

# Statistics 514: Design of Experiments

## Topic 7

### Topic Overview

This topic will cover

- Testing contrasts
- Multiple comparisons
- Error rates in multiple testing
- $P$ -value methods

### One-Sample Hypothesis Test of Mean (Review)

- **Data:**  $Y_1, Y_2, \dots, Y_n \sim N(\mu, \sigma^2)$  ( $\sigma^2$  known)

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- **Test Statistic:**

$$Z_0 = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

- **Decision Rule:** Reject  $H_0$  if  $Z_0 > Z$ .
- **Type I Error:**  $Pr(\text{reject } H_0 | H_0)$ .
- **Type II Error:**  $Pr(\text{accept } H_0 | H_0 \text{ is false})$ .
- **Power:**  $1 - Pr(\text{Type II error})$

### Linear Combinations of Means

$$\begin{aligned} y_{i,j} &= \mu + \tau_i + e_{i,j} \\ &= \mu_i + e_{i,j} \end{aligned}$$

- Often hypotheses of interest different from  $H_0 : \text{all } \tau_i = 0$ .
- Would like to test  $H_0 : L = \sum c_i \mu_i = L_0$ 
  - Pairwise comparisons

- Treatments vs control
  - Comparing combinations of treatments
  - (Curvi)linear relationship between means and treatments
- Hypotheses may be planned or “after the fact.”

## Estimation and Test Statistic

$$\begin{aligned}\hat{L} &= \sum c_i \bar{y}_i \\ \text{Var}(\hat{L}) &= \text{Var}\left(\sum c_i \bar{y}_i\right) \\ &= \sum c_i^2 \text{Var}(\bar{y}_i) \\ &= \sigma^2 \sum \frac{c_i^2}{n_i} \\ SE_{\hat{L}} &= \sqrt{MS_E \sum c_i^2 / n_i} \\ t_0 &= \frac{\hat{L} - L_0}{\sqrt{\text{Var}(\hat{L})}}\end{aligned}$$

Under  $H_0 : t_0 \sim t_{N-a}$

## Contrasts

- $\Gamma = \sum c_i \mu_i$  is a contrast if  $\sum c_i = 0$ .
- Example:
  - To compare Treatment 1 and Treatment 2:  $c = \{1, -1, 0, \dots, 0\}$
- Two contrasts  $\{c_i\}$  and  $\{d_i\}$  are *orthogonal* if  $\sum c_i d_i / n_i = 0$ .
- By Cochran’s Theorem, estimates based on orthogonal contrasts are independent.
- Other possible decompositions (linear effect, quadratic effect).

## SAS Procedures (cont.sas)

Example 3-6

```
options ls=80;
title1 'Contrast Comparisons';
data one;
```

```
infile 'h:\System\Desktop\tensile.dat';
input percent strength time;
```

```
proc glm data=one;
class percent;
model strength=percent;
contrast 'C1' percent 0 0 0 1 -1;
contrast 'C2' percent 1 0 1 -1 -1;
contrast 'C3' percent 1 0 -1 0 0;
contrast 'C4' percent 1 -4 1 1 1;
```

-----

Dependent Variable: strength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

R-Square	Coeff Var	Root MSE	strength Mean
0.746923	18.87642	2.839014	15.04000

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
C1	1	291.6000000	291.6000000	36.18	<.0001
C2	1	31.2500000	31.2500000	3.88	0.0630
C3	1	152.1000000	152.1000000	18.87	0.0003
C4	1	0.8100000	0.8100000	0.10	0.7545

## Issues

- Why *t*-test instead of overall *F*-test?
  - *t*-test specifically addresses hypothesis of interest
  - Overall *F*-test *jointly* tests all possible contrasts (important in assessing “importance” of variable with > 1 factor).
  - Why include contrasts of no interest in a test?
    - If overall error rate equal to  $\alpha$ ,
    - Error rate for single comparison  $\leq \alpha$  (sharp inequality).
    - F*-test will reduce power (“water down” test)
    - May find individual test significant while *F* not
  
- Problems with linear combinations
  - *Multiple tests* inflate “overall” error rate
  - Not all combinations independent

# Testing Multiple Contrasts

- If  $m$  orthogonal (independent) contrasts

$$Pr(\text{at least one type I error}) = 1 - (1 - \alpha')^m$$

Can control overall error rate using equation

**Example:** Consider  $m = 5$  independent tests and overall error rate 0.05 .  
For this to hold, each test must use

$$\alpha' = 1 - (1 - 0.05)^{1/5} = 0.012$$

- Scheffé's Method
  - Set up  $1 - \alpha$  simultaneous CI for all contrasts
  - Protects for unplanned comparisons
  - Overall error rate at most  $\alpha \leftrightarrow$  low power
  - Uses  $SS_C/\sigma^2 \sim \chi_{a-1}^2$  instead of  $\chi_1^2$ .
  - Compare  $|C|$  to  $SE_{\hat{L}}\sqrt{(a-1)F_{\alpha, a-1, N-a}}$
  - Relationship with overall  $F$ -test: if  $p$ -value of  $F$ -test is  $\gamma$ , then Scheffé method will find no contrasts significant for  $\alpha$  level less than  $\gamma$ .

