Lecture 20:
Multiple Hypothesis Testing

GENOME 560

Su-In Lee, CSE & GS (suinlee@uw.edu)

Outline

- Why multiple hypothesis testing matters?
  - A motivating example
  - R-session
- Basic concepts on multiple hypothesis testing
  - Type I error and Type II error
  - Define the multiple testing problem and related concepts
- Methods for multiple hypothesis testing correction
  - 2 methods that control the family-wise error rate (FWER)
  - 1 method that controls the false discovery rate (FDR)
  - *Our goal is to understand what type of error is controled in each method, such that we know which one to use in our research.*

Motivating example

- Storey and Akey (2007). **Gene-expression variation within and among human populations.** *AJHG*
  - Understand patterns of gene-expression variation within and among human populations.
  - Gain insights into the molecular basis of phenotypic diversity.
- The authors measured expression levels of 5194 genes in 16 human individuals
  - 8 European + 8 African individuals
- Apply t-test to each gene. This results in 5194 p-values!

Expression Data

- Here is the expression data from Storey & Akey (2007):

Replicates from an individual CEU_1

16 samples from 8 Eurpean individuals

16 samples from 8 African individuals

32 samples

5419 genes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ProbeSet	CEU_1_1	CEU_1_2	CEU_2_1	CEU_2_2	CEU_3_1	CEU_3_2	CEU_4_1	CEU_4_2	CEU_5_1	CEU_5_2	CEU_6_1	CEU_6_2	CEU_7_1	CEU_7_2	CEU_8_1	CEU_8_2	YRI_1_1	YRI_1_2	YRI_2_1	YRI_2_2
2	1007_s_at	5.98169	5.67082	6.41715	6.08762	5.91509	6.17978	6.21335	6.03895	6.06861	6.07892	5.78884	5.74019	6.29691	6.49425	5.96735	6.01498	5.6815	5.77367	5.77471	6.1653
3	1053_at	7.79122	8.07729	8.14663	7.64775	7.89844	8.12242	8.36584	8.19187	8.22189	8.28195	7.82862	8.01362	7.89883	7.58709	7.74083	7.90788	7.56851	7.79912	8.3541	8.08236
4	117_at	4.97942	4.71225	5.10184	5.04141	4.90681	5.14985	4.86161	4.78877	4.95509	5.23931	4.96736	4.85927	4.84583	5.11296	5.01919	4.84037	4.91608	5.04789	4.81784	4.82377
5	121_at	7.13332	7.02673	7.26751	6.90441	7.23218	7.12554	7.36271	7.12975	7.45557	7.49174	7.32974	7.11061	7.39483	7.35183	7.39411	7.26759	7.42888	7.263	6.9638	7.31065
6	1294_at	7.52736	7.21817	7.19655	7.76186	7.5295	6.91137	7.11729	6.42918	7.3382	7.29095	7.47051	6.95853	7.34192	7.50438	7.39755	6.94965	7.66957	7.20394	7.02302	7.10666
7	1316_at	4.26332	4.17658	4.27741	4.30384	4.26586	4.20854	4.34486	4.22165	4.40363	4.5199	4.28163	4.27318	4.45466	4.44125	4.52665	4.38887	4.4778	4.31582	4.23707	4.34228
8	1487_at	6.74382	6.70351	6.91644	7.04603	6.60947	6.63286	6.72988	6.69927	6.8179	6.80211	6.85259	6.7266	6.52409	6.59425	6.63326	6.52908	6.77679	6.72127	6.71948	6.90661
9	1494_f_at	5.16651	4.99188	5.12793	5.25252	5.03059	5.06722	5.27456	4.84703	5.05162	5.44135	5.14389	4.87616	5.11117	5.25574	5.33691	5.13767	5.3711	5.14092	4.89588	5.17297
10	160020_at	6.02283	5.89353	5.98744	5.9435	5.97474	6.04196	6.04496	5.80167	5.97236	6.30119	5.97172	5.84176	6.04939	6.22937	6.15538	6.06261	6.18919	5.89937	5.70936	6.06758
11	1729_at	7.69774	7.40541	7.31776	7.51505	7.54614	7.32557	7.28511	7.01336	7.18561	7.0188	7.57075	7.3487	7.60301	7.71821	7.5925	7.57401	7.71808	7.60594	7.24643	7.20956
12	1773_at	5.86558	6.15341	5.98571	5.14151	5.90395	6.13465	5.8256	5.42099	5.78256	5.92947	5.77468	5.77462	5.7773	5.98557	6.07014	5.89564	5.83511	5.72583	5.99022	5.77984
