STAT 500

ANOVA - Multiple Comparisons

Pairwise Comparisons

- Compare means for each pair of treatments
  - Each comparison is a contrast: $\mu_i - \mu_k$ for all $i \neq k$.
  - There are $\binom{r}{2}$ possible pairwise comparisons.
  - Set of $\binom{r}{2}$ comparisons are NOT orthogonal.
    - Example: $\mu_1 - \mu_2$ and $\mu_1 - \mu_3$.

Contrast	μ_1	μ_2	μ_3
1	1	-1	0
2	1	0	-1

Pairwise Comparisons

Using MS_{error} as the estimate of the common variance:

- $(1-\alpha) \times 100\%$ confidence intervals for difference in two means:

$$(\bar{Y}_i - \bar{Y}_k) \pm t_{N-r,1-\alpha/2}\sqrt{MS_{error}\left(\frac{1}{n_i}+\frac{1}{n_k}\right)}$$

- Hypothesis test to compare two-means:

$$t = \frac{\bar{Y}_i - \bar{Y}_k}{\sqrt{MS_{error}\left(\frac{1}{n_i}+\frac{1}{n_k}\right)}} \quad \text{with } N-r \text{ df}$$

Pairwise Comparisons

- Each pairwise comparison has type I error level α or confidence level $100(1-\alpha)\%$
- we do $\binom{r}{2}$ such comparisons!
- If r is large, some significant differences are expected by chance even if all of the means are the same.

Multiple Comparisons

- Known as the multiple comparisons problem

- When many comparisons are made, how should one interpret the p-value for a single comparison?

- Reminder: traditional p-value interpretation is derived from P(observe more extreme result | H_o is true).
  small p-value $\Rightarrow$ observed statistic unlikely if H_o is true
  i.e. reject H_o if observed result is "unusual"

- Doesn't have the same interpretation when many comparisons are made.

Multiple Comparisons Example

- Experiment with 10 treatments: $\binom{10}{2} = 45$ possible tests

- Case 1 - Prespecified contrast:

  Test # 10 is the only test you want to do:

  p-value for test # 10 has the usual interpretation.

  e.g. p-value $= 0.032 \Rightarrow$ evidence of significant difference between two means since p-value < 0.05.

Multiple Comparisons

- Case 2 - Post-hoc test (no pre-specified contrast):
  Test # 10 has the smallest p-value, p-value $= \mathbf{0.032}$,

  – With 45 *independent* tests, one would expect (45)(0.05) = 2.25 of the p-values to be smaller than 0.05 if all H_os are true.

  – A p-value of $\mathbf{0.032}$ is no longer unusual!

Multiple Comparisons

- Comparison-wise type I error rate:
  P[reject $\boldsymbol{H_0}$ for **one** test | $\boldsymbol{H_0}$ is true for that test]
- Experiment-wise type I error rate:
  P[reject at least one of the $\boldsymbol{H_0}$s | all $\boldsymbol{H_0}$s are true]
- Multiple comparisons adjustment: avoid too many *false* significant findings
- Make experiment-wise type I error rate reasonably small.
- Equivalent to simultaneous confidence intervals, i.e.
  all confidence intervals in a set include their individual targets with a specified probability.

Multiple Comparisons

- Basic approach is to adjust the $t_{N-r,1-\alpha/2}$ critical value used in individual $100(1-\alpha)\%$ confidence intervals or individual t-tests of size α.

- Cost is lower power: less likely to detect a non-zero effect.

- Benefit is that the experiment-wise type I error rate is no larger than the specified α

Multiple Comparisons

- Comparison-wise Type I error rate
  - Least Significant Difference
- