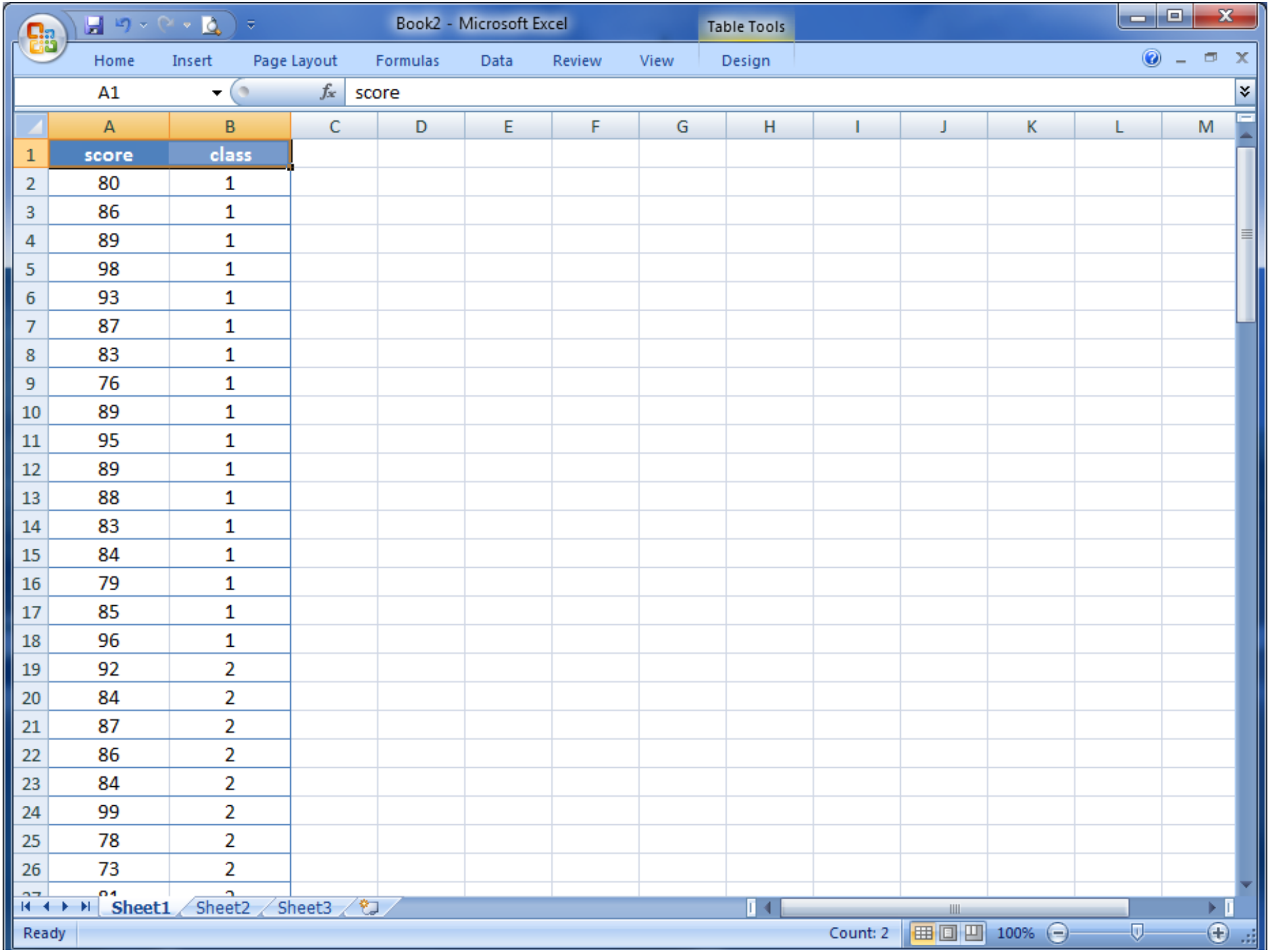
第8章 多重检验问题

感谢清华大学自动化系江瑞教授提供PPT



	A	B
1	score	class
2	80	1
3	86	1
4	89	1
5	98	1
6	93	1
7	87	1
8	83	1
9	76	1
10	89	1
11	95	1
12	89	1
13	88	1
14	83	1
15	84	1
16	79	1
17	85	1
18	96	1
19	92	2
20	84	2
21	87	2
22	86	2
23	84	2
24	99	2
25	78	2
26	73	2