13	177_at	5.85235	6.63166	4.89025	5.12546	6.51927	7.15841	4.99199	5.49555	5.28412	5.93808	5.25714	5.63431	5.80036	6.40891	5.5988	5.53586	5.13165	6.14363	5.56086	6.36961
14	1861_at	6.15471	6.38081	6.09384	6.02244	5.70657	5.73772	6.16299	5.57907	6.24892	6.00059	6.01572	6.18548	6.34677	6.14858	5.87171	6.038	6.02631	5.88272	6.34075	5.97614
15	200000_s_at	9.13278	9.21692	8.99018	9.63054	9.55528	8.97999	9.13531	8.57438	9.4126	9.23347	9.27055	9.22298	9.01712	9.15443	9.31366	9.01806	9.3888	9.17289	9.1096	9.11917
16	200001_at	9.83966	9.35552	9.2242	9.42504	9.58869	9.7412	9.38165	8.85948	9.52156	9.42842	9.64574	9.32279	9.77189	10.11824	9.62728	9.45831	9.65025	9.58449	9.3861	9.71814
17	200002_at	12.11383	11.94651	12.16294	11.85746	12.08855	11.98535	12.174	11.93448	12.13459	12.12292	12.16459	11.96674	12.21628	12.06945	12.18677	12.04452	12.1959	12.09627	12.12228	11.99147
18	200003_s_at	13.07069	13.01511	13.03935	12.87746	12.93673	12.75523	13.01374	13.02525	13.00242	13.07201	13.00242	13.02889	13.09854	12.93907	13.0692	13.05037	13.153	13.17386	13.02241	13.0163
19	200004_at	10.55842	11.03104	10.96424	10.72764	10.82515	10.91759	10.91447	11.2844	11.31299	11.0198	10.92223	11.04594	10.81108	11.0729	10.94669	11.05366	10.63201	10.99445	11.07644	11.01879
20	200005_at	10.97043	10.65963	10.97516	11.06102	10.72875	10.71474	10.91149	10.76657	10.66952	10.66663	10.79352	10.72233	11.09684	10.88053	10.67535	10.73676	10.80804	10.71755	10.74218	10.72491
21	200006_at	11.48915	11.61778	11.81477	11.40593	11.25149	11.59819	11.71656	11.80532	11.65274	11.53486	11.45718	11.33958	11.74612	11.6041	11.48975	11.57558	11.43568	11.64114	11.71386	11.57568
22	200007_at	11.32918	11.49135	11.43332	10.69464	11.19353	11.30759	11.36443	11.47085	11.32257	11.22004	11.42069	11.34796	11.37548	11.29618	11.39315	11.319	11.40391	11.48604	11.35568	11.34042
23	200008_s_at	10.05924	10.19904	10.08833	7.31779	9.20387	9.1517	10.10063	10.33664	10.04564	10.2272	10.09962	9.39115	10.07026	10.05453	9.64721	9.51131	10.27698	9.74414	10.2275	10.25982
24	200009_at	11.0563	11.12816	11.13257	11.07872	11.58731	11.47549	11.06448	11.30978	11.09186	11.26098	11.22962	11.3609	11.10342	10.95636	11.117	11.078	11.14835	11.18163	11.20217	11.27175
25	200010_at	12.38319	12.36744	12.46796	12.03685	11.80795	11.88995	12.45424	12.45923	12.36846	12.30105	12.4119	12.28217	12.60342	12.51502	12.33517	12.30307	12.50114	12.5221	12.30057	12.22573
26	200011_s_at	8.9967	9.07854	8.65446	9.1407	9.18669	8.98629	8.89977	8.92349	9.13694	9.06953	9.10752	8.78164	9.18106	9.59519	9.15071	9.00216	8.63081	9.01313	9.08474	9.20584
27	200012_x_at	13.2107	13.09015	13.19736	13.30851	13.21507	13.22693	13.27011	13.20257	13.19096	13.20106	13.26242	13.41315	13.26763	13.16261	13.31986	13.36066	13.25649	13.37189	13.14087	13.05253
28	200013_at	12.39168	12.39763	12.54363	12.26828	12.32055	12.41859	12.45704	12.67021	12.43595	12.42122	12.35361	12.33161	12.56036	12.49851	12.46436	12.51449	12.31502	12.44044	12.47497	12.38138
29	200014_s_at	9.94128	10.20595	10.27985	10.01113	10.12948	10.34191	10.26396	10.39993	10.23167	10.22328	10.07787	10.45789	10.22766	9.65756	10.23254	10.41426	10.02346	10.23485	10.25513	9.92751
30	200015_s_at	10.08432	10.36298	10.04279	9.60846	10.45268	10.52381	9.9906	10.22174	10.28308	10.47019	10.19026	10.40726	10.05219	10.05065	10.16536	10.30576	10.11721	10.44556	10.17652	10.29149
31	200016_x_at	12.02516	12.15283	12.20907	12.04757	12.28712	12.26901	12.18126	12.26704	12.27663	12.04564	12.17173	12.31069	12.11635	12.05232	12.20532	12.26479	12.03764	12.19688	12.20176	12.08424

Questions of interest

- How many genes show a significant difference in expression levels between European and African individuals?

- Which genes show a significant difference?

How would you answer these questions?

Let's try…

- Load the data

```
a <- read.table(header = T,
file="http://www.cs.washington.edu/homes/suinlee/genome560/RMA_Filtered.txt")
b <- a[,2:33]
```

- Define a function of performing the t-test

```
fun <- function(d){return(t.test(d[1:16],d[17:32])$p.value)}
```

- Obtain the p-values

```
p <- apply(b, 1, fun)
```

- See the distribution

```
hist(p, breaks=20)
```

We got many p-values!

- Which genes show significant difference in expression levels between European and African individuals?

- How many p-values < 0.05?

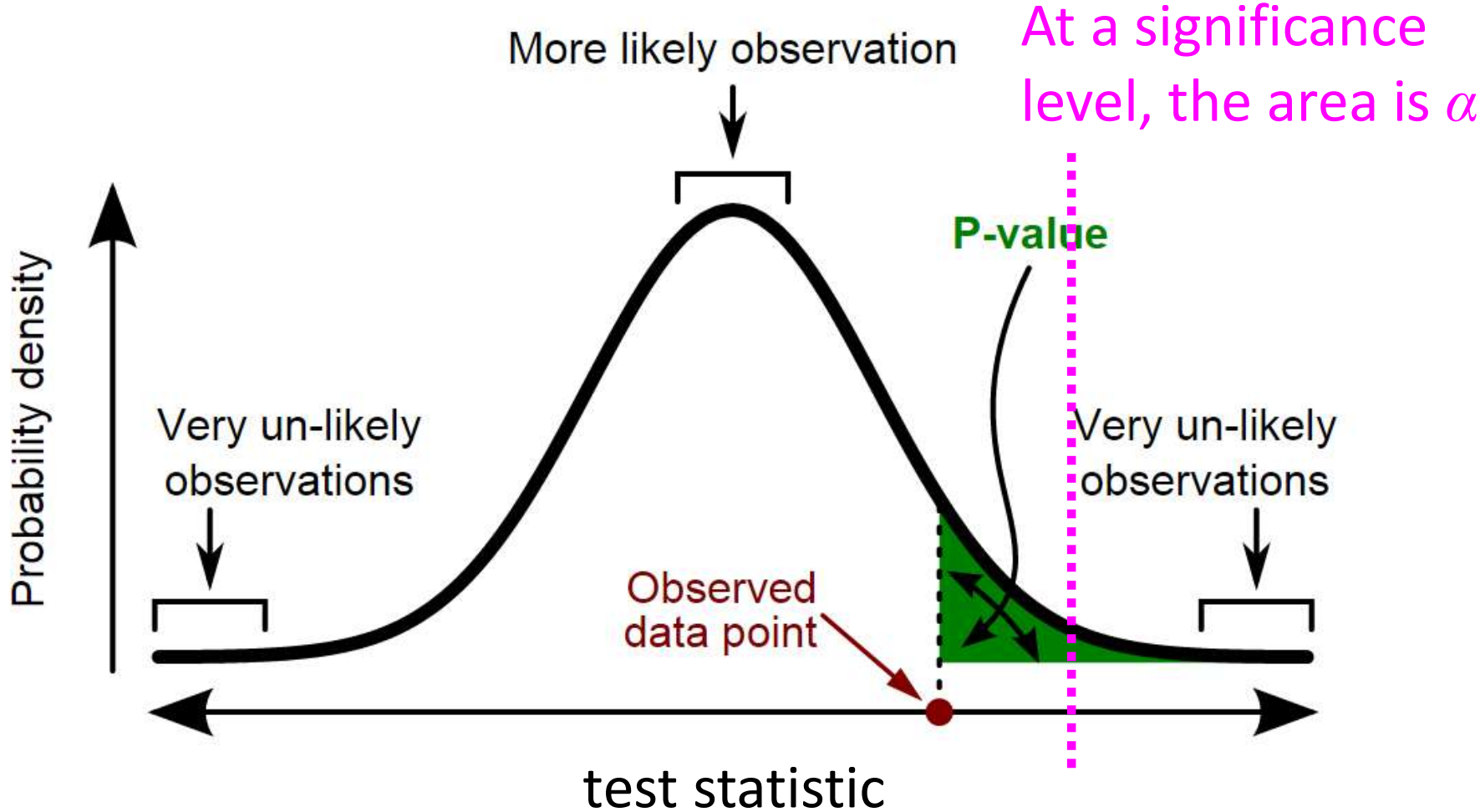
```
alpha = 0.05
tabulate(as.numeric(p < alpha))
```

