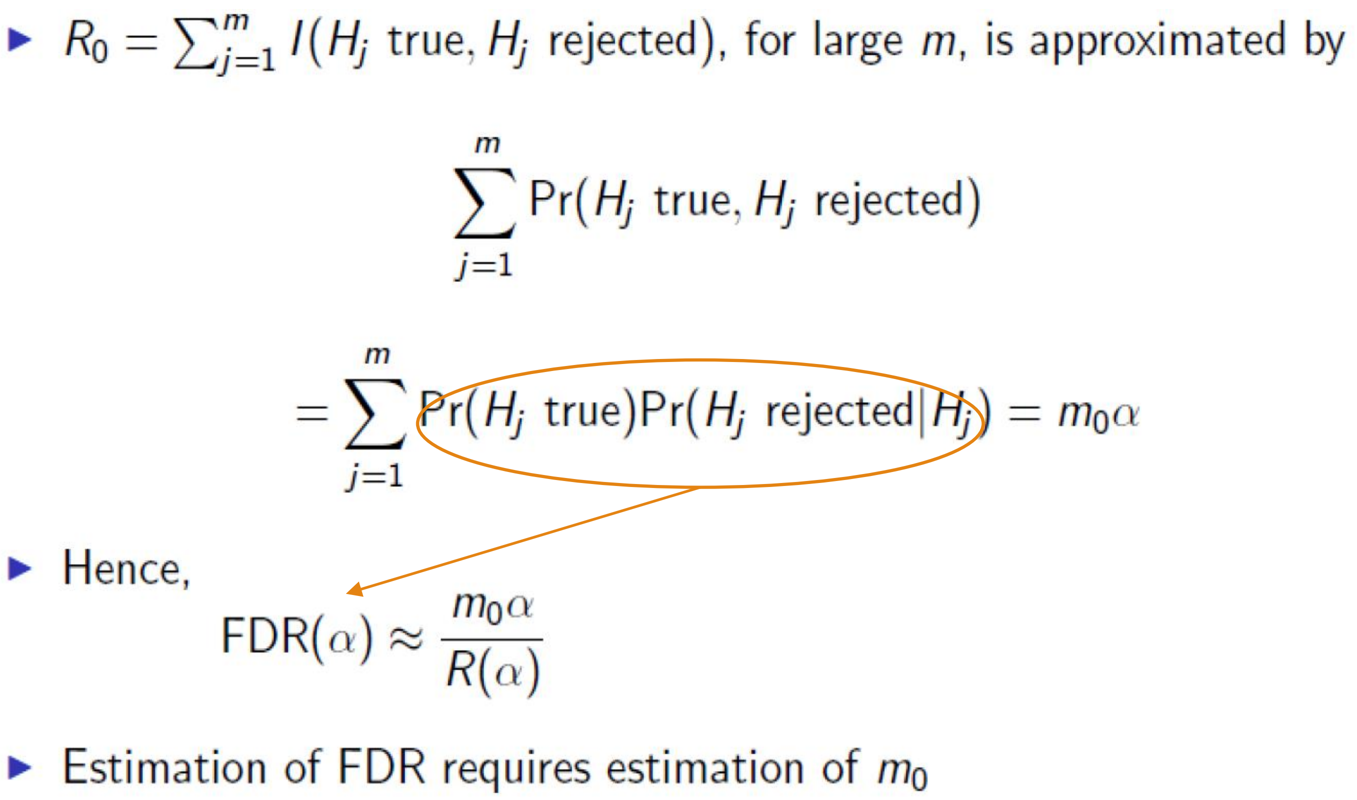
- Testing results:

		Accept		
		H_j	$\bar{H}_j$	Total
Truth	H_j	A_0	R_0	m_0
	$\bar{H}_j$	A_1	R_1	m_1
	Total	A	R	m

- False Discovery Rate:

$$\mathrm{FDR} = \mathrm{E}\left(\frac{R_0}{R}\right) = \mathrm{E}\left(\frac{R_0}{R_0 + R_1}\right)$$