Another presentation of the data

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Score 11	Score 21	Score 31	Score 41	Score 51	Score 61
		Score $3n_3$			
					Score $6n_6$
Score $1n_1$					
	Score $2n_2$				
			Score $4n_4$		
				Score $5n_5$	

Hypothesis testing

"Whether different classes have different scores?"

We need to test

H_0: $\theta_i = \theta_j$ for all i-j pairs versus H_1: $\theta_i \neq \theta_j$ for some i-j pair

Union-intersection tests (UIT)

If the null hypothesis is

$$H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$$

Then the rejection region is

$$R = \bigcup_{\gamma \in \Gamma} \{\mathbf{x} : T(\mathbf{x}) \in R_\gamma\}$$

R_γ is the rejection region for the hypothesis testing problem

$$H_{0\gamma} : \theta \in \Theta_\gamma \quad \text{versus} \quad H_{1\gamma} : \theta \in \Theta_\gamma^c$$

Pairwise tests

"Whether different classes have different scores?"

We need to test

H_0: $\theta_i = \theta_j$ for all i, j pairs versus H_1: $\theta_i \neq \theta_j$ for some i, j pair

We can
Run pair-wise two-sample t test over all possible combinations of the classes with the same hypotheses:

$$H_{0ij}: \theta_i = \theta_j \quad \text{versus} \quad H_{1ij}: \theta_i \neq \theta_j$$

and reject H_0 if **any** H_{0ij} is rejected at a certain significant level α (union-intersection test).

Real data

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	80	92	94	96	75	78
2	86	84	92	91	76	90
3	89	87	81	90	85	94
4	98	86	92	92	85	85
5	93	84	90	84	82	83
6	87	99	94	84	80	96
7	83	78	100	89	88	84
8	76	73	80	88	83	92
9	89	81	90	83	80	99
10	95	84	76	79	90	76
11	89	74	95	96	84	75
12	88	82	93	91	76	88
13	83	81	87	83	76	76
14	84	84	91	77	85	83
15	79	92	91	77	94	89
16	85	86		78	91	81
17	96	87		89	90	
18		92		100	94	
19				82	100	
20					84	
Mean	87.06	84.78	89.73	86.79	84.90	85.56
Std	6.10	6.47	6.33	6.77	6.84	7.43

Pairwise *t* test

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	0.291	-	-	-	-	-
Class 3	0.235	0.036	-	-	-	-
Class 4	0.901	0.362	0.201	-	-	-
Class 5	0.317	0.955	0.039	0.392	-	-
Class 6	0.533	0.746	0.103	0.616	0.785	-

```
> data <- read.table("/path/to/your/file.txt", header=T);
> pairwise.t.test(data$score,data$class,p.adj="none",pool.sd=F);
```

Any problem?

Error probability

The rejection region is

$$R = \{p_{11} \le \alpha \text{ OR } p_{12} \le \alpha \text{ OR } \cdots\} = \left\{\bigcup_{\text{all } i,j \text{ pair}} p_{ij} \le \alpha\right\}$$

That is to say

$$\begin{aligned} A = R^c = & \{p_{11} > \alpha \text{ AND } p_{12} > \alpha \text{ AND } \cdots\} \\ = & \left\{\bigcap_{\text{all } i,j \text{ pair}} p_{ij} > \alpha\right\} \\ = & \{\min p_{ij} > \alpha\} \end{aligned}$$

Therefore

$$\begin{aligned} P(\text{accept}|H_0) &= P(\min p_{ij} > \alpha) \\ P(\text{rejection}|H_0) &= 1 - P(\min p_{ij} > \alpha) \end{aligned}$$

Can we still obtain the original significance level (0.05)?

Bonferroni inequality

For two events

If P is a probability function, then

1. $P(A \cap B) \geq P(A) + P(B) - 1$;
2. $P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$ for any sets $A_1, A_2, \ldots$.

$$P\left(\bigcup_{i=1}^{n} A_i^c\right) \leq \sum_{i=1}^{n} P(A_i^c) \qquad \text{(Boole's inequality)},$$

$$P(A_i^c) = 1 - P(A_i), \qquad P\left(\bigcup_{i=1}^{n} A_i^c\right) = 1 - P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$1 - P\left(\bigcap_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} (1 - P(A_i)) = n - \sum_{i=1}^{n} P(A_i) \Rightarrow P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Error probability

The accept region is

$$A = R^c = \{p_{11} > \alpha \text{ AND } p_{12} > \alpha \text{ AND } \cdots\}$$
$$= \left\{\bigcap_{\text{all } i,j \text{ pair}} p_{ij} > \alpha\right\}$$
$$= \{\min p_{ij} > \alpha\}$$

Therefore

$$P(\text{accept}|H_0) = P(\min p_{ij} > \alpha)$$
$$= P(p_{ij} > \alpha, \text{ all } i,j \text{ pair})$$
$$\geq \sum_{i,j} P(p_{ij} > \alpha) - (\#(i,j \text{ pairs}) - 1) \quad \text{Bonferroni inequality}$$
$$\geq \sum_{i,j} (1-\alpha) - (\#(i,j \text{ pairs}) - 1) \qquad P(p_{ij} > \alpha) \geq 1-\alpha$$
$$= m(1-\alpha) - (m-1) \qquad (m = \#(i,j \text{ pair}))$$
$$= 1 - m\alpha$$

$$P(\text{reject}|H_0) = 1 - P(\text{accept}|H_0) \leq m\alpha$$

What does this mean?

$$P(\text{reject}|H_0) \leq m\alpha$$

The type-I error probability could be much larger than expected (***mα*** instead of ***α***).

To correct this problem, an simple choice is to run pair-wise two-sample t test over all possible combinations of the classes with the same hypotheses:

$$H_{0ij}: \theta_i = \theta_j \qquad \text{versus} \qquad H_{1ij}: \theta_i \neq \theta_j$$

but reject H_0 **if any H_{0ij} is rejected at a more stringent significance level, say, *α/m*. Equivalently, multiply the *p*-values of each H_{0ij} by *m*.**