- Would a standard p-value cutoff α = 0.05 (or 0.01) be useful when there are many hypotheses?

- Definition of p-value:
  - The estimated percentage of observations more extreme than the one observed under the assumption that the null hypothesis is true

Review: p-value

- Distributin of the test statistic (e.g., f-ratio in ANOVA) when the null hypothesis is true:



Why Multiple Testing Matters I

- We have 5194 hypothesis tests in this problem
  - A typical microarray experiment might result in performing 10,000 separate hypothesis tests.
  - Genomics: Lots of data, Lots of hypothesis tests
- We would expect ***~260*** (5194 x 0.05) genes to be deemed "significant" by chance.
- A standard p-value cutoff $\alpha = 0.05$ needs to be adjusted.

Why Multiple Testing Matters II

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive?
  - Assume that all the null hypotheses are true

Rejecting the null hypothesis when that hypothesis is true.

$$P(\text{Making an error}) = \alpha$$

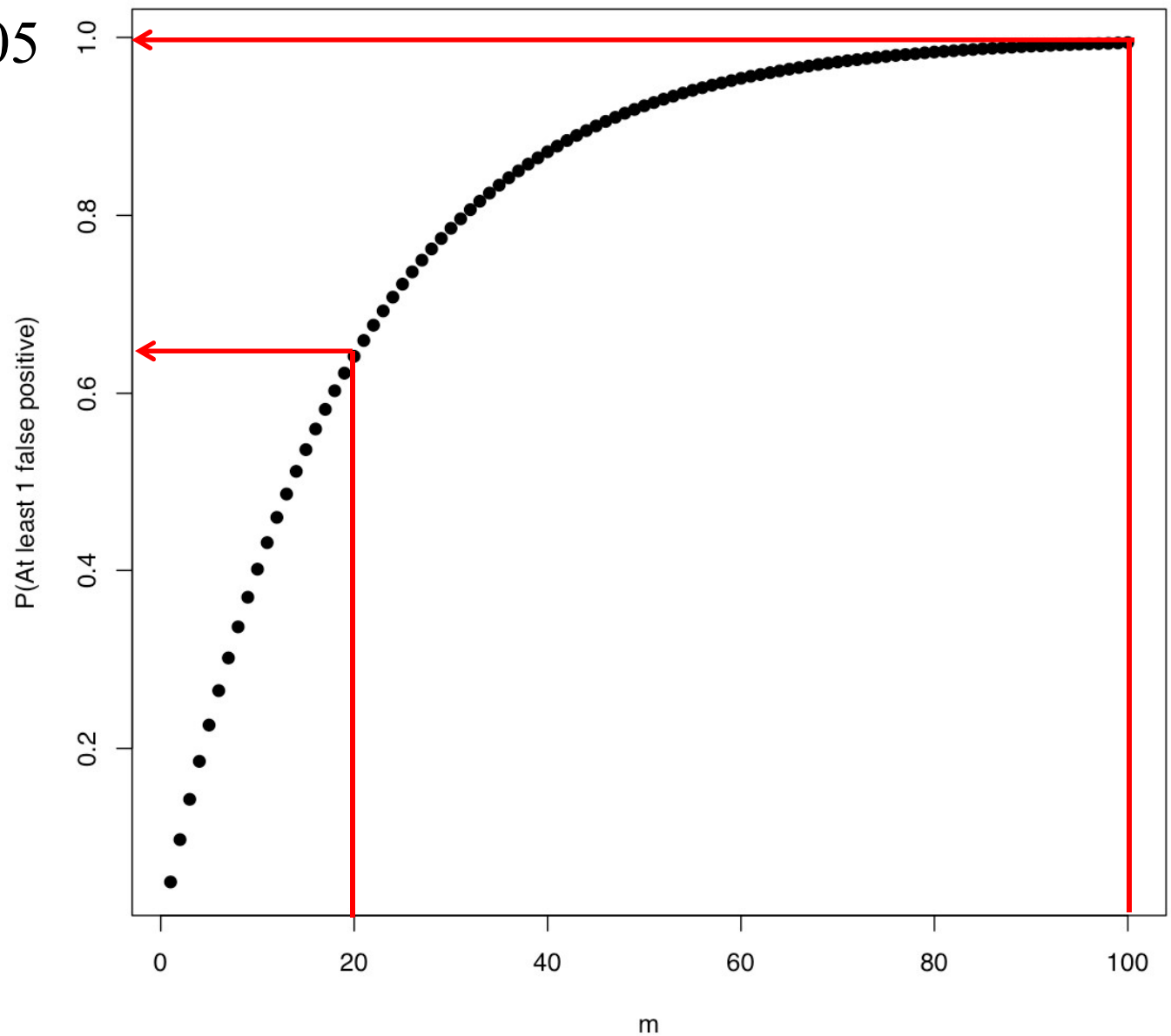
$$P(\text{Not making an error}) = 1 - \alpha$$

$$P(\text{Not making an error in } m \text{ tests}) = (1 - \alpha)^m$$

$$P(\text{Making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$

Probability of At Least 1 False Positive

When $\alpha = 0.05$



"Correcting" for Multiple Testing

- We need to *adjust p-values for the number of hypothesis tests performed"*

- Very active area of statistics – many different methods have been described

- Although these varied approaches have the same goal, they go about it in fundamentally different ways
  - *Our goal is to understand what errors each method controls, such that we know which method to use in research.*

- Many of them control the *Type I error*

Outline

- Why multiple hypothesis testing matters?
  - A motivating example
  - R-session
- Basic concepts on multiple hypothesis testing
  - Type I error and Type II error
  - Define the multiple testing problem and related concepts
- Methods for multiple hypothesis testing correction
  - 2 methods that control the family-wise error rate (FWER)
  - 1 method that controls the false discovery rate (FDR)

