

Multiple Comparisons

Omnibus test

Omnibus or global test: is explained variance statistically larger than unexplained variance?

- What does it tell us?
 - population Means/Medians are different
- How?
 - Don't know.

Getting the specifics

- Multiple comparisons
 - which population means are different?
 - more generally which medians, means, trimmed means are different?
- Omnibus test is usually not required first
- Possible for omnibus test to not find anything, while multiple comparisons procedure finds significant results!

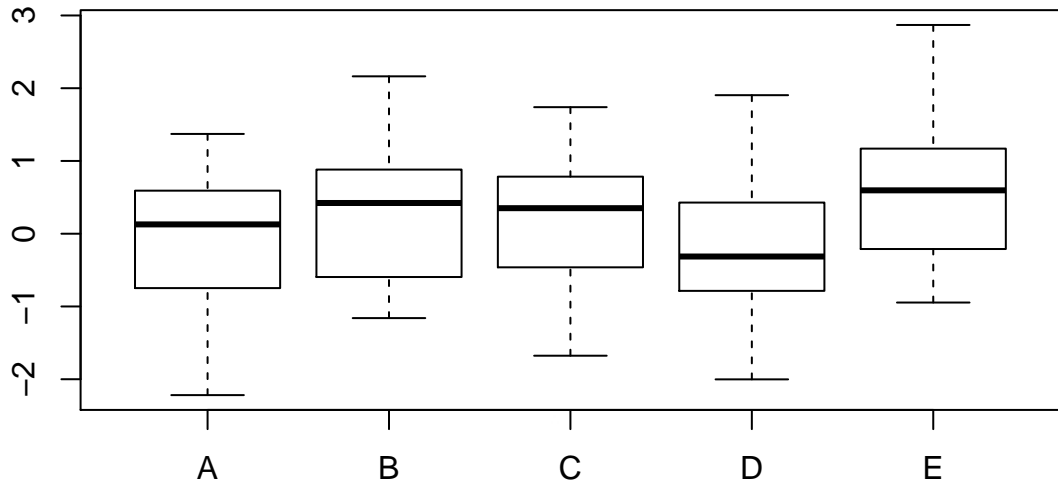
A five group example

```
set.seed(554)
x <- rnorm(100); group <- factor( rep( LETTERS[1:5], each=20 ) )
tapply(x, group, mean)
```

##	A	B	C	D	E
##	-0.09851133	0.24606290	0.22823887	-0.24223872	0.59744611

A five group example, pg 2

```
boxplot(x~group)
```



Test each pair?

```
t.test( x[ group=='A' ], x[ group=='B' ])
```

```
## Welch Two Sample t-test
```

```
##
```

```
## data: x[group == "A"] and x[group == "B"]
```

```
## t = -1.0839, df = 37.85, p-value = 0.2853
```

```
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## -0.9882436 0.2990951
```

How many pairs?

k groups	$m = \frac{k(k-1)}{2}$ pairs
3	3
4	6
5	10
6	15
7	21
\vdots	\vdots

Examine All Pairs

```
pairwise.t.test( x, group, p.adjust.method='none',  
                 conf.level=.95)$p.value
```

##	A	B	C	D
## B	0.28380983	NA	NA	NA
## C	0.30931125	0.9556525	NA	NA
## D	0.65401777	0.1299588	0.1443946	NA
## E	0.03195149	0.2744607	0.2510068	0.01005151

- $\mu_A \neq \mu_E$ and $\mu_D \neq \mu_E \Rightarrow 2$ Type I errors
- Multiple tests \Rightarrow increased risk of errors

Error Rates

- Individual: α_I
 - prob. Type I error for single comparison
- Experimentwise Error Rate: α_E
 - also called Familywise Error Rate (FWER or FW_{α})
 - prob. at least one Type I error for m comparisons
 - strongly controlling FWER can lower the power of a test

Goal: control FWER (α_E)

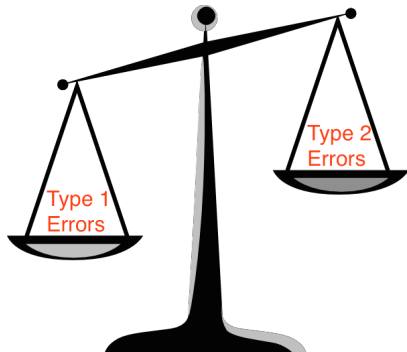
No procedure works well in all scenarios

- Considerations

- Normal distributions?
- Equal variances?
- Same shaped distributions?
- All pairs or a few contrasts or pairs?
- Confidence intervals needed or just hypothesis tests?

A side note

- Genomics and other fields make 100's or 1000's comparisons
- False Discovery Rate $\approx \frac{\text{\#false positives}}{\text{\#positives}}$



Going beyond pairs

- mean of groups 1 and 2 different than mean of groups 3 and 4?
- textbook
 - Ott Section 9.2
 - don't worry about orthogonal contrasts
 - Example 9.3
 - F-test is same as t-test here, because $F = t^2$
- `onewayComp()` code
 - handles arbitrary linear contrasts
 - uses Welch type corrections for unequal variances
- formulas for t -tests of contrasts at end of slides

Bonferroni Correction

- Bonferroni Inequality: $\alpha_E \leq m\alpha_I$
- To make α_E small, reduce α_I
- To get $\alpha_E \leq \alpha$, set $\alpha_I = \alpha/m$.