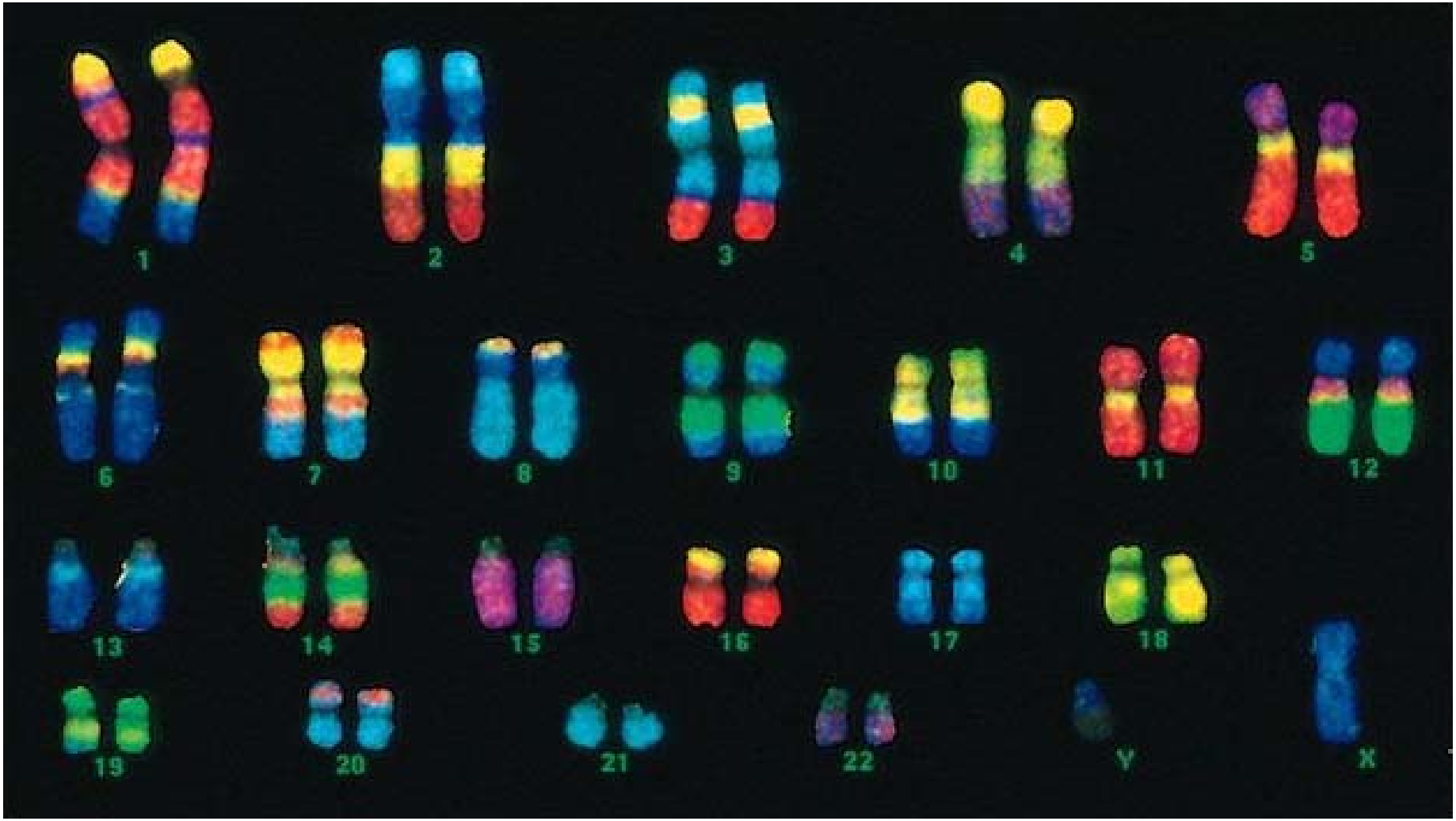
Bonferroni correction

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	1.00	-	-	-	-	-
Class 3	1.00	0.51	-	-	-	-
Class 4	1.00	1.00	1.00	-	-	-
Class 5	1.00	1.00	0.58	1.00	-	-
Class 6	1.00	1.00	1.00	1.00	1.00	-

```
> pairwise.t.test(data$score,data$Class,p.adj="bonf",pool.sd=F);
```