Type I and II Errors

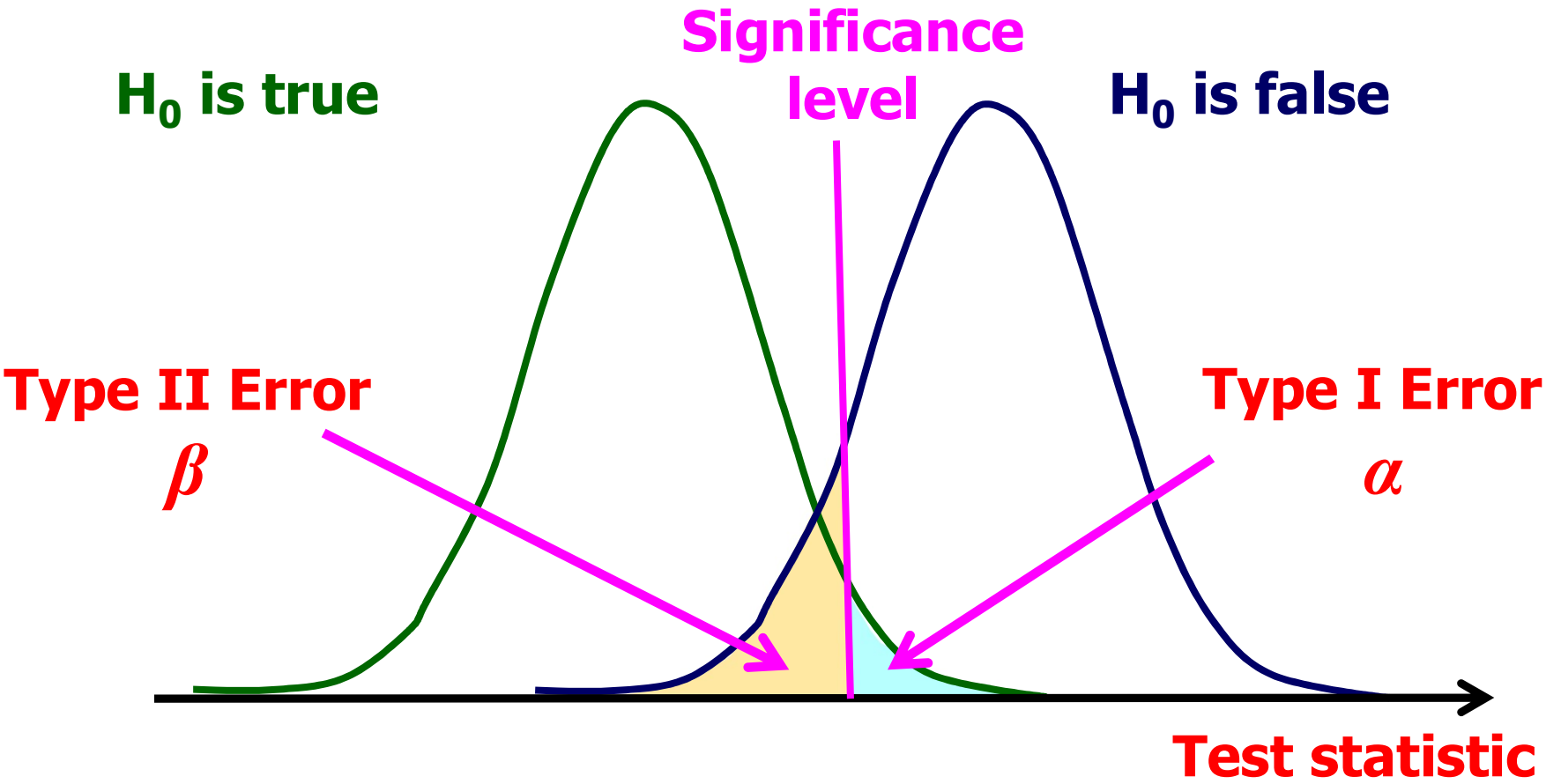
Decision	Actual Situation "Truth" H_0 True	H_0 False
Do Not Reject H_0	**Correct Decision (True Negative) 1-α**	**Incorrect Decision (False Negative) Type II Error β**
Reject H_0	**Incorrect Decision (False Positive) Type I Error α**	**Correct Decision (True Positive) 1-β**

α = P(Type I Error) β = P(Type II Error)

Power = 1 - β

Type I and Type II Errors

- Consider the distribution of your test statistic



Counting Errors

- Assume that we are testing m hypotheses: $H^1, \ldots, H^m$
  - m_0 = # of **true** null hypotheses
  - R = # of rejected null hypotheses

	Null True	**Alternative True**	**Total**
Not Called Significant	U	T	$m-R$
Called Significant	V	S	R
	m_0	$m - m_0$	m

  - V = # Type I errors [false positives]

Outline

- Why multiple hypothesis testing matters?
  - A motivating example
  - R-session
- Basic concepts on multiple hypothesis testing
  - Type I error and Type II error
  - Define the multiple testing problem and related concepts
- Methods for multiple hypothesis testing correction
  - 2 methods that control the *family-wise error rate*
  - 1 method that controls the false discovery rate (FDR)

Family-Wise Error Rate (FWER)

- Assume that we are testing *m* hypotheses: $H^1, \ldots, H^m$
- Family-Wise Error Rate (FWER)
  - Here, the term "family" refers to the collection of hypotheses $H^1, \ldots, H^m$
  - The probability of making one type I error among all the hypotheses, $P(V \geq 1)$
- Two general types of FWER corrections:
  - **Single step:** equivalent adjustments made to each p-value
  - **Sequential:** adaptive adjustment made to each p-value

Single Step Approach: Bonferroni

- By assuring FWER ≤ α, *the probability of making even one Type I error in the family is controlled at level* α.