## Comparison of Means

- Often only interested in pairwise comparisons
- Can be expressed as contrasts
- If comparing Treatment  $j$  to Treatment  $k$

$$c_i = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

- Pairwise comparisons a subset of all contrasts
- Want test to consider only that subset
- Usually compare  $|\bar{y}_i - \bar{y}_j|$  with a **Critical Difference** (CD) to declare significance.
- Trade-off between power and  $Pr(\text{Type I Error})$ .
- No one “best” test
- Comparison methods vary in protection
  - Experimentwise error (overall Type I)
  - Comparisonwise error (individual Type I)

## Pairwise Comparison Methods

- Fisher's Least Significant Difference (LSD)
  - Use  $MS_E$  and df and performs usual  $t$ -test

$$t_0 = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MS_E \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t_{N-a}$$

to compare  $\mu_i, \mu_j$ .

- If  $|t_0| > LSD$  (based on  $t$  distribution), conclude means differ.
- Does not control experimentwise error, controls  $\alpha'$ . Can greatly inflate experimentwise errors.
- Still might be the case that none of the tests comes out significant if  $F$  is significant. (Some other contrast might be.)
- Protected Least Significant Difference (LSD)
  - Only do comparison if  $F$ -test significant
- Tukey's (Tukey-Kramer) Method
  - Sometimes called Honest Significance Distance Test
  - Uses studentized range distribution  $q$  instead of  $t$ .
  - Distribution based on  $\bar{y}_{max} - \bar{y}_{min}$  in  $a$  trts.
  - Assuming group sizes  $n_i$  are the same, choose  $q_{m,\nu,\alpha/2}$  such that

$$Pr\left(\max_{1 \leq t \leq m} \bar{y}_t - \min_{1 \leq t \leq m} \bar{y}_t \geq q se_E / \sqrt{n}\right) = \alpha,$$

(where  $\nu$  is the df of  $MS_E$ ) under the null hypothesis with underlying normal distribution.

- (Tukey-Kramer): If sample size  $n_i$ 's different in each group, can approximate  $q$  with  $\frac{q}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ .
- Accounts for any possible pair being selected
- Controls overall experimentwise error rate  $\alpha$ .
- Reject if ( $q$  available in Table VI)

$$|\bar{y}_i - \bar{y}_j| > \frac{q_{\alpha,a,N-a}}{\sqrt{2}} \sqrt{MS_E \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

## Other Methods

- Bonferroni Confidence Intervals for Contrasts

$$\sum c_j^{(i)} \bar{y}_j. \pm t_{\alpha/2m}(N - a) \sqrt{MS_E \sum \frac{c_j^2}{n_j}}$$

- For planned comparisons only
- Looks only at subset of contrasts
- Extremely conservative (with respect to overall error rate  $\alpha'$ ) if  $m$  is large
- Generalizes easily to case where comparing importance of factors

- Duncan's Multiple Range Test

- Sort  $\{\bar{y}_i.\}_{i=1}^m: \bar{y}_{i(1)}. \leq \bar{y}_{i(2)}. \leq \dots \leq \bar{y}_{i(m)}.$
- For  $|\bar{y}_{i(j)}. - \bar{y}_{i(k)}.|$ , where  $j < k$ :

$$CD = r_{\alpha,p,f} \sqrt{MS_E/n},$$

where  $r_{\alpha,p,f}$  come from table.

- If  $(i_j, i_k)$  is not significant, skip the pairs nested within them.
- When sample sizes unequal,  $n = m / \sum 1/n_i$ .
- Does not control overall error rate.

- Newman-Keuls Test

- Similar testing approach to Duncan's scheme.
- Based on Tukey's test.
- $CD = q_{\alpha,p,N-m} \sqrt{\frac{MS_C}{n}}$ .
- Controls FDR.

- Dunnett's Test

- Specifically designed for trts vs control situation.
- Distribution based on  $\max(\bar{y}_{control} - \bar{y}_{trt})$  for  $m - 1$  treatments.
- Similar in approach to Tukey's test:

$$CD = d_{\alpha}(m - 1, N - m) \sqrt{MS_E(1/n_i + 1/n_1)},$$

where  $d_{\alpha}(p, f)$  can be found in Table VIII.

- Controls experimentwise error rate.

# SAS Procedures

```
options ls=80;

title1 'Means Comparison';

data one;
  infile 'h:\System\Desktop\tensile.dat';
  input percent strength time;

proc glm data=one;
  class percent;
  model strength=percent;
  means percent / alpha=.05 lines bon snk tukey;
  means percent / lines duncan lsd scheffe;
  means percent / dunnett;
  means percent / lsd clm;
run;
```

## Student-Newman-Keuls Test for strength

NOTE: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

	Alpha	0.05		
Number of Means	2	3	4	5
Critical Range	3.7454539	4.5427095	5.0256316	5.3729604

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	percent
A	21.600	5	30
B	17.600	5	25
B	15.400	5	20
C	10.800	5	35
C			
C	9.800	5	15

## Tukey's Studentized Range (HSD) Test for strength

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	5.373

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	percent
A	21.600	5	30
A			
B A	17.600	5	25
B			
B C	15.400	5	20
C			
D C	10.800	5	35
D			
D	9.800	5	15

Bonferroni (Dunn) t Tests for strength

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Critical Value of t	3.15340
Minimum Significant Difference	5.6621

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	percent
A	21.600	5	30
A			
B A	17.600	5	25
B			
B C	15.400	5	20
C			
C	10.800	5	35
C			
C	9.800	5	15

t Tests (LSD) for strength

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Critical Value of t	2.08596
Least Significant Difference	3.7455

Means with the same letter are not significantly different.

t Grouping	Mean	N	percent
A	21.600	5	30
B	17.600	5	25
B			

B	15.400	5	20
C	10.800	5	35
C			
C	9.800	5	15

Duncan's Multiple Range Test for strength

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	8.06

Number of Means	2	3	4	5
Critical Range	3.745	3.931	4.050	4.132

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	percent
A	21.600	5	30
B	17.600	5	25
B			
B	15.400	5	20
C	10.800	5	35
C			
C	9.800	5	15

Scheffe's Test for strength

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Critical Value of F	2.86608
Minimum Significant Difference	6.0796

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	percent
A	21.600	5	30
A			
B A	17.600	5	25
B			
B C	15.400	5	20
C			
C	10.800	5	35
C			
C	9.800	5	15

Dunnett's t Tests for strength

NOTE: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha 0.05  
 Critical Value of Dunnett's t 2.65112  
 Minimum Significant Difference 4.7602

Comparisons significant at the 0.05 level are indicated by \*\*\*.

percent	Difference Between Means	Simultaneous 95% Confidence Limits		
30 - 15	11.800	7.040	16.560	***
25 - 15	7.800	3.040	12.560	***
20 - 15	5.600	0.840	10.360	***
35 - 15	1.000	-3.760	5.760	

t Confidence Intervals for strength

Alpha 0.05  
 Critical Value of t 2.08596  
 Half Width of Confidence Interval 2.648434

percent	N	Mean	95% Confidence Limits	
30	5	21.600	18.952	24.248
25	5	17.600	14.952	20.248
20	5	15.400	12.752	18.048
35	5	10.800	8.152	13.448
15	5	9.800	7.152	12.448

## Comparison with Best (Hsu 1996)

- Assumes “highest” is “best”. (What if lowest is best?)
- Other procedures give narrower intervals but do not give information about how much better the best group is than other groups.

- **Procedure**

1. Identify group with largest sample mean.
2. Use Dunnett's method to form one-sided confidence intervals (see SAS code), comparing group with largest sample mean to each of the other  $a - 1$  groups.
  - Provides confidence intervals for all groups except the group with the largest mean.
3. *To obtain endpoints for group with largest mean:* Identify group with the second largest mean. Reverse the signs of the lower and upper endpoints of the Dunnett confidence intervals comparing this to the group with the largest mean.

- If interval for group with largest mean does not contain 0, group is deemed “best treatment.” If not, all groups with intervals containing 0 included in a set of groups, any one of which is possibly the best treatment. Any group whose interval is entirely below 0 is ruled out as the best treatment.

## Example

In `tensile.dat` data, percent level 30 has largest mean (21.6).

```
proc glm data = one;
  class percent;
  model strength=percent;
  means percent / dunnett('30');
  means percent / dunnett1('30');
```

percent		Difference			
Comparison		Between	Means	Simultaneous 95% Confidence Limits	
25	- 30	-4.000		-8.138	Infinity
20	- 30	-6.200		-10.338	Infinity
35	- 30	-10.800		-14.938	Infinity
15	- 30	-11.800		-15.938	Infinity

percent		Difference			
Comparison		Between	Means	Simultaneous 95% Confidence Limits	
25	- 30	-4.000		-Infinity	0.138
20	- 30	-6.200		-Infinity	-2.062 ***
35	- 30	-10.800		-Infinity	-6.662 ***
15	- 30	-11.800		-Infinity	-7.662 ***

- Confidence Intervals ( $\mu_i - \mu_{best} = \mu_i - \max_{j \neq i} \mu_j$ )

$$\mu_{25} - \mu_{best} : (-8.138, 0.138)$$

$$\mu_{20} - \mu_{best} : (-10.338, 0)$$

$$\mu_{35} - \mu_{best} : (-14.938, 0)$$

$$\mu_{15} - \mu_{best} : (-15.938, 0)$$

$$\mu_{30} - \mu_{best} : (-0.138, 8.138)$$

- Possible “best treatments”: 30, 25
- Ruled out as possible “best treatment”: 20, 35, 15

## Flow Chart for Choosing FWER-controlling Procedure

- Is goal to identify the best treatment?

- **Yes:** Use Hsu’s method.
  - **No:** Go to Step 2.
2. Contrasts between  $a - 1$  groups vs control group?
- **Yes:** Use Dunnett’s method.
  - **No:** Go to Step 3.
3. Testing all pairwise and no “complex” comparisons (either planned or post-hoc) or choosing to test only some pairwise comparisons post-hoc?
- **Yes:** Use Tukey’s method.
  - **No:** Go to Step 4.
4. Are all comparisons planned?
- **No:** Use Scheffé’s method.
  - **Yes:** Go to Step 5.
5. Is Bonferroni critical value less than Scheffé critical value?
- **Yes:** Use Bonferroni’s method.
  - **No:** Go to Step 6.
6. Use Scheffé’s method (or, prior to collecting the data, reduce the number of contrasts to be tested)

## General Multiple Testing Setting

- $m$  tests:

$$\begin{array}{rcl}
 H_{0,1} & \text{vs} & H_{1,1} \\
 H_{0,2} & \text{vs} & H_{1,2} \\
 & \vdots & \\
 H_{0,m} & \text{vs} & H_{1,m}
 \end{array}$$

With  $m_0$  and  $m_1$  fixed and unknown, a given test will allocate in the following way:

	FTR	Reject	Total
Null True	$U$	$V[\text{I}]$	$m_0$
Alt True	$T[\text{II}]$	$S$	$m_1$
	$W$	$R$	$m$