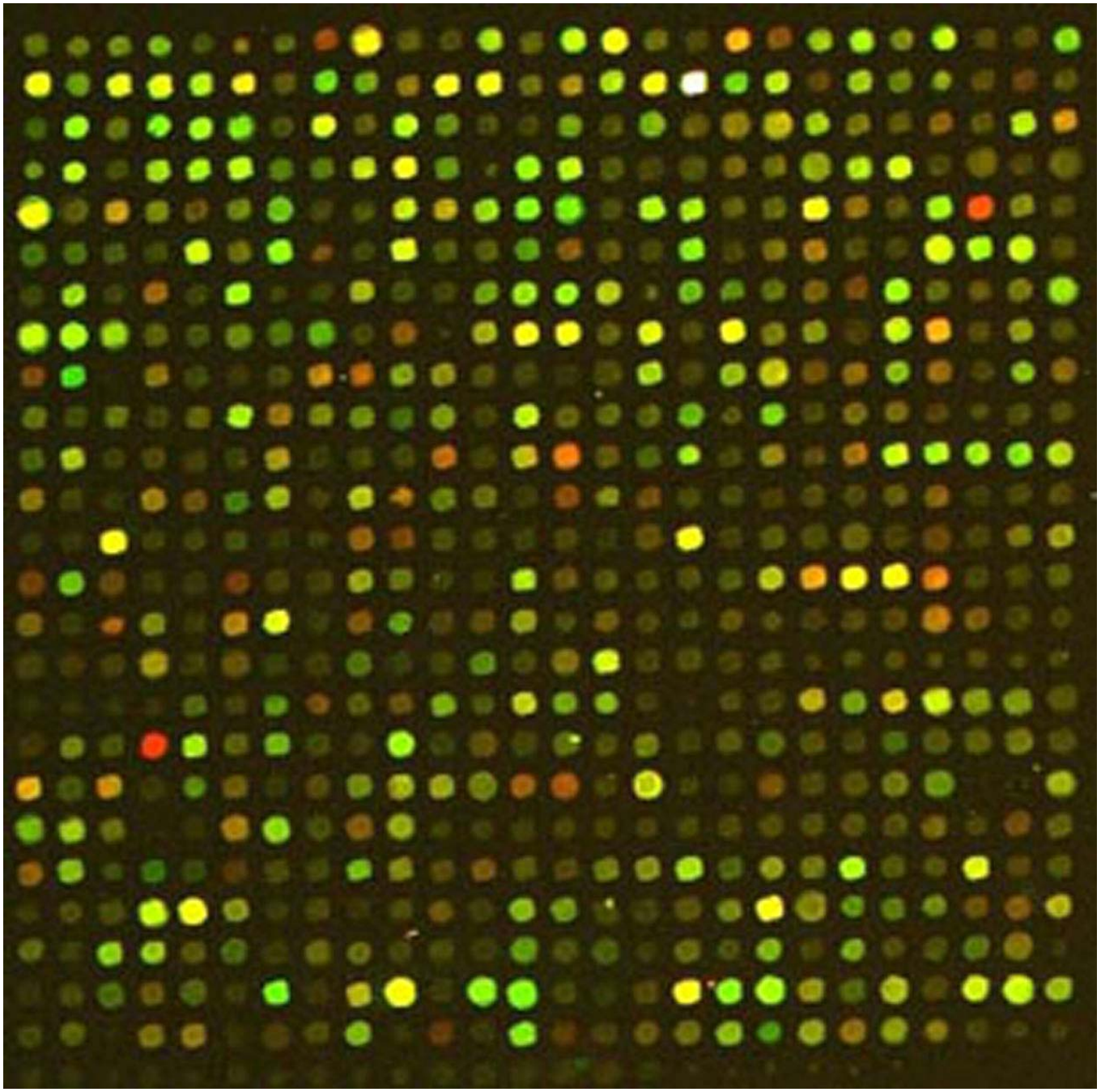
Bonferroni correction is very stringent (too conservative).

Each *p*-value is multiplied by 15, the total number of tests performed.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Y
X

Chromosome	Genes	Total bases	Sequenced bases
1	4,220	247,199,719	224,999,719
2	1,491	242,751,149	237,712,649
3	1,550	199,446,827	194,704,827
4	446	191,263,063	187,297,063
5	609	180,837,866	177,702,766
6	2,281	170,896,993	167,273,993
7	2,135	158,821,424	154,952,424
8	1,106	146,274,826	142,612,826
9	1,920	140,442,298	120,312,298
10	1,793	135,374,737	131,624,737
11	379	134,452,384	131,130,853
12	1,430	132,289,534	130,303,534
13	924	114,127,980	95,559,980
14	1,347	106,360,585	88,290,585
15	921	100,338,915	81,341,915
16	909	88,822,254	78,884,754
17	1,672	78,654,742	77,800,220
18	519	76,117,153	74,656,155
19	1,555	63,806,651	55,785,651
20	1,008	62,435,965	59,505,254
21	578	46,944,323	34,171,998
22	1,092	49,528,953	34,893,953
X (sex chromosome)	1,846	154,913,754	151,058,754
Y (sex chromosome)	454	57,741,652	25,121,652