- Adjust p-values as following:

$$\tilde{p}_j = \min[mp_j, 1]$$

- Rejects any hypothesis with p-value ≤ α/m.

- Example
  - Say that we want the probability of making at least one Type I error to be <0.05 when we perform 10,000 hypothesis tests.
  - Then, we need a p-value of 0.05/10,000 = 5 x 10^{-6} to declare significance

The Bonferroni correction controles the FWER

- Notations
  - $H^1, \ldots, H^m$: a family of hypotheses
  - $p_1, \ldots, p_m$: the corresponding p-values.
  - I_0 : set of (unknown)true null hypotheses, having m_0 members
- The FWER is the probability of rejecting at least one of the members in I_0
- The Bonferroni correction states that rejecting all $p_i \leq \alpha/m$ will control the FWER $\leq \alpha$

The proof is based on the Boole's inequality

- In probability theory, Boole's inequality says that *for any finite set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events*.

- Boole's inequality is named after George Boole.

- Formally, for a set of events, A_1, A_2,…, we have:

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

E.g., probability that it rains on at least one day this week.

E.g., sum of the probabilities that it rains each day.

The Bonferroni correction controles the FWER

- Notations
  - $H^1, \ldots, H^m$: a family of hypotheses
  - $p_1, \ldots, p_m$: the corresponding p-values.
  - I_0 : set of the true null hypotheses, having m_0 members

- The FWER is bounded above by α:

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\}$$

FWER = P(V≥1)
The probability that at least one of the true null hypotheses has a p-value ≤ α/m

The Bonferroni correction controles the FWER

- Notations
  - $H^1, \ldots, H^m$: a family of hypotheses
  - $p_1, \ldots, p_m$: the corresponding p-values.
  - I_0 : set of the true null hypotheses, having m_0 members

- The FWER is bounded above by α:

$$FWER = \boxed{Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\}}$$

The probability that *at least one of the true null hypotheses has a p-value* $\leq \alpha/m$

Sum of the probabilities that *each of the true null hypotheses has a p-value* $\leq \alpha/m$

The Bonferroni correction controles the FWER

- Notations
  - $H^1, \ldots, H^m$: a family of hypotheses
  - $p_1, \ldots, p_m$: the corresponding p-values.
  - I_0 : set of the true null hypotheses, having m_0 members

- The FWER is bounded above by α:

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\boxed{\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\}} \leq m_0\frac{\alpha}{m}$$

The probability that a true null hypothesis test gets a p-value $\leq \alpha/m = \alpha/m$

(see the definition of p-value on slide 8)

The Bonferroni correction controles the FWER

- Notations
  - $H^1, \ldots, H^m$: a family of hypotheses
  - $p_1, \ldots, p_m$: the corresponding p-values.
  - I_0 : set of the true null hypotheses, having m_0 members

- The FWER is bounded above by α:

$$FWER = Pr\left\{\bigcup_{I_o}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{I_o}\left\{Pr\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha$$

An alternative method

- Drawback of the Boferroni correction
  - It often leaves very few hypotheses that are deemed significant
- "Holm method" a.k.a "Holm-Bonferroni method"
- It is known to be more powerful than the Bonferroni method
  - More "lenient" correction than Bonferroni method
- Basic idea:
  - The Bonferroni correction p-value cut-off is α/m.
  - This could have been α/m_0, where m_0 is the number of true null hypotheses.
  - We do not know m_0 , but could estimate it.

FWER: Sequential Adjustments

- Simplest sequential method is Holm's Method
  - Order the unadjusted p-values such that $p_1 \leq p_2 \leq \ldots \leq p_m$
  - For control of the FWER at level α, the step-down Holm adjusted p-values are

$$\tilde{p}_j = \min[\boxed{(m-j+1)} \cdot p_j, 1]$$

  - The point here is that *we do not multiply every p_i by the same factor m*

- For example, when m = 10000:

$$\tilde{p}_1 = 10000 \cdot p_1, \quad \tilde{p}_2 = 9999 \cdot p_2, \cdots, \tilde{p}_m = 1 \cdot p_m$$

Philosophical Objections to Bonferroni Corrections

- "Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference" Perneger (1998)
- Counter-intuitive: interpretation of finding depends on the number of other tests performed (shared by all methods)
- The general null hypothesis (that all the null hypotheses are true) is rarely of interest
- High probability of Type II errors, i.e., of not rejecting the general null hypothesis when important effects exist

Who Cares About Not Making ANY Type I Errors?