4. FDR : Storey의 방법



- $R_0 = \sum_{j=1}^{m} I(H_j \text{ true}, H_j \text{ rejected})$, for large m, is approximated by

$$\sum_{j=1}^{m} \Pr(H_j \text{ true}, H_j \text{ rejected})$$

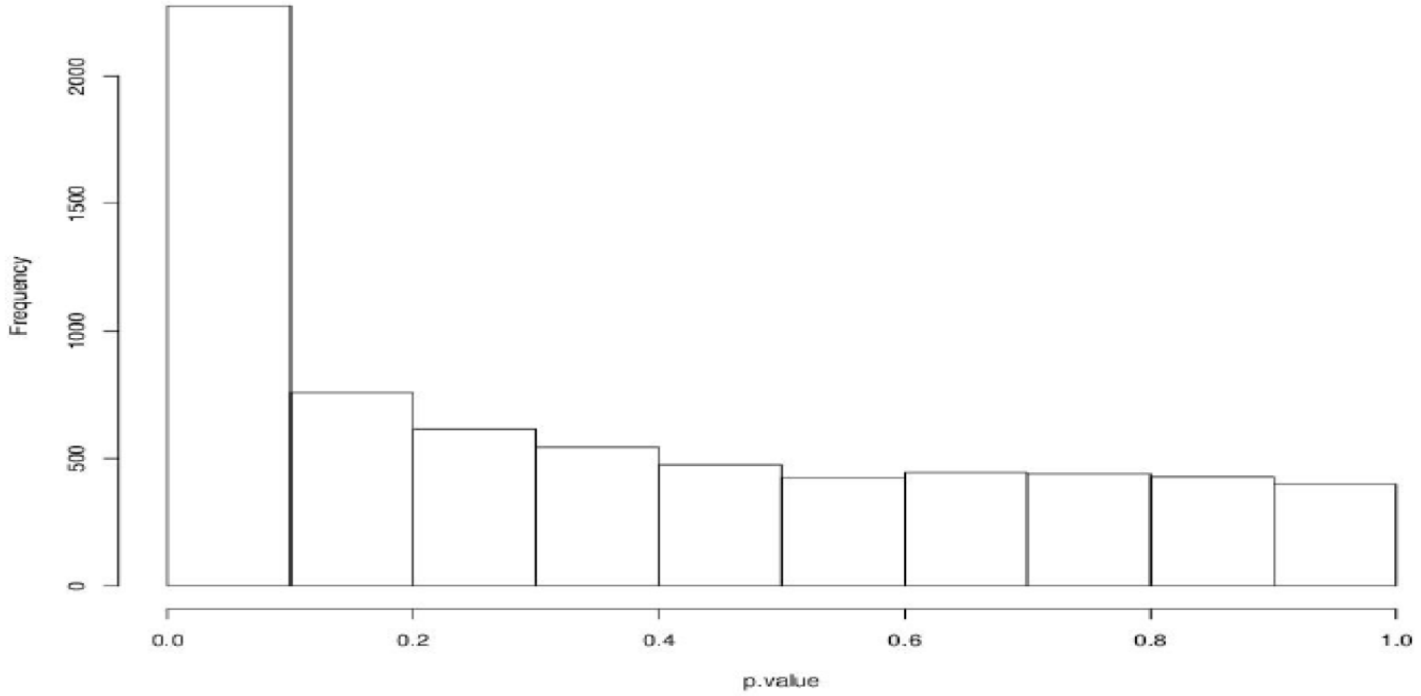
$$= \sum_{j=1}^{m} \Pr(H_j \text{ true}) \Pr(H_j \text{ rejected} | H_j) = m_0 \alpha$$