AFFYMETRIX
GeneChip®
Human Genome
U133 Plus 2.0
P/N: 520019
Lot #: 4010008
Exp. Date: 05/07/06
For Research Use Only
AFFYMETRIX
GeneChip®
Mouse Genome
430 2.0 Array
P/N: 520029
Lot #: 4015112
Exp. Date: 11/24/06
For Research Use Only
Microarray



Whether a gene is differentially expressed

Gene	Normal cell				Cancer cell				*p*-value
	I_1	I_2	…	I_k	I_1	I_2	…	I_l	
1	e_{11}	e_{12}	…	e_{1k}	f_{11}	f_{12}	…	f_{1l}	**0.0001**

Do a two-sample *t* test, and then set a *p*-value cutoff 0.05.

Two random samples $X_1,\ldots X_m$ and $Y_1,\ldots Y_n$ are obtained from **two normal** populations $N(\mu_X,\sigma_X^2)$ and $N(\mu_Y,\sigma_Y^2)$, respectively, where σ_X^2 and σ_Y^2 are **unknown** but an assumption that $\sigma_X^2=\sigma_Y^2$ holds. We like to test

$$H_0 : \mu_X = \mu_Y \qquad \text{versus} \qquad H_1 : \mu_X \neq \mu_Y$$

In a single test

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	Correct rejection	Type I error (0.05) $1 \times 0.05 = 0.05$
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

$$\alpha = 0.05$$

Probability of Type I error = 0.05 $\Leftrightarrow$
We expect **0.05** genes out of one to be statistically significant by chance.

Gene selection

Gene	Normal cell				Cancer cell				p-value
	I_1	I_2	…	I_k	I_1	I_2	…	I_l	
1	e_{11}	e_{12}	…	e_{1k}	f_{11}	f_{12}	…	f_{1l}	**0.0001**
2									0.0888
3									0.6984
4									0.3276
…									
…									
19997									**0.0300**
19998									0.8743
19999									**0.0499**
20000									0.8498

Do a two-sample t test for each gene, and then set a p-value cutoff 0.05?

In a total of 20000 tests

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	Correct rejection	Type I error (0.05) $20000 \times 0.05 = 1000$
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

$$\alpha = 0.05$$

Probability of Type I error = 0.05 $\Leftrightarrow$
We expect **1000** genes out of 20000 to be statistically significant by chance!

But we like to select, e.g., at most **1000** genes! **(Too many rejections)**

Family-wise error rate (FWER)

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	True Positive (TP)	False Positive (FP)
	Accept H_0 ($X \in R^c$)	False Negative (FN)	True Negative (TN)

$$FWER = P(FP \ ? \ 1) = 1 - P(FP \quad 0)$$

Family-wise error rate (FWER) is
the probability of making one or more false positives,
or type I errors, among all the hypotheses,
when performing multiple tests.

Bonferroni inequality says

$$
\begin{aligned}
P(FP = 0) &= P(p_i > \alpha,\ i = 1,\ldots,m) \\
&\geq \sum_{i=1}^{m} P(p_i > \alpha) - (m-1) && \text{Bonferroni inequality} \\
&\geq \sum_{i=1}^{m} (1-\alpha) - (m-1) && P(p_i > \alpha) \geq 1-\alpha \\
&= m(1-\alpha) - (m-1) && (m = \#(\text{tests})) \\
&= 1 - m\alpha
\end{aligned}
$$

$$
P(FP \geq 1) = 1 - P(FP = 0) \leq m\alpha
$$

In other words, to ensure

$$
P(FP \geq 1) \leq \alpha
$$

We need to reject each null at a more strigent significant level α / m, such that

$$
P(p_i \leq \alpha / m) \leq \alpha / m \text{ for } i = 1,\ldots,m
$$

equivalently, we can multiply each p_i by m and reject at threshold α.