- Usual hypothesis test ( $m = 1$ )

$$\begin{aligned} &\text{Try to maximize } E(S|m_1 > 0) \\ &\text{such that } E(V|m_0 > 0) \leq \alpha \end{aligned}$$

Type I error:  $Pr(V > 0|m_0 = 1)$

Power:  $E(S|m_1 = 1)$

## What if $m > 1$ ?

### Case of $m$ (Independent) Level- $\alpha$ Tests

$$Pr(\text{"false alarm"}) = 1 - (1 - \alpha')^m$$

Could control overall error rate to be  $\alpha$ .

$$\begin{aligned} \alpha &= Pr(\text{at least one type I error}) \\ &= 1 - (1 - \alpha')^m. \end{aligned}$$

Therefore, we set each test to level

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

$\alpha = 0.05$

$m$	3	5	10	15	25	100
$\alpha'$	0.017	0.010	0.005	0.003	0.002	0.0005

### Criteria to Control ( $m > 1$ )

	FTR	Reject	Total
Null True	$U$	$V[\text{I}]$	$m_0$
Alt True	$T[\text{II}]$	$S$	$m_1$
	$W$	$R$	$m$

- *Per-family Error Rate* (PFER):  $E(V)$ 
  - What Bonferroni actually controls
  - Greater than or equal to FWER
- *Family-wise Error Rate* (FWER):  $Pr(V > 0|m_1 = 0)$  (called *Experimentwise Error*)
- *Strong Familywise Error Rate* (SFWER):  $Pr(V > 0)$
- *k-FWER*:  $P(V > k)$
- *False Discovery Proportion* (FDP):  $P(\frac{V}{R} > \gamma)$  (for some  $\gamma < 1$ )
- *False Discovery Rate* (FDR):  $E(\frac{V}{R}|R > 0)$
- *Comparisonwise Error Rate*: no adjustment (specify  $\alpha$  for each comparison)

Most methods try to control FWER.

## Confidence Intervals

- If overall rate is  $\alpha$ , then set each confidence interval to be at  $100(1 - \alpha_{adj})$  level.

## Implications for Confidence Regions

Controlling	$\frac{\text{CWER}}{\text{FDR}}$ $\frac{\text{SFWER}}{\text{FWER}}$	means	$\frac{\text{the probability of making an incorrect statement for a given comparison is bounded by}}{\text{as a fraction of total number of statements, the number of incorrect statements will on average be no more than}} \alpha.$ $\frac{\text{probability of making any incorrect statement is bounded by}}{\text{probability of making any incorrect statement when all means are nearly equal is}}$
-------------	---	-------	--

## Bonferroni Approach to CI's

Let  $E_i$  be event that the  $i$ th confidence interval does not cover true mean. If

$$Pr(E_1) = Pr(E_2) = \dots = Pr(E_m) = \alpha',$$

then

$$\begin{aligned} Pr(\text{both correct}) &= 1 - Pr(E_1 \cup E_2) \\ &= 1 - Pr(E_1) - Pr(E_2) + Pr(E_1 \cap E_2) \\ &\geq 1 - Pr(E_1) - Pr(E_2) \\ &= 1 - 2\alpha' \end{aligned}$$

By induction

$$\begin{aligned} Pr(\text{all } m \text{ intervals are correct}) &\geq 1 - \sum_{i=1}^m Pr(E_i) \\ &= 1 - m\alpha' \end{aligned}$$

Easily generalizes to case where  $Pr(E_i)$  unequal.

**Bonferroni Adjustment:** Use  $\alpha_{adj} = \alpha/m$ .

- Adjusts for all possible dependence structures among hypotheses.

## Scheffé's Method for Confidence Intervals

Consider (linearly independent) contrasts:

$$\begin{aligned}\Gamma_1 &= \sum_i c_{i,1} \mu_{i,1} \\ \Gamma_2 &= \sum_i c_{i,2} \mu_{i,2} \\ &\vdots \\ \Gamma_m &= \sum_i c_{i,m} \mu_{i,m}\end{aligned}$$

Estimates:  $C_j = \sum_i c_{i,j} \bar{y}_i$ .

Standard Errors:  $se_C = \sqrt{MSE \sum_i \frac{c_i^2}{n_i}}$

Assuming  $SS_C/\sigma^2 \sim \chi_{m-1}^2$  (instead of  $\chi_1^2$ ), then

$$C \pm \sqrt{(a-1)F_{\alpha, a-1, N-a}} se_C$$

is a  $1 - \alpha$  confidence set for all  $\Gamma_j$ .

$$\begin{aligned}Pr(\text{CI contains true parameter for any contrast}) &\geq 1 - \alpha \\ Pr(\text{at least one CI does not contain true parameter}) &\leq \alpha\end{aligned}$$

### Remarks:

- Can “project” contrasts to get set of intervals for  $\mu_i - \mu_j$ .
- No assumption of equal  $n_i$ 's.
- Generalizes easily.
- Protects against “unplanned” comparisons.

## Tukey and Scheffé

- Both control overall error rate.
- Confidence Regions
  - *Tukey*: rectangles
  - *Scheffé*: ellipses
  - One is not contained in other.
- Tukey might disagree with  $F$ -test.

- Number of Comparisons:
  - *Tukey*: all possible pairs
  - *Scheffé*: subset of contrasts (hard)
  - *Bonferroni*: any comparisons (easy but conservative)

## Procedures based on $p$ -values

For the suite of tests  $H_1, H_2, \dots, H_m$  with respective  $p$ -values  $P_1, P_2, \dots, P_m$ , let the ordered  $p$ -values  $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$  correspond to respective hypotheses  $H_{(1)}, \dots, H_{(m)}$ . Want to keep error rate less than  $q^*$ .