- Hence,

$$\text{FDR}(\alpha) \approx \frac{m_0 \alpha}{R(\alpha)}$$

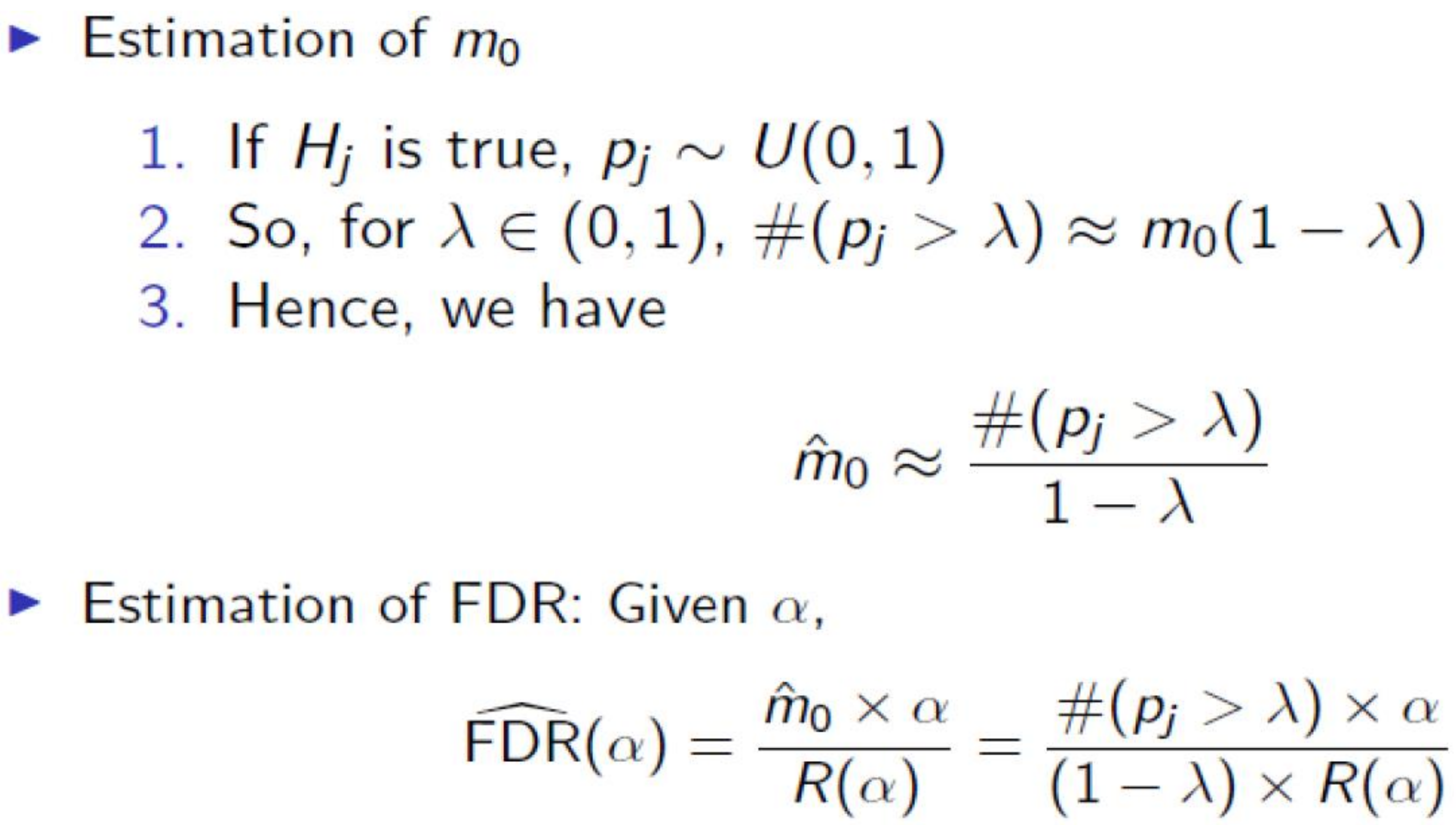
- Estimation of FDR requires estimation of m_0

4. FDR : Storey의 방법



- Figure: P-values by two-sample t-tests.