Bonferroni Example - compare P to reduced α

5 groups, 10 comparisons, $\alpha_I = .05/10 = .005$

```
pairwise.t.test( x, group, p.adjust.method='none',  
                 conf.level=.95)$p.value
```

##		A	B	C	D
## B	0.28380983		NA	NA	NA
## C	0.30931125	0.9556525		NA	NA
## D	0.65401777	0.1299588	0.1443946		NA
## E	0.03195149	0.2744607	0.2510068	0.01005151	

For $\alpha_E = .05$ we reject each individual HT if $p < .005$

Bonferroni Example - adjust P and compare to α_E

5 groups, 10 comparisons, $p_{\text{adj}} = \min(10p, 1)$

```
pairwise.t.test( x, group, p.adjust.method='bonferroni',  
                 conf.level=.95)$p.value
```

##		A	B	C	D
## B	1.0000000	NA	NA	NA	
## C	1.0000000	1	NA	NA	
## D	1.0000000	1	1	NA	
## E	0.3195149	1	1	0.1005151	

For $\alpha_E = .05$ we reject each individual HT if $p_{\text{adj}} = 10p < .05$

Bonferroni Pros and Cons

- Pros
 - any number of simultaneous hypothesis tests or confidence intervals
 - any procedures allowed
- Cons
 - overly conservative / reduced power - too many Type II errors
 - avoid, if possible, unless the number of comparisons is small

Procedures for CI's

<i>Simultaneous Confidence Intervals</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games- Howell	Bonferroni corrected bootstrapped CI's or other more advanced procedures
	Few pairs or contrasts	t-intervals with Bonferroni correction	

CI's for All Pairs and Normal

<i>Simultaneous Confidence Intervals</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games- Howell	Bonferroni corrected bootstrapped CI's or other more advanced procedures
	Few pairs or contrasts	t-intervals with Bonferroni correction	

Tukey-Kramer

$$\bar{x}_i - \bar{x}_j \pm q_{\text{crit}} s_p \sqrt{\frac{\frac{1}{n_i} + \frac{1}{n_j}}{2}}$$

q_{crit} is the upper-tail critical value of the Studentized range distribution (`qtukey()` in R).

$$s_p = \sqrt{MSE}, df = N - k$$

exact control of FWER if samples balanced and population variances equal.

Games-Howell

$$\bar{x}_i - \bar{x}_j \pm q_{\text{crit}} \sqrt{\frac{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}{2}}$$

q_{crit} is the upper-tail critical value of the Studentized range distribution (`qtukey()` in R).

“Welch” corrected degrees of freedom: $v_i = \frac{s_i^2}{n_i}$, $v_j = \frac{s_j^2}{n_j}$, $df = \frac{(v_i + v_j)^2}{\frac{v_i^2}{n_i - 1} + \frac{v_j^2}{n_j - 1}}$

approximate control of FWER

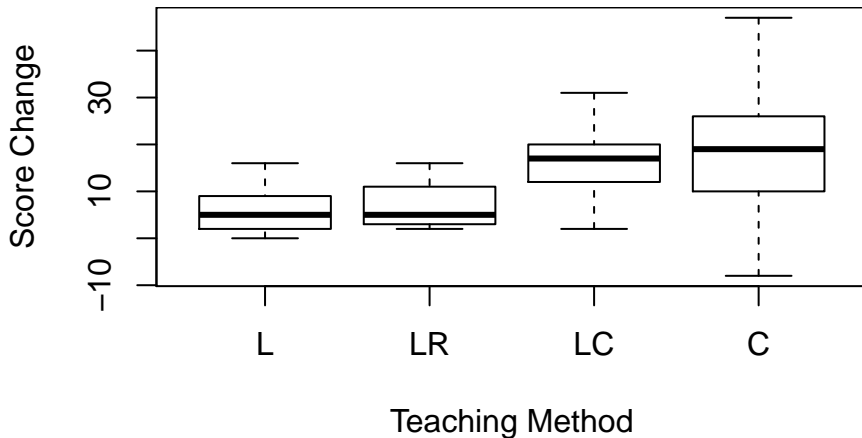
Tukey-Kramer vs. Games-Howell

- Tukey-Kramer
 - approximately balanced (equal) sample sizes
 - and approximately equal variances
- Games-Howell
 - unbalanced sample sizes
 - and/or unequal variances
 - don't transform, use this instead

Reading Example

- data from Ott, problem 7.22, page 393
- compare improvement in reading score for 4 teaching methods:
 - (L) Lecture only
 - (LR) Lecture and Remedial textbook
 - (LC) Lecture and Computer assistance
 - (C) Computer Assistance only

Reading Example (2)



Note: variances are different!

Reading Example (3)

Ignoring the equal variances for a moment ...

```
TukeyHSD(aov(scoreInc~group, data=d))$group
```

##		diff	lwr	upr	p adj
##	LR-L	1.385714	-11.7198702	14.49130	0.99159633
##	LC-L	10.100000	-2.1190235	22.31902	0.13399125
##	C-L	11.766667	-0.4523568	23.98569	0.06243555
##	LC-LR	8.714286	-4.6877469	22.11632	0.30906387
##	C-LR	10.380952	-3.0210802	23.78299	0.17477900
##	C-LC	1.666667	-10.8697869	14.20312	0.98361676

Reading Example (4)

Tukey-Kramer with onewayComp()

```
source('./onewayComp.R') # included in weekly download
onewayComp(scoreInc~group, data=d, var.equal=T)$comp[,c(1,2,3,6)]
```

##		diff	lwr	upr	p adj
##	LR-L	1.385714	-11.7198702	14.49130	0.99159633
##	LC-L	10.100000	-2.1190235	22.31902	0.13399125
##	C-L	11.766667	-0.4523568	23.98569	0.06243555
##	LC-LR	8.714286	-4.6877469	22.11632	0.30906387
##	C-LR	10.380952	-3.0210802	23.78299	0.17477900
##	C-LC	1.666667	-10.8697869	14.20312	0.98361676

No significant differences (all the intervals contain 0)

Reading Example (5)

Unequal variances so use Games-Howell:

```
onewayComp(scoreInc~group, data=d, var.equal=F)$comp[,c(1,2,3,6)]
```

##		diff	lwr	upr	p adj
##	LR-L	1.385714	-6.2096361	8.981065	0.94875286
##	LC-L	10.100000	0.6763831	19.523617	0.01776619
##	C-L	11.766667	-5.3317545	28.865088	0.16624444
##	LC-LR	8.714286	-1.3500848	18.778656	0.07550173
##	C-LR	10.380952	-6.9040031	27.665908	0.27902060
##	C-LC	1.666667	-16.0794366	19.412770	0.99231005

LC reading improvement is 0.7 to 19.5 greater than L reading improvement!