### Procedures for Controlling Error Rates ( $q^*$ )

	$p_{(i)} < q^*/m$	Bonferroni	FWER
Reject $H_{0,(i)}$ if	$p_{(j)} < q^*/(m - j + 1)$ for $j = 1, \dots, i$	Holm	SFWER
	$p_{(j)} < jq^*/m$ for some $j \geq i$	B&H	FDR (independent tests)

#### Facts:

- Newman-Keuls controls FDR.
- There is a range test by Ryan-Einot-Gabriel-Welsch (REGWR) like N-K which controls SFWER.

## Controlling False Discovery Rate

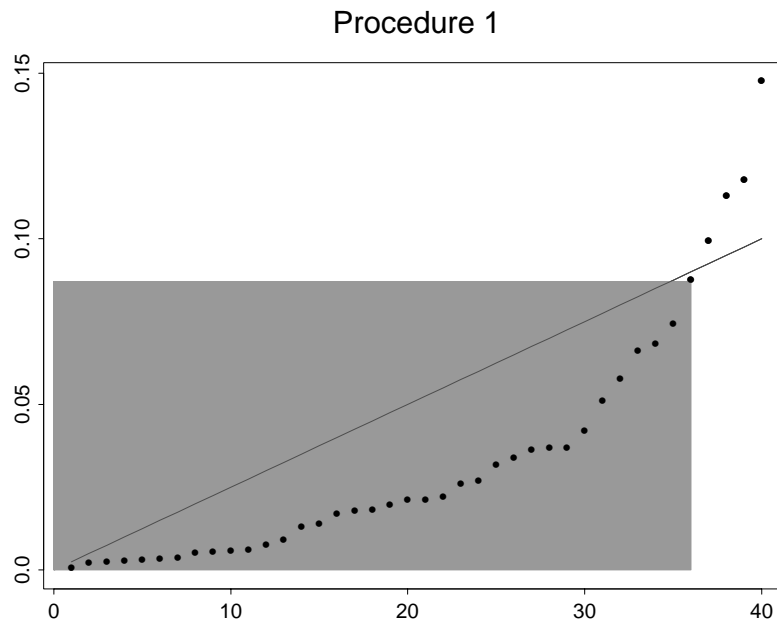
	FTR	Reject	Total
Null True	$U$	$V$	$m_0$
Alt True	$T$	$S$	$m_1$
	$W$	$R$	$m$

**False Discovery Rate (FDR):**  $E(\frac{V}{R} | R > 0) Pr(R > 0)$  or  $E(\frac{V}{R})$  (sometimes force  $\frac{V}{R}$  to be zero when  $R = 0$ ).

- $FDR \leq Pr(V \geq 1)$  (FWER)
- Equality when  $m = m_0$ .
- FDR-controlling procedures (at level  $\alpha$ )
  - Typically more powerful than FWER-controlling procedures.
  - *Strongly controls* if  $FDR \leq \alpha$  for all  $m_0$ .
  - *Weakly controls* if  $FDR \leq \alpha$  for a particular  $m_0$ .
- Used when evaluating which subset of experiments to run again.

### Procedure 1 (most common)

1. Let  $k$  be the largest  $i$  for which  $P_{(i)} \leq \frac{i}{m}q^*$ .
2. Reject all  $H_{(i)}$  for  $i = 1, 2, \dots, k$ .



- Known as a *step-up* procedure.
- Depends on independence (although works well with some dependence among test statistics).

### Procedure 2

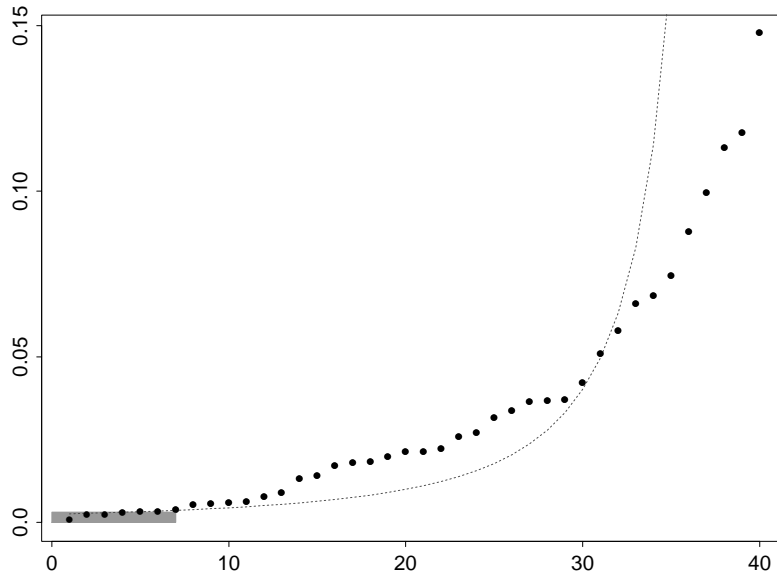
1. Define

$$\delta_i \equiv 1 - \left[ 1 - \min \left( 1, \frac{m}{m-i+1} q^* \right) \right]^{\frac{1}{m-i+1}}$$

for  $i = 1, \dots, m$ .