4. FDR : Storey의 방법



- Estimation of m_0

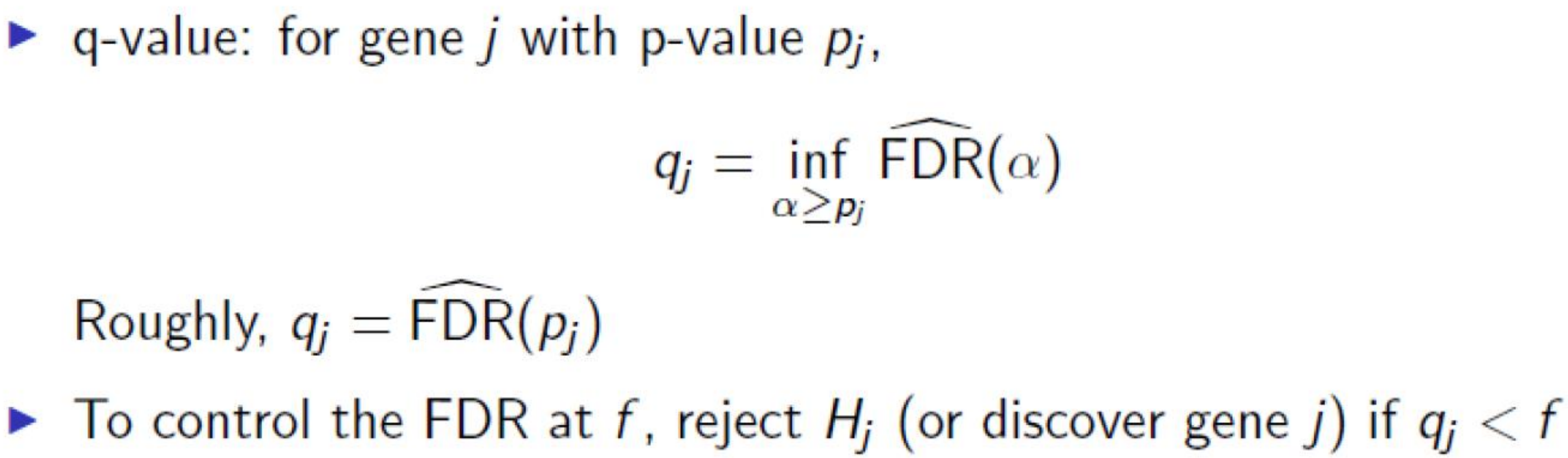
  1. If H_j is true, $p_j \sim U(0,1)$
  2. So, for $\lambda \in (0,1)$, $\#(p_j > \lambda) \approx m_0(1-\lambda)$
  3. Hence, we have

$$\hat{m}_0 \approx \frac{\#(p_j > \lambda)}{1-\lambda}$$

- Estimation of FDR: Given α,

$$\widehat{\mathrm{FDR}}(\alpha) = \frac{\hat{m}_0 \times \alpha}{R(\alpha)} = \frac{\#(p_j > \lambda) \times \alpha}{(1-\lambda) \times R(\alpha)}$$