CI's for few pairs or contrasts

<i>Simultaneous Confidence Intervals</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games- Howell	Bonferroni corrected bootstrapped CI's or other more advanced procedures
	Few pairs or contrasts	t-intervals with Bonferroni correction	

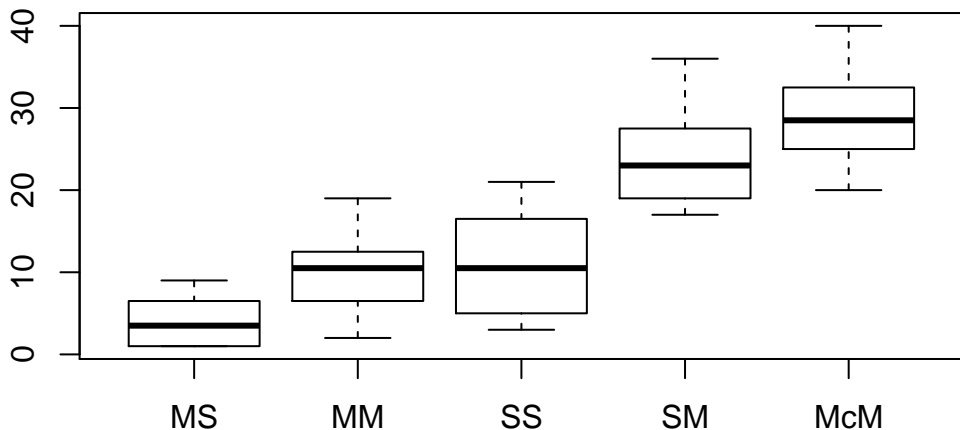
Bonferroni Correction for CI's

- 4 comparisons
- want familywise error rate $\alpha_E = 0.05$
- familywise confidence level $1 - \alpha_E = 0.95$
- individual comparison error rate $\alpha_I = 0.05/4 = 0.0125$
- individual comparison confidence level $1 - \alpha_I = 0.9875$

Morphine Tolerance Example

- Record pain sensitivity after rats developed morphine tolerance
- 5 treatment groups: MS, MM, SS, SM, McM
- from David Howell's book: *Statistical Methods for Psychology* - Chapter 12 (included with download)
- original study: www.eou.edu/psych/re/morphinetolerance.doc

Morphine Tolerance Example (2)



- Normality and equal variances seem OK

Morphine Tolerance Example (3)

- Particular comparisons of interest
 - average of MS, MM, SS vs average of SM, McM
 - MM vs McM
 - MS vs SS
 - MM vs SS
- build a contrast matrix with weights (next slide)

Morphine Tolerance Example (4)

- contrast matrix
 - each contrast is a row
 - weights in each row add to zero
 - absolute value of weights in each row adds to 2 (standard contrasts)

```
K = rbind('ave2 -ave3'=c(-1/3,-1/3,-1/3,1/2,1/2),  
          'McM - MM'=c( 0, -1, 0, 0, 1),  
          'SS - MS'=c(-1, 0, 1, 0, 0),  
          'MM - SS'=c( 0, 1, -1, 0, 0) )
```


Morphine Tolerance Example (5)

- Approach 1: use `glht()` from package 'multcomp'.
- Bonferroni with 4 comparisons, want familywise 0.95, so individual 0.9875.

```
require(multcomp) # install package if necessary  
# setup oneway ANOVA  
amod <- aov(pain ~ treat, data = morph)  
comp.glht <- glht( amod, linfct = mcp( treat = K) )
```

Output on next slide.

Morphine Tolerance Example (6)

```
confint( comp.glht )
```

```
## 95% family-wise confidence level
```

```
##
```

```
## Linear Hypotheses:
```

##		Estimate	lwr	upr
##	ave2 -ave3 == 0	18.1667	13.4668	22.8665
##	McM - MM == 0	19.0000	11.7190	26.2810
##	SS - MS == 0	7.0000	-0.2810	14.2810
##	MM - SS == 0	-1.0000	-8.2810	6.2810

- multcomp family of confidence intervals use a “Tukey” like procedure
- requires approximately balanced and equal variances

Morphine Tolerance Example (7)

Can do the Bonferroni correction manually:

```
confint( comp.glm, calpha=univariate_calpha(), level = 1-.05/4 )
```

```
## 98.75% confidence level
```