2. Let  $k$  be the smallest  $i$  for which  $P_{(i)} \geq \delta_i$ .
3. Reject  $H_{(1)}, \dots, H_{(k-1)}$ .

### Procedure 2

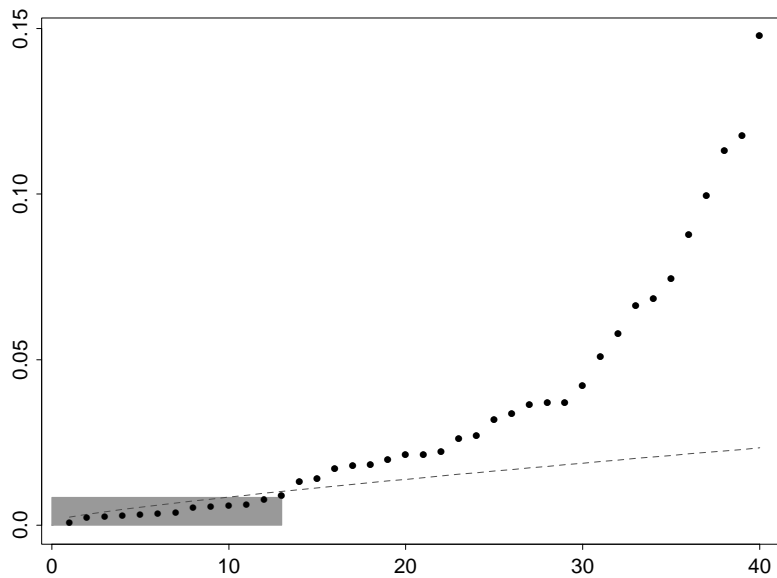


- Known as a *step-down* procedure.
- More powerful than Procedure 1 when number of hypotheses is small and many hypotheses are “clearly” true.

### Procedure 3

1. Let  $k$  be the largest  $i$  for which  $P_{(i)} \leq \frac{i}{m} (q^* / (\sum_{i=1}^m \frac{1}{i}))$ .
2. Reject all  $H_{(i)}$  for  $i = 1, 2, \dots, k$ .

### Procedure 3



- Controls FDR under “any” dependence structure among the hypotheses.
- Less powerful than first procedure.

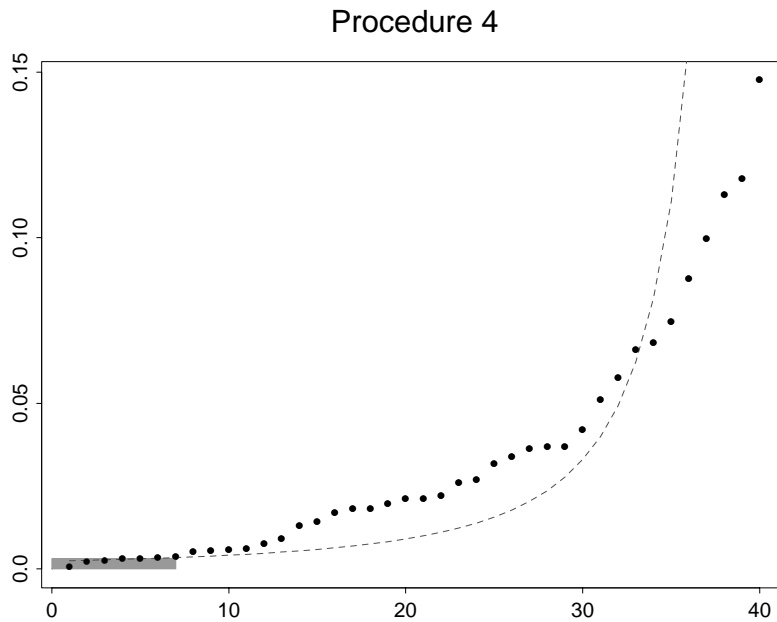
#### Procedure 4

1. Define

$$\delta_i \equiv \min \left( 1, \frac{m}{(m-i+1)^2} q^* \right)$$

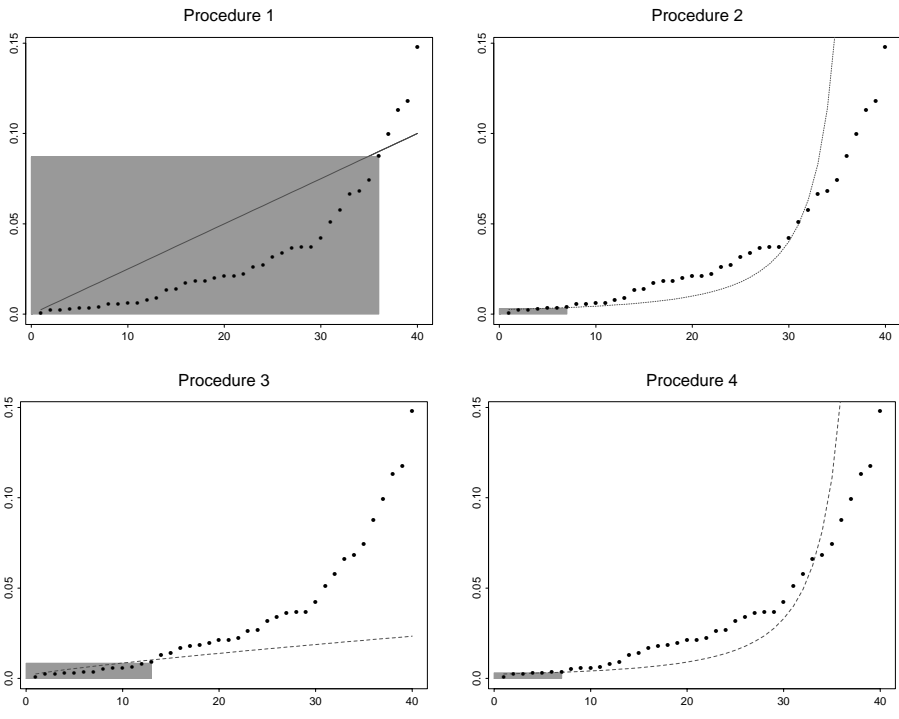
for  $i = 1, \dots, m$ .