4. FDR : Storey의 방법



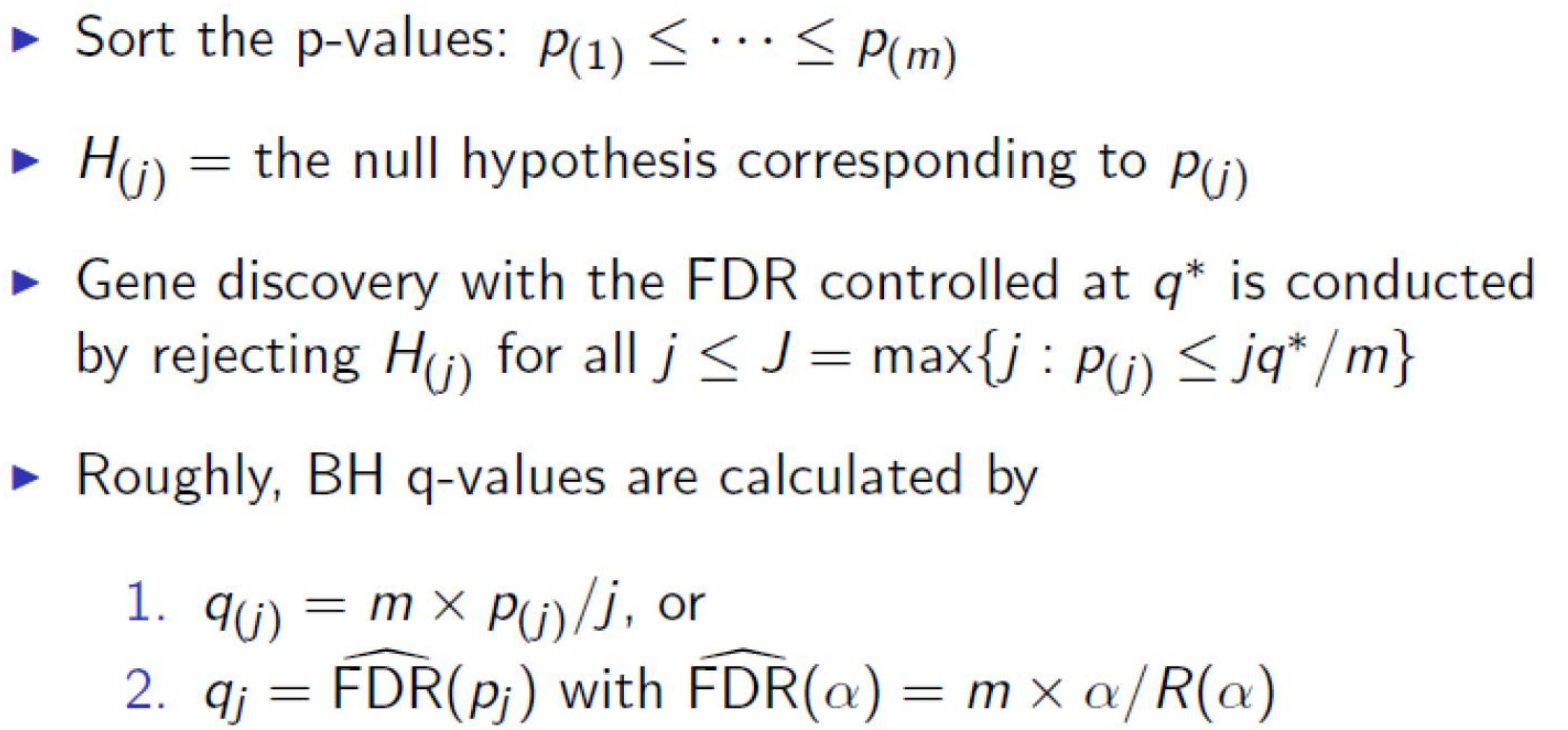
- q-value: for gene j with p-value p_j,

$$q_j = \inf_{\alpha \geq p_j} \widehat{\text{FDR}}(\alpha)$$

  Roughly, $q_j = \widehat{\text{FDR}}(p_j)$

- To control the FDR at f, reject H_j (or discover gene j) if $q_j < f$

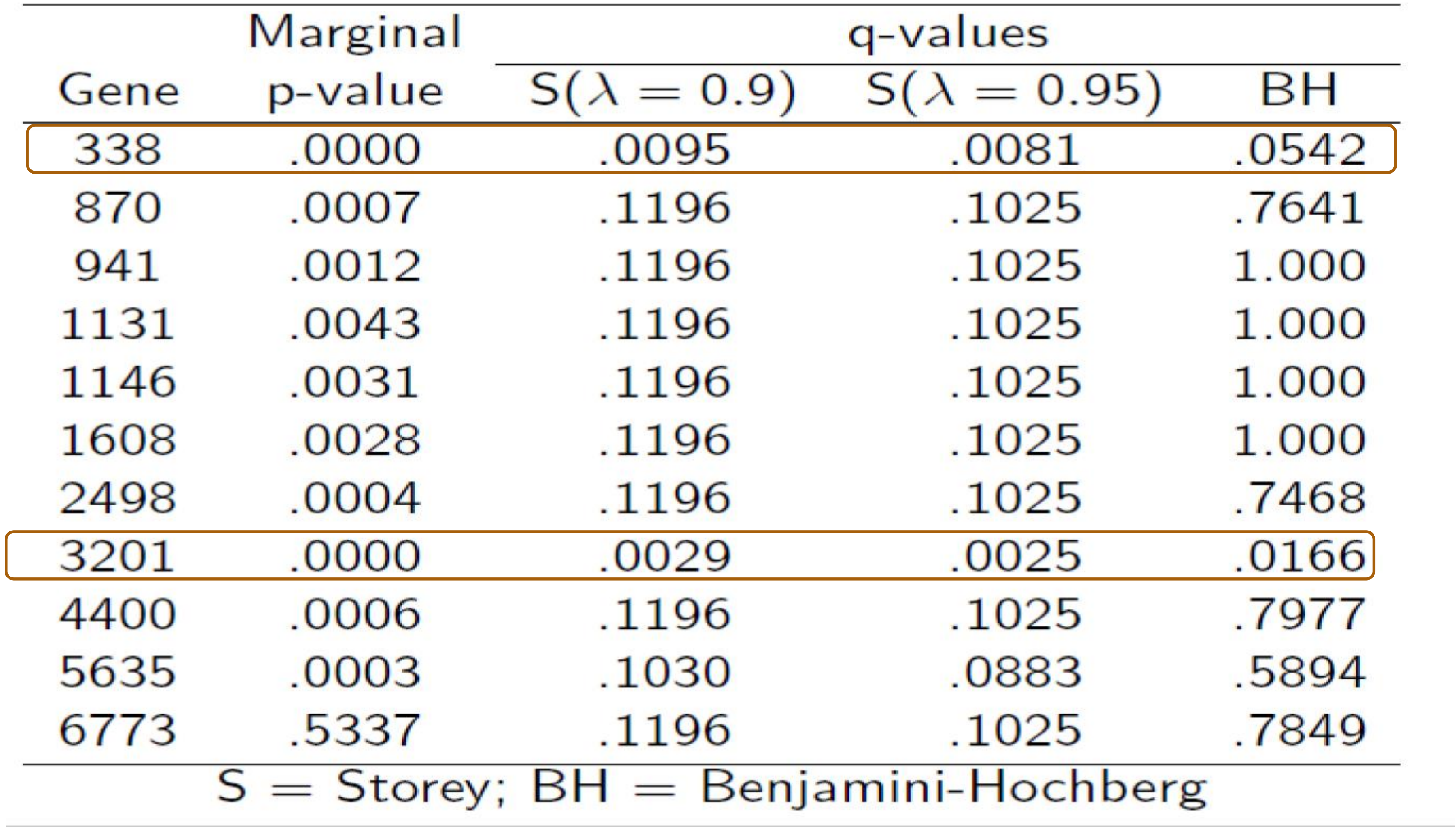
4. FDR : Benjamin-Hochberg procedure



- Sort the p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$
- $H_{(j)}$ = the null hypothesis corresponding to $p_{(j)}$
- Gene discovery with the FDR controlled at q^* is conducted by rejecting $H_{(j)}$ for all $j \leq J = \max\{j : p_{(j)} \leq jq^*/m\}$
- Roughly, BH q-values are calculated by

  1. $q_{(j)} = m \times p_{(j)}/j$, or
  2. $q_j = \widehat{\mathrm{FDR}}(p_j)$ with $\widehat{\mathrm{FDR}}(\alpha) = m \times \alpha/R(\alpha)$

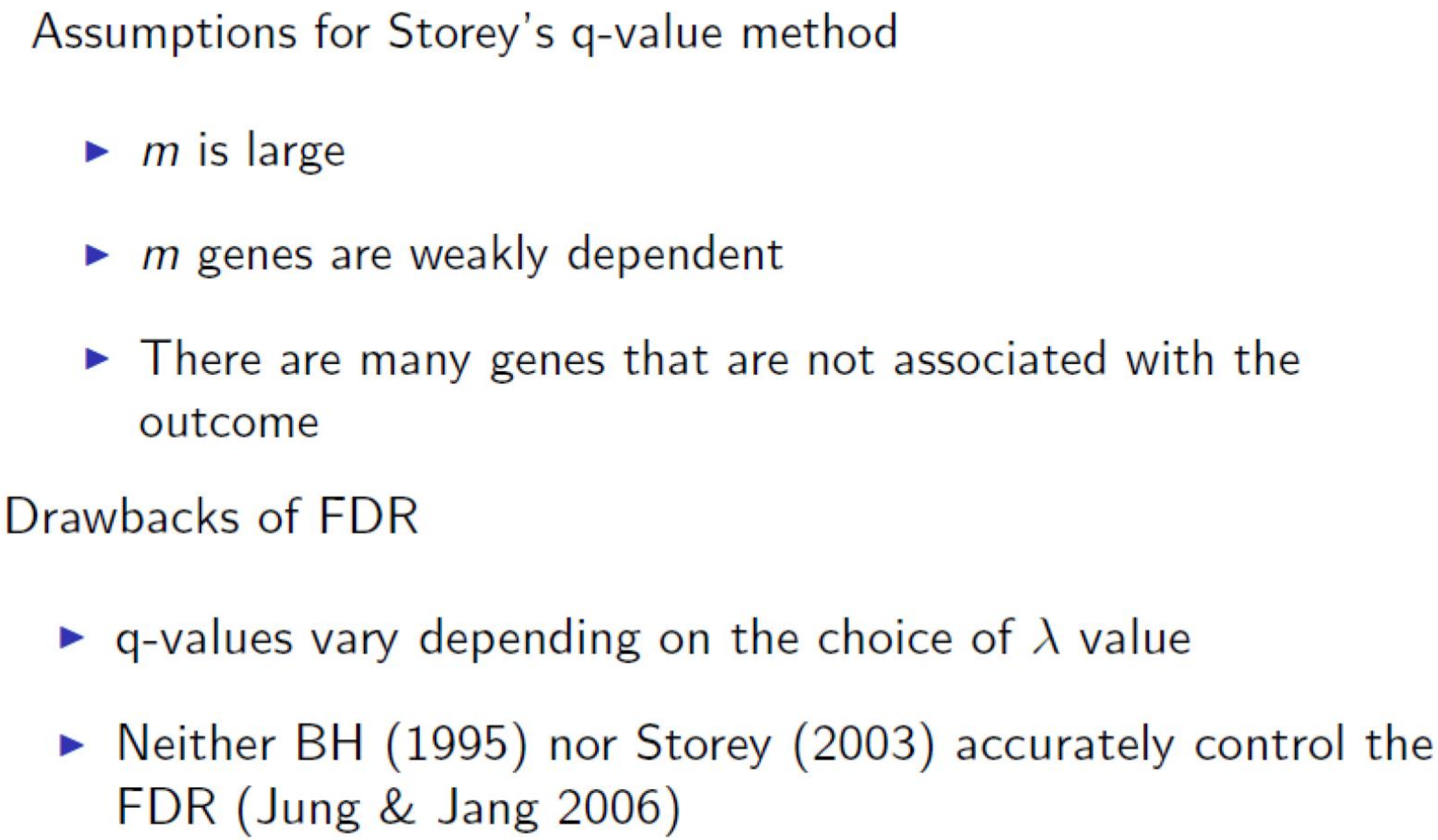
4. FDR



	Marginal	q-values		
Gene	p-value	$S(\lambda = 0.9)$	$S(\lambda = 0.95)$	BH
338	.0000	.0095	.0081	.0542
870	.0007	.1196	.1025	.7641
941	.0012	.1196	.1025	1.000
1131	.0043	.1196	.1025	1.000
1146	.0031	.1196	.1025	1.000
1608	.0028	.1196	.1025	1.000
2498	.0004	.1196	.1025	.7468
3201	.0000	.0029	.0025	.0166
4400	.0006	.1196	.1025	.7977
5635	.0003	.1030	.0883	.5894
6773	.5337	.1196	.1025	.7849

S = Storey; BH = Benjamini-Hochberg

4. FDR



Assumptions for Storey's q-value method

- m is large
- m genes are weakly dependent
- There are many genes that are not associated with the outcome

Drawbacks of FDR

- q-values vary depending on the choice of λ value
- Neither BH (1995) nor Storey (2003) accurately control the FDR (Jung & Jang 2006)