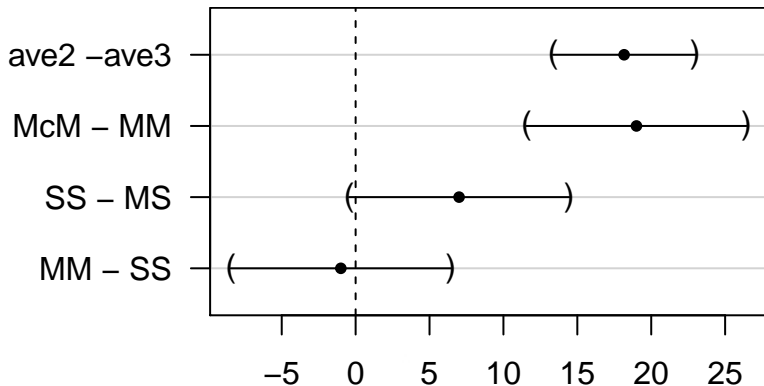
```
##
```

```
## Linear Hypotheses:
```

##		Estimate	lwr	upr
##	ave2 -ave3 == 0	18.1667	13.3587	22.9746
##	McM - MM == 0	19.0000	11.5516	26.4484
##	SS - MS == 0	7.0000	-0.4484	14.4484
##	MM - SS == 0	-1.0000	-8.4484	6.4484

Morphine Tolerance Example (8)

```
plot(confint( comp.glht,  
             calpha=univariate_calpha(), level= 1-.05/4 ))
```



Morphine Tolerance Example (9)

```
onewayComp( pain~treat, data = morph, con=K,  
            adjust = 'bonferroni')$comp[,c(1:3,4,6)]
```

##			diff	lwr	upr	t	p adj
##	ave2	-ave3	18.16667	13.3587469	22.974586	9.9502931	3.862945e-11
##	McM	- MM	19.00000	11.5516028	26.448397	6.7175144	3.548925e-07
##	SS	- MS	7.00000	-0.4483972	14.448397	2.4748737	7.327455e-02
##	MM	- SS	-1.00000	-8.4483972	6.448397	-0.3535534	1.000000e+00

Morphine Tolerance Example (10)

- With 95% confidence
 - Average population mean for SM and McM is 13.4 to 23.0 higher than average population mean of MS, MM, and SS
 - population mean for McM is 11.6 to 26.4 larger than population mean for MM

CI's for Non-normal Distributions

<i>Simultaneous Confidence Intervals</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games- Howell	Bonferroni corrected bootstrapped CI's or other more advanced procedures
	Few pairs or contrasts	t-intervals with Bonferroni correction	

Bootstrap with onewayComp

If distributions deviate significantly from normal, then set `nboot > 0`.

```
onewayComp( pain~treat, data = morph, con=K, nboot=10000,  
            adjust = 'bonferroni')$comp[,c(1:3,4,6)]
```

##			diff	lwr	upr	t	p	adj
##	ave2	-ave3	18.16667	13.1171931	23.201579	9.9502931	0.0000	
##	McM	- MM	19.00000	11.7194181	26.386214	6.7175144	0.0000	
##	SS	- MS	7.00000	-0.1206754	14.187423	2.4748737	0.0556	
##	MM	- SS	-1.00000	-9.0565723	7.108379	-0.3535534	1.0000	

CI's based on Studentized estimates. `var.equal = TRUE` by default.

Bootstrap using boot package

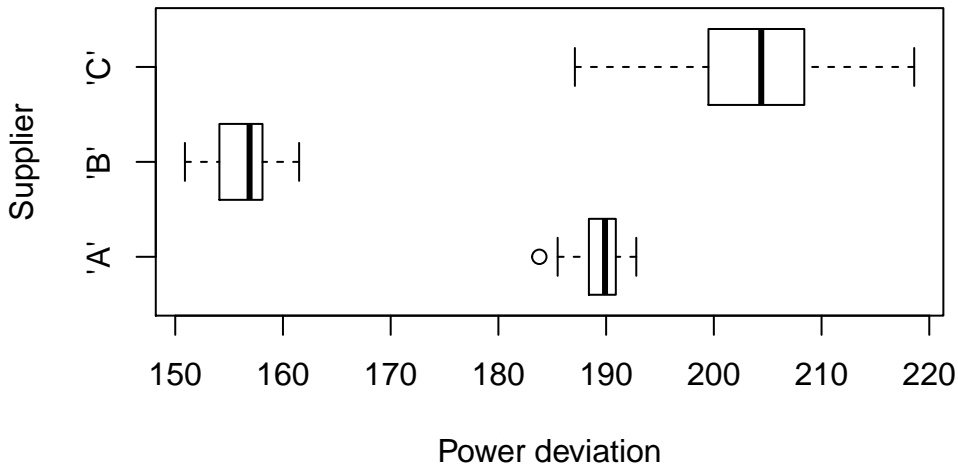
CI's for contrasts with
medians, trimmed means, and
means.



Watch presentation about
'boot' package.

Contact Lenses Example

See Problem 8.27 in Ott (page 442).



Contact Lenses Example (2)

- Estimate differences in population medians: C-A, C-B, and A-B

```
bootMedDiff <- function(d,i){  
  # d is a dataframe with  
  #   quantitative variable in column 1  
  #   factor variable in column 2  
  meds <- tapply(d[i,1],d[,2],median)  
  c( meds[3]-meds[1], meds[3]-meds[2], meds[1]-meds[2])  
}
```

Contact Lenses Example (3)

```
boot.object <- boot(contacts, bootMedDiff, R = 5000,  
                   strata = contacts$Supplier)  
  
# med_C - med_A  
boot.ci(boot.object, conf = 1 - .05/3, type='bca', index=1)$bca[4:5]
```

```
## [1] 6.7 22.1
```

```
# med_C - med_B  
boot.ci(boot.object, conf = 1 - .05/3, type='bca', index=2)$bca[4:5]
```

```
## [1] 39.2 54.5
```

```
# med_A - med_B (= 6)  
boot.ci(boot.object, conf = 1 - .05/3, type='bca', index=3)$bca[4:5]
```

```
## [1] 27.8 27.8
```

Contact Lenses Example (4)

For power deviations, we are 95% confident that

- Pop. median C is 6.7 to 22.1 larger than pop. median A
- Pop. median C is 39.2 to 54.5 larger than pop. median B
- Pop. median A is 27.8 to 37.8 larger than pop. median B

Statistically significant, YES. Practically significant?