Bonferroni correction

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	Correct rejection	FWER (0.05) $20000 \times 0.0000025 = 0.05$
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

FWER = 0.05

Family-wise error rate (FWER) = 0.05 $\Leftrightarrow$

We expect **0.05** genes out of 20000 to be statistically significant by chance.

But α **= 0.0000025** is such a stringent threshold!

Probability of Type II error must be very high! **(Too few rejections)**

False discovery rate (FDR)

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	True Discovery	False Discovery
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

$$FDR = \mathrm{E}\left[\frac{FD}{TD + FD}\right]$$

Control FDR so that it is less than a given threshold value.

(Medium number of rejections)

positive False discovery rate (*p*FDR)

Hypothesis testing procedure		Truth	
		H_1 ($\theta \in \Theta_0^c$)	H_0 ($\theta \in \Theta_0$)
Decision	Reject H_0 ($X \in R$)	True Discovery	False Discovery
	Accept H_0 ($X \in R^c$)	Type II error	Correct acceptance

$$pFDR = 0.05$$

Positive false discovery rate (*p*FDR) =
E(#{False discovery} / #{All discovery} | All discovery > 0)
The rate that discoveries are false.
Control *p*FDR so that it is at most 0.05. **(Even more reasonable)**

Bonferroni correction (for FWER)

The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant:

$$\textbf{Corrected } p\textbf{-value} = p\textbf{-value} \times n$$

where n is the number of genes in the test.

As a consequence, if testing 20000 genes at a time, the highest accepted individual p-value is 0.0000025, making the correction very stringent. With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.

Bonferroni step-down (Holm) correction (for FWER)

This correction is very similar to the Bonferroni, but a little less stringent:

1. The *p*-value of each gene is ranked from the smallest to the largest.
2. The first *p*-value is multiplied by the number of genes present in the gene list: if the end value is less than 0.05, the gene is significant:

   $$\textbf{Corrected } p\textbf{-value} = p\textbf{-value} \times n \qquad (< 0.05)$$

3. The second *p*-value is multiplied by the number of genes less 1:

   $$\textbf{Corrected } p\textbf{-value} = p\textbf{-value} \times (n-1) \qquad (< 0.05)$$

4. The third *p*-value is multiplied by the number of genes less 2:

   $$\textbf{Corrected } p\textbf{-value} = p\textbf{-value} \times (n-2) \qquad (< 0.05)$$

It follows that sequence until no gene is found to be significant.

Westfall and Young Permutation (for FWER)

The Westfall and Young permutation follows a step-down procedure similar to the Holm method, combined with a bootstrapping method to compute the *p*-value distribution:

1. *p*-values are calculated for each gene based on the original data set and ranked from the smallest to the largest.
2. The permutation method creates a pseudo-data set by dividing the data into artificial treatment and control groups.
3. *p*-values for all genes are computed on the pseudo-data set.
4. The **successive minima** of the new *p*-values are retained and compared to the original ones.
5. This process is repeated a large number of times, and the proportion of resampled data sets where the minimum pseudo-*p*-value is less than the original *p*-value is the adjusted *p*-value.

This method has a similar Family-wise error rate as the Bonferroni and Holm correction. Because of the permutations, the method is **very slow**.

Benjamini and Hochberg correction (for FDR)

This correction will yield less false negative genes. Here is how it works:

1. The *p*-values are ranked from the smallest to the largest.
2. For a given significance level α, find the largest k such that

$$P_{(k)} \leq \frac{k}{m}\alpha$$

3. Reject all null hypotheses for $i = 1,\ldots,k$

Cited by 20188

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

John storey's *p*FDR

Based on a Bayesian model, a little complicated.

Please Google

"John storey"

"q-value"

for details.

Whether different classes have different scores?

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	80	92	94	96	75	78
2	86	84	92	91	76	90
3	89	87	81	90	85	94
4	98	86	92	92	85	85
5	93	84	90	84	82	83
6	87	99	94	84	80	96
7	83	78	100	89	88	84
8	76	73	80	88	83	92
9	89	81	90	83	80	99
10	95	84	76	79	90	76
11	89	74	95	96	84	75
12	88	82	93	91	76	88
13	83	81	87	83	76	76
14	84	84	91	77	85	83
15	79	92	91	77	94	89
16	85	86		78	91	81
17	96	87		89	90	
18		92		100	94	
19				82	100	
20					84	
Mean	**87.06**	**84.78**	**89.73**	**86.79**	**84.90**	**85.56**
Squared error	**37.18**	**41.83**	**40.07**	**45.84**	**46.83**	**55.20**
Pooled error	**44.54**					

Pairwise *t* test

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	0.291	-	-	-	-	-
Class 3	0.235	0.036	-	-	-	-
Class 4	0.901	0.362	0.201	-	-	-
Class 5	0.317	0.955	0.039	0.392	-	-
Class 6	0.533	0.746	0.103	0.616	0.785	-

```
> data <- read.table("/path/to/your/file.txt", header=T);
> pairwise.t.test(data$score,data$class,p.adj="none",pool.sd=F);
```

Multiple testing without correction may produce incorrect conclusion!

Pairwise *t* test without correction

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	0.315	-	-	-	-	-
Class 3	0.261	0.036	-	-	-	-
Class 4	0.904	0.362	0.205	-	-	-
Class 5	0.329	0.955	0.036	0.379	-	-
Class 6	0.521	0.733	0.085	0.589	0.768	-

```
> pairwise.t.test(data$score,data$class,p.adj="none",pool.sd=T);
```

Multiple testing without correction may produce incorrect conclusion!

Pairwise *t* test with Bonferroni correction

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	1.00	-	-	-	-	-
Class 3	1.00	0.54	-	-	-	-
Class 4	1.00	1.00	1.00	-	-	-
Class 5	1.00	1.00	0.55	1.00	-	-
Class 6	1.00	1.00	1.00	1.00	1.00	-

```
> pairwise.t.test(data$score,data$Class,p.adj="bonf",pool.sd=T);
```

Bonferroni correction is too stringent.
Each *p*-value is multiplied by 15, the total number of tests performed.

Pairwise *t* test with Bonferroni step-down (Holm) correction

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	1.00	-	-	-	-	-
Class 3	1.00	0.54	-	-	-	-
Class 4	1.00	1.00	1.00	-	-	-
Class 5	1.00	1.00	0.54	1.00	-	-
Class 6	1.00	1.00	1.00	1.00	1.00	-

```
> pairwise.t.test(data$score,data.Class,p.adj="holm",pool.sd=T);
```

Bonferroni step-down correction is less stringent than Bonferroni correction.

Pairwise *t* test with Benjamini and Hochberg correction

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Class 1	-	-	-	-	-	-
Class 2	0.63	-	-	-	-	-
Class 3	0. 63	0.27	-	-	-	-
Class 4	0.96	0.63	0.63	-	-	-
Class 5	0. 63	0.96	0.27	0.63	-	-
Class 6	0.78	0.89	0.43	0.80	0.89	-

```
> pairwise.t.test(data$score,data$Class,p.adj="BH", pool.sd=T);
> pairwise.t.test(data$score,data$Class,p.adj="fdr",pool.sd=T);
```

Benjamini and Hochberg correction is even less stringent.

In summary

Bonferroni
Bonferroni step-down (Holm)
Westfall and Young Permutation
Benjamini and Hochberg (FDR)
John Storey (*p*FDR)
None