2. Reject  $H_{(1)}, \dots, H_{(k-1)}$ , where  $k$  is the smallest  $i$  for which  $P_{(i)} > \delta_i$ .

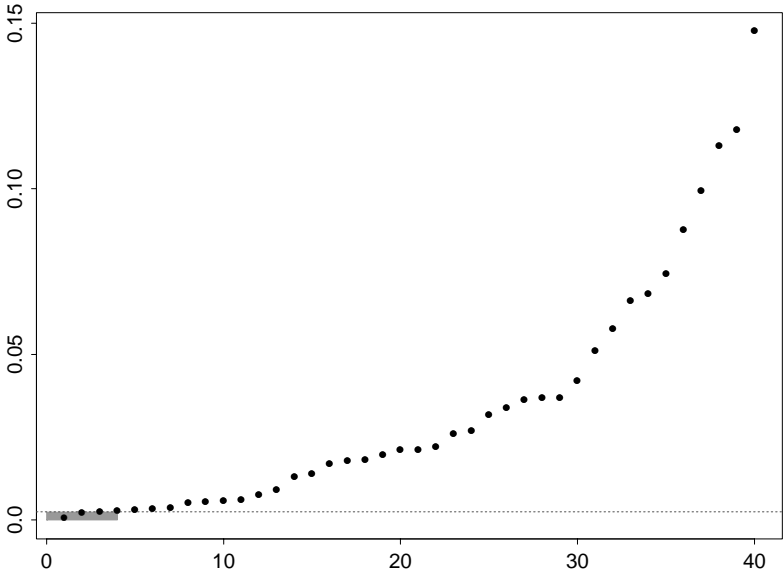


- More powerful than Procedure 3 when number of hypotheses  $m$  is small and most are far from being true.
- Also controls FDR under any dependence structure.

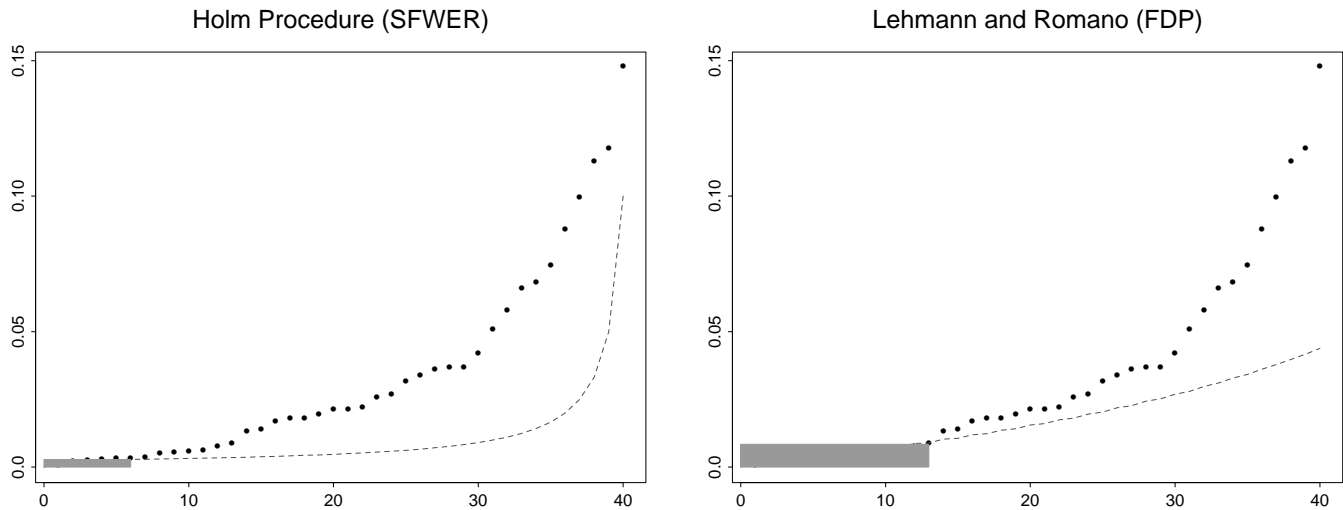
# Comparison with Bonferroni



## Bonferroni Procedure



## Comparison with Other Procedures



- Lehmann and Romano's procedure for controlling FDP:

$$\alpha_{adj} = \frac{(\lfloor \gamma j \rfloor + 1)\alpha}{(m + \lfloor \gamma j \rfloor - j)C_{\lfloor \gamma m \rfloor + 1}},$$

where  $C_k = \sum_{s=1}^k \frac{1}{s}$  and  $\lfloor x \rfloor$  is the greatest integer smaller than or equal to  $x$ .

- In this case,  $P(FDP > \gamma) \leq \alpha$ .