Procedures for HT's

<i>Simultaneous Hypothesis Tests</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games-Howell (use REGWQ if balanced and equal variances)	Dunn Test if appropriate, other more advanced procedures such as bootstrap
	Few pairs or contrasts	F or t tests with Bonferroni-Holm adjusted p-values	Pairwise test like Mann Whitney or bootstrap tests + adjusted p-values

Sequential Procedures

- Adjust α and/or p sequentially to account for
 - number of tests remaining
 - possibly the number of means between the pair being tested
- No confidence intervals.
- Less conservative, more power!

Bonferroni-Holm Stepdown Procedure

- Compare smallest p -value to α/m
- Second smallest p -value to $\alpha/(m - 1)$, etc.
- Stop at first non-rejection and do not reject any remaining hypotheses.
- Strengths: Guaranteed FWER, arbitrary contrasts, more power than Bonferroni
- Cons: still conservative

Bonferroni-Holm Example

null hyp.	p-value
H1	0.01
H2	0.04
H3	0.03
H4	0.002

Bonferroni-Holm Example - adjusted p -values

null hyp.	p-value
H1	0.01
H2	0.04
H3	0.03
H4	0.002

```
p.adjust( p = c( .01, .04, .03, .002), method = 'holm')
```

```
## [1] 0.030 0.060 0.060 0.008
```

Benjamin-Hochberg

- less conservative sequential method
- controls False Discovery Rate (FDR)
- allows more false positives, but more powerful
- use for Exploratory Analysis

Ryan Procedure (REGWQ)

- Sequential, more powerful version of Tukey Test
- Requires samples
 - are from normal distributions with equal variances
 - have (approx) equal sample sizes
- guarantees FWER
- Available in package 'mutoss', but difficult to use.
- Use downloaded `regwqComp()`.
 - also in DS705data package.

Multiple HT with Normal Distributions

<i>Simultaneous Hypothesis Tests</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games-Howell (use REGWQ if balanced and equal variances)	Dunn Test if appropriate, other more advanced procedures such as bootstrap
	Few pairs or contrasts	F or t tests with Bonferroni-Holm adjusted p-values	Pairwise test like Mann Whitney or bootstrap tests + adjusted p-values

Which procedure?

For independent samples from (approx.) normal distributions:

- approx. balanced sample sizes AND approx. equal variances
 - Tukey-Kramer or REGWQ if available
 - `onewayComp()` with `var.equal = TRUE`, or `regwqComp()`
- unbalanced sample sizes and/or unequal variances
 - Games-Howell
 - `onewayComp()` with `var.equal = FALSE`

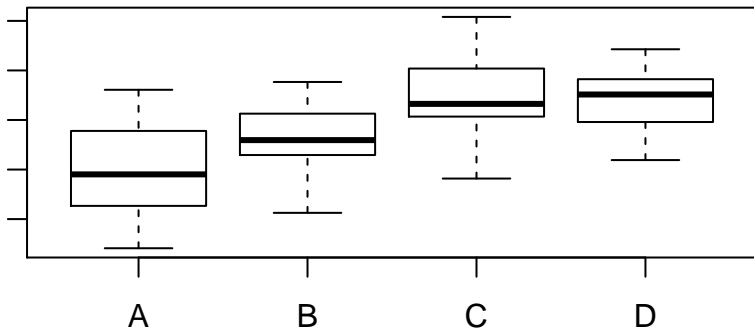
Example REGWQ

4 samples from normals with $\sigma = 1$ and $\mu_A = 0, \mu_B = 0.5, \mu_C = 1, \mu_D = 1.5$

```
set.seed(4321)
```

```
x <- rnorm(80)+rep(c(0,.5,1,1.5),each=20)
```

```
g <- factor(rep(LETTERS[1:4],each=20)); d <- data.frame(x,g)
```



Example REGWQ (2)

If REGWQ is not available, can use Tukey-Kramer:

```
posthoc<-onewayComp(x~g,adjust='one.step')  
posthoc$comp[,c(1,6,7)]
```

##		diff	p adj	rej	H_0
##	B-A	0.63120069	5.730278e-02		0
##	C-A	1.52019844	1.659970e-07		1
##	D-A	1.46525313	4.237883e-07		1
##	C-B	0.88899775	2.902847e-03		1
##	D-B	0.83405244	5.858867e-03		1
##	D-C	-0.05494531	9.960118e-01		0

Can't separate μ_A and μ_B or μ_C and μ_D

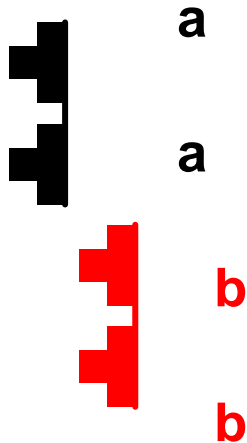
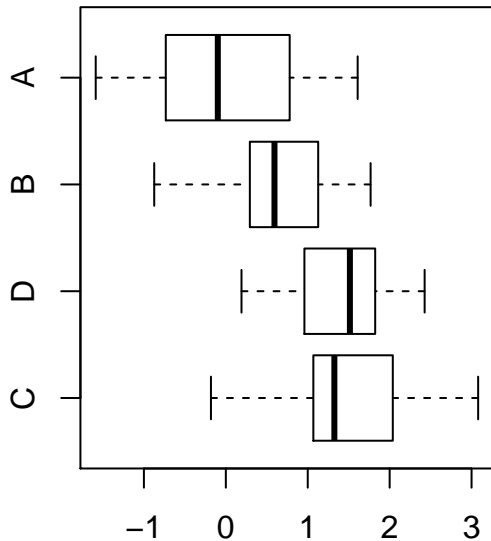
Example REGWQ (3)

Visualize pairs with multcompBoxlot

```
require(multcompView)
padj_extract <- function(formula,data){posthoc$comp[, 'p adj']}
multcompBoxplot(x~g,data=d,
                 horizontal=TRUE,compFn="padj_extract")
```

See next slide for plot

Example REGWQ (4)



Example REGWQ (5)

With REGWQ

```
source('./regwqComp.R')  
posthoc<-regwqComp(x~g)  
posthoc[,c(1,4,5,6)]
```

##		diff	p adj	alpha adj	rej	H_0
##	B-A	0.63120069	1.208868e-02	0.02532057		1
##	C-A	1.52019844	1.659970e-07	0.05000000		1
##	D-A	1.46525313	2.127664e-07	0.05000000		1
##	C-B	0.88899775	1.511694e-03	0.05000000		1
##	D-B	0.83405244	1.085063e-03	0.02532057		1
##	D-C	-0.05494531	8.234992e-01	0.02532057		0

Now μ_A and μ_B are also separated!

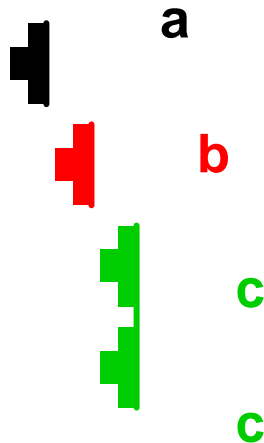
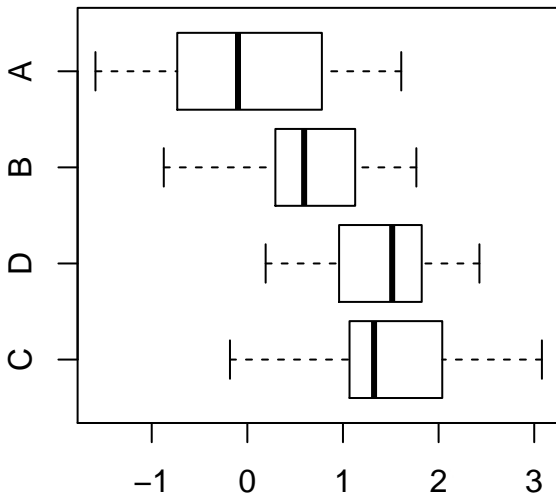
Example REGWQ (6)

Visualize pairs with multcompBoxlot

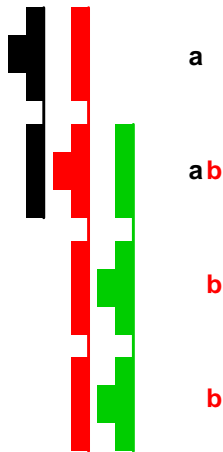
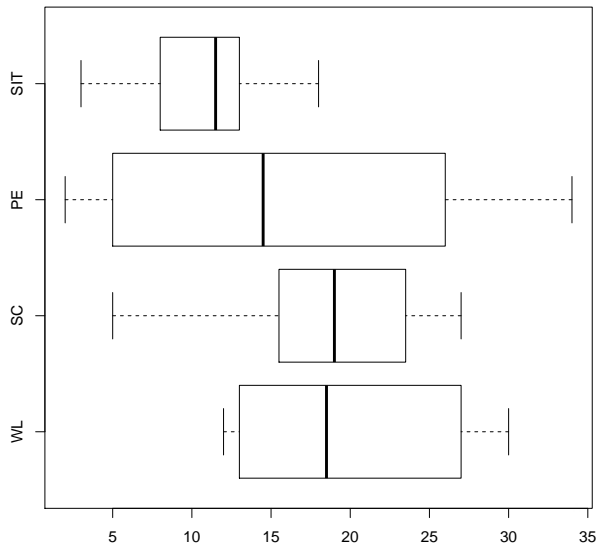
```
require(multcompView)
padj_extract <- function(formula,data){posthoc[, 'p adj']}
par(mar=c(4,4,4,4))
multcompBoxplot(x~g,data=d,horizontal=TRUE,compFn="padj_extract")
```

See next slide for plot

Example REGWQ (8)



One more visualization



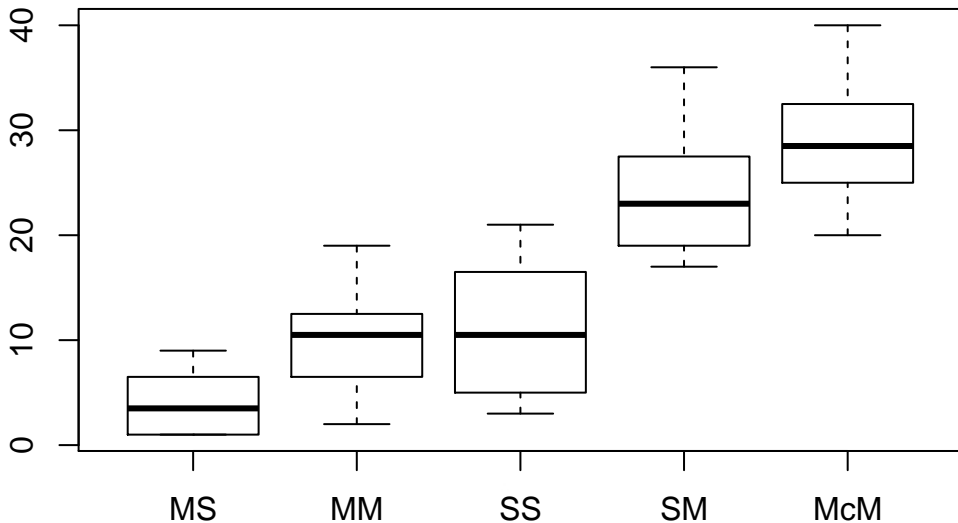
HT's for few pairs or contrasts

<i>Simultaneous Hypothesis Tests</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games-Howell (use REGWQ if balanced and equal variances)	Dunn Test if appropriate, other more advanced procedures such as bootstrap
	Few pairs or contrasts	F or t tests with Bonferroni-Holm adjusted p-values	Pairwise test like Mann Whitney or bootstrap tests + adjusted p-values

Morphine Again

- Particular comparisons of interest
 - average of MS, MM, SS vs average of SM, McM
 - MM vs McM
 - MS vs SS
 - MM vs SS

Morphine Again (2)



Morphine Again (3)

Test with one-step Bonferroni

```
onewayComp( pain~treat, data=morph,  
            con=K, adjust='bonferroni' )$comp[,c(1,4,5:7)]
```

##			diff	t	p	p adj	rej	H_0
##	ave2	-ave3	18.16667	9.9502931	9.657362e-12	3.862945e-11		1
##	McM	- MM	19.00000	6.7175144	8.872312e-08	3.548925e-07		1
##	SS	- MS	7.00000	2.4748737	1.831864e-02	7.327455e-02		0
##	MM	- SS	-1.00000	-0.3535534	7.257945e-01	1.000000e+00		0

The first two contrasts are significant. Note, the third is close ($P_{\text{adj}} \approx 0.07$)

Morphine Again (4)

Test with sequential Bonferroni-Holm

```
onewayComp( pain~treat, data=morph,  
            con=K, adjust='holm' )$comp[,c(1,4,5:7)]
```

##			diff	t	p	p adj	rej	H_0
##	ave2	-ave3	18.16667	9.9502931	9.657362e-12	3.862945e-11		1
##	McM	- MM	19.00000	6.7175144	8.872312e-08	2.661693e-07		1
##	SS	- MS	7.00000	2.4748737	1.831864e-02	3.663728e-02		1
##	MM	- SS	-1.00000	-0.3535534	7.257945e-01	7.257945e-01		0

The first three contrasts are significant.

Morphine Again (5)

Test with sequential Benjamin-Hochberg (controls FDR)

```
onewayComp( pain~treat, data=morph,  
            con=K, adjust='BH')$comp[,c(1,4,5:7)]
```

##			diff	t	p	p adj	rej	H_0
##	ave2	-ave3	18.16667	9.9502931	9.657362e-12	3.862945e-11		1
##	McM	- MM	19.00000	6.7175144	8.872312e-08	1.774462e-07		1
##	SS	- MS	7.00000	2.4748737	1.831864e-02	2.442485e-02		1
##	MM	- SS	-1.00000	-0.3535534	7.257945e-01	7.257945e-01		0

The first three contrasts are significant.

HT's for non-normal distributions

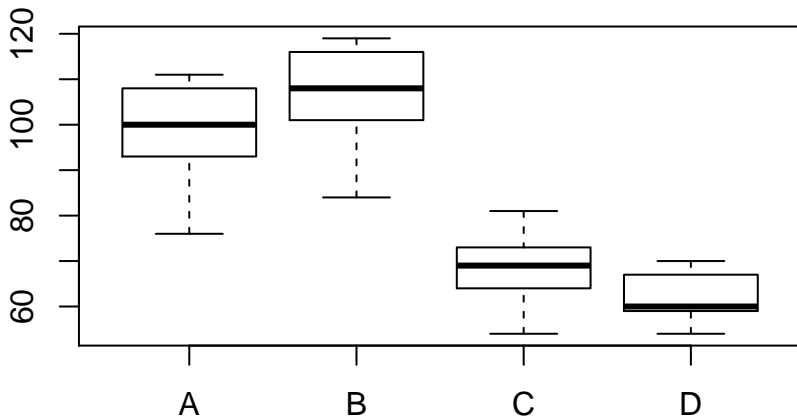
<i>Simultaneous Hypothesis Tests</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games-Howell (use REGWQ if balanced and equal variances)	Dunn Test if appropriate, other more advanced procedures such as bootstrap
	Few pairs or contrasts	F or t tests with Bonferroni- Holm adjusted p-values	Pairwise test like Mann Whitney or bootstrap tests + adjusted p-values

Dunn Test

- nonparametric posthoc pairwise test of location for Kruskal-Wallis Test
- requires same conditions as Kruskal-Wallis Test
- independent samples from distributions
 - with same shape and scale
 - possibly shifted
- p-values still need to be adjusted for multiple comparisons
- p-values are compared to $\alpha/2$ not α

Rain Example

4 cities A,B,C,D have monthly rainfall (mm) recorded for 6 different months



Some skewness, similar shapes and scales \Rightarrow Kruskal-Wallis and Dunn Test OK.

Rain Example (2)

```
require(dunn.test) # install package 'dunn.test' if needed  
dunn.test(rainfall,city,method='holm',alpha=.05)
```

```
##    Kruskal-Wallis rank sum test
```

```
##
```

```
## data: rainfall and city
```

```
## Kruskal-Wallis chi-squared = 17.8032, df = 3, p-value = 0
```

Output continued on next slide ...