- FWER is appropriate when you want to guard against ANY false positives
- However, in many cases (particularly in genomics) we can live with a certain number of false positives
- In these cases, the more relevant quantity to control is the **false discovery rate (FDR)**

$$\frac{\#\,\text{falsely rejected}}{\#\,\text{rejected in total}}$$

Outline

- Why multiple hypothesis testing matters?
  - A motivating example
  - R-session
- Basic concepts on multiple hypothesis testing
  - Type I error and Type II error
  - Define the multiple testing problem and related concepts
- Methods for multiple hypothesis testing correction
  - 2 methods that control the family-wise error rate
  - 1 method that controls the *false discovery rate (FDR)*

False Discovery Rate

	Null True	**Alternative True**	**Total**
Not Called Significant	***U***	***T***	***m-R***
Called Significant	***V***	***S***	***R***
	m_0	$m - m_0$	***m***

- ***V*** = # Type I errors [false positives]

- False discovery rate (FDR) is designed to control the proportion of false positives ***among the set of rejected hypotheses*** (R) -- V/R

Benjamini and Hochberg FDR

- To control FDR at level δ:
  - 1. Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$
  - 2. Then find the test with the highest rank, j, for which the p-value, p_j, is less than or equal to (j/m) x δ
  - 3. Declare the tests of rank 1, 2, …, j as significant

$$p(j) \leq \delta \frac{j}{m}$$

- Adjust p-values as following (now called q-values):

$$\tilde{p}_j = \min[\frac{m}{j} p_j, 1]$$

B&H FDR Example

- Controlling the FDR at $\delta = 0.05$

Rank (j)	P-value	(j/m) x δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Summary

- Why care about multiple hypoithesis testing?
- Controling the family-wise error rate (FWER)
  - Bonferroni correction
  - Holm method
- Controling the false discovery rate (FDR)
  - B-F FDR correction

Let's quickly do the following

- Load the data

```
a <- read.table(header = T,
file="http://www.cs.washington.edu/homes/suinlee/genome560/RMA_Filtered.txt")
b <- a[,2:33]
```

- Define a function of performing the t-test

```
fun <- function(d){return(t.test(d[1:16],d[17:32])$p.value)}
```

- Obtain the p-values `p <- apply(b, 1, fun)`
- See the distribution `hist(p, breaks=20)`
- Check how many are < 0.05 `tabulate(as.numeric(p<0.05))`

R commands

- tabulate(as.numeric(p<0.05))
  - tabulate & as.numeric
  - Counting how many elements are < 0.05
- as.numeric
  - What is the output of "p<0.05"
  - What is the output of "as.numeric(p<0.05)" ?
  - as.numeric (which is identical to as.double) coerces to the class
- tabulate
  - tabulate takes the integer-valued vector bin and counts the number of times each integer occurs in it.
- Alternatively, we can use a different command "???"

Let's apply the correction methods

- What methods are available?

```
p.adjust.methods
```

- Let's apply each of them

```
padj1 <- p.adjust(p, "bonferroni")
padj2 <- p.adjust(p, "holm")
padj3 <- p.adjust(p, "BH")
```

- How many of the adjusted p-values are less than 0.05?

```
alpha = 0.05
tabulate(as.numeric(padj1 < alpha) )
tabulate(as.numeric(padj2 < alpha) )
tabulate(as.numeric(padj3 < alpha) )
```

Let's compare the adjusted p-values!

- Plot the histograms of the original and adjusted p-values.

```
par( mfrow = c(1,2) )
hist( p, breaks = 50)
hist( padj1, breaks = 50)
```

- Let's plot the histogram of 4 sets of p-values.

```
par( mfrow = c(1,4) )
hist( p, breaks = 50)
hist( padj1, breaks = 50)
hist( padj2, breaks = 50)
hist( padj3, breaks = 50)
```

- Which method is the most lenient and which one is the most harsh adjustment?