### Summary of Error Rates

- *Comparisonwise Error Rate (CWER) or Per Comparison Error Rate*
  - Expected fraction of individual tests that reject  $H_{0,i}$  when  $H_{0,i}$  are true
  - Makes no correction for multiple comparisons
- *Experimentwise Error Rate or Per Experiment Error Rate or Familywise Error Rate (FWER)*
  - Probability of rejecting one or more  $H_{0,i}$  when all  $H_{0,i}$  are true
  - Doesn't take into account any possible correct rejections
  - Stronger than CWER
- *False Discovery Rate (FDR)*
  - Expected number of Type I errors made
  - Same as FWER if all  $H_{0,i}$  are true
  - Allows more incorrect rejections as the number of true rejections increases

- *Strong Familywise Error Rate* (SFWER)
  - Probability that the number of false discoveries (Type I errors) is positive, irrespective of the number of correct rejections

## Levels of Inference for Multiple Comparisons (Hsu 1996)

Weakest	Per comparison	Present inference without adjustment
↓	Test of homogeneity	$F$ -test: difference exists but not sure which specific means differ from each other
↓	Confidence inequalities	Report which specific means are different from each other
↓	Confidence directions	Justification for statement that one mean is larger (or smaller) than another
Strongest	Confidence Intervals	Not just direction – report size of difference

## Power in multiple-testing framework

- *Minimal power* – probability of rejecting at least one false null hypothesis. Not particularly appropriate, since goal is to reject as many false hypotheses as possible.
- *Global power* – probability of rejecting all false null hypotheses. Too strict: it is not necessarily failure to miss a single false null hypotheses
- *Average power* – average of the individual probabilities of rejecting each null hypothesis. Equivalent to expected number of rejected false null hypotheses. Appropriate in many circumstances
- *Relative power* – expected proportion of false null hypotheses that will be rejected
- Probability of rejecting at least  $\gamma \times 100\%$  false null hypotheses, where  $\gamma \in (0, 1]$  specified by user

Different notions of power lead to different testing methods.

## Criticism of Notion of $p$ -value

- Often misunderstood.
  - “ $p$ -value is probability that null hypothesis is true.”
  - Strong psychological/philosophical pull (see Bayesian methods)
- Null hypothesis is *never* true.
  - $p$ -value measures extent of effort
  - $p$ -value measures sample size

- $p$ -values divert attention from other important issues, like Type II errors.
- Partial solution: power studies/confidence intervals more important; where should our resources go?
- Ultimately,  $p$ -value (and confidence regions) are almost always approximate.
  - Relies on technical assumptions we can't verify (or understand)

## Data Snooping

- Practice related to having many tests
- first look at the data and choose null hypotheses based on “interesting features” of the data; discard uninteresting behavior
- reporting smallest  $p$ -value is not like doing just one test
- use simultaneous inference techniques to decide which error rate to control and choosing procedure to control desired error rate

## Contrary View (to adjusting)

- Multiple comparison procedures are applied arbitrarily in practice (for instance, rarely seen in factorial designs)
  - Matter of perspective – appropriate error rate and family of tests can depend on user
- Adjusting
  - makes it harder to detect real differences
  - encourages the idea that data can be used for two things at once: generating hypotheses and testing them at the same time without running a new experiment.
  - yields different results for different adjustments.

“Two extremes of behavior are open to anyone involved in statistical inference. A non-multiple comparisonist regards each separate statistical statement as a family, and does not give increased protection to any group of statements through group error rates. At the other extreme is the ultraconservative statistician who has just a single family consisting of every statistical statement he might make during his lifetime. If all statisticians operated in this latter fashion at the 5 percent level, then 95 percent of the world's statisticians would never falsely reject a null hypothesis, and 5 percent would be guilty of some sin against nullity. There are a few statisticians who would adhere to the first principle, but the author has never met any of the latter variety.”

- Scenario: Two scientists –  $A$  and  $B$  – do the same experiments, except  $B$  adds several “spurious” experiments that he doubts will be significant.
  - controlling FWER -  $A$  will reject more hypotheses.
  - controlling FDR -  $B$  will reject more hypotheses.

Some conclude that FDR can be “manipulated.”

- However, FDR is a criteria for suggesting which experiments should be done next, not which should be published. It is in the scientists’ best interest not to add experiments that aren’t worthwhile.
  - Ultimately, any prior predictions that can be made about how many null hypotheses are true should be reflected in the controlling parameter ( $\alpha$  or  $q^*$ ).
- Testing many things at once in case of “hidden” factors: can we *select* hypotheses that we “meant” to test.
    - Partial (unsatisfactory) answer: there are new methods (bootstrap-based) for adjusting estimates of statistical uncertainty in the face certain types of “automated data-snooping.”

Still...

Science is about discovery, and sometimes identifying “promising” areas for further research is more important than guaranteeing that every “discovery” can be confidently cited with reasonable certainty.

Ultimately, we count on the belief that incorrect conclusions will eventually be identified and discredited with further investigation.

## The Copernican Principle and the Rise of Alternative Science

“The ‘Copernican principle’ states that we are unlikely to be in any ‘special place’ in space-time. Therefore, things have a 95% likelihood of being in the middle 95% of their lifetime. This interesting theory was discovered by J. Richard Gott III (*Nature* 27 May 93), who used it to prove – at 95% confidence level – that *Homo sapiens* will be around for another 200 thousand to 8 million years! Eight days later, WHAT’S NEW use the same principle to predict how long his theory would be around (*W. N.* 4 Jun 93). Its calculated demise was between 8pm that evening and 20 April 94 – with 95% certainty! So, we can be 95% confident that if the theory was right, it’s already dead. But if it’s wrong, it has a 95% likelihood of still being alive!”