Rain Example (3)

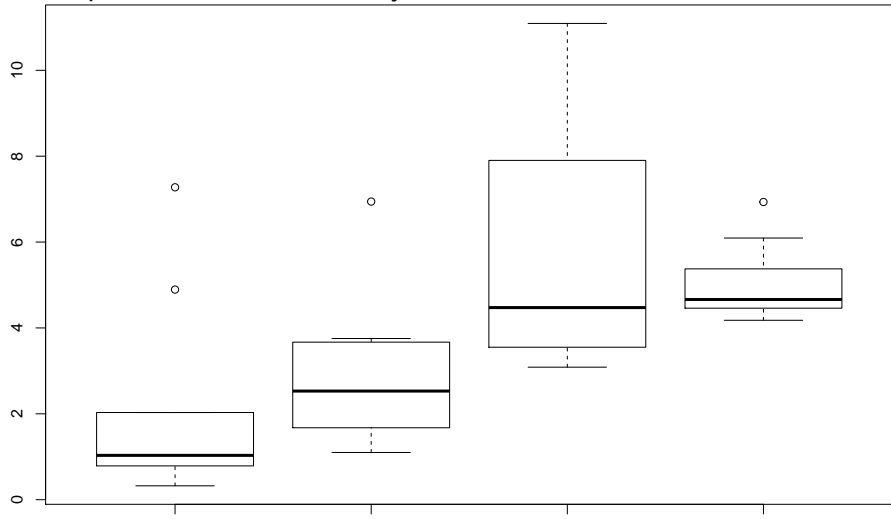
```
##                                     Comparison of rainfall by city
##                                     (Holm)
## Col Mean-|
## Row Mean |           A           B           C
## -----+-----
##          B |    0.612772
##          |    0.2700
##          |
##          C |   -2.205979   -2.818752
##          |    0.0411    0.0096*
##          |
##          D |   -2.982158   -3.594930   -0.776178
##          |    0.0072*    0.0010*    0.4376
```

Non-normal distributions + few contrasts

<i>Simultaneous Hypothesis Tests</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise	Tukey-Kramer or Games-Howell (use REGWQ if balanced and equal variances)	Dunn Test if appropriate, other more advanced procedures such as bootstrap
	Few pairs or contrasts	F or t tests with Bonferroni-Holm adjusted p-values	Pairwise test like Mann Whitney or bootstrap tests + adjusted p-values

Synthetic Example

4 samples of size 10 from very skewed distributions.



Synthetic Example (2)

- compare mean of groups A and B to mean of groups C and D
- compare means of A and B
- compare means of C and D
- use bootstrap t-tests in `onewayComp()`
- can also bootstrap with 'boot' package or 'WRS' package

Synthetic Example (3)

Build the contrast matrix:

```
K = rbind('aveCD -aveAB'=c(-1/2, -1/2, 1/2, 1/2),  
          '      B - A'=c( -1,    1,    0,    0),  
          '      D - C'=c(  0,    0,   -1,    1))
```

Synthetic Example (4)

```
onewayComp( x~g, var.equal=F, nboot=10000, con=K,  
            adjust = 'holm')$comp[,c(1,4,6,7)]
```

##		diff	t	p	adj	rej	H_0
##	aveCD -aveAB	2.9203362	4.4349448	0.0003			1
##	B - A	0.9142911	1.0158534	0.6748			0
##	D - C	-0.7050426	-0.7333225	0.6748			0

$$\text{Reject } H_0 : \frac{\mu_A + \mu_B}{2} = \frac{\mu_C + \mu_D}{2}$$

The other contrasts are not significant.

Big data

Exploring lots of groups looking for differences?

Use sequential procedures that control the FDR.

Recap

<i>Multiple Comparisons</i>		Distributions	
		Normal	Not Normal
Compare Type	All pairwise (<i>posthoc</i>)	Tukey-Kramer, Games-Howell, REGWQ	Dunn Test + adjust, resampling with adjustment, others
	Few pairs or contrasts (<i>planned</i>)	t-intervals + Bonferroni or sequential	Bootstrap or other with adjustment

Appendix: t-tests for contrasts

- Let k be the number of groups and N the pooled sample size.
- The formula for equal variances (next slide) is equivalent to the F test statistic in your textbook when all the samples are of size n (balanced).

Appendix: t-tests for contrasts (2)

For equal variances, use MSE from ANOVA for pooled variance estimate

$$s_p^2 = \text{MSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots$$

Now for contrast $\Psi = \sum_{i=1}^k a_i \mu_i$ we have $t = \frac{\sum_{i=1}^k a_i \bar{X}_i}{\sqrt{s_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}}$ with $\text{df} = N - k$.

If samples sizes are balanced with $n = n_1 = n_2 = \dots = n_k$, then the square of t is the same as F in the textbook. If there are only two weights $a_m = 1, a_n = -1$ this becomes the ordinary pooled t .

Appendix: t-tests for contrast (3)

For unequal variances (or quite unbalanced sample sizes). No pooled variance, but

Welch-like corrected degrees of freedom is harder. $t = \frac{\sum_{i=1}^k a_i \bar{X}_i}{\sqrt{\sum_{i=1}^k \frac{a_i^2 s_i^2}{n_i}}}$

For df, set $w_i = \frac{a_i^2 s_i^2}{n_i}$, then $df = \frac{\left(\sum_{i=1}^k w_i\right)^2}{\sum_{i=1}^k \frac{w_i^2}{n_i - 1}}$.

If there are only two weights $a_m = 1, a_n = -1$ this becomes the ordinary unpooled t .

A Bad Idea

- It's not OK to inspect the data and then test only the groups with the largest differences
- This is called data snooping.
- You've implicitly tested all of the pairs and so you have an increased risk of Type I errors due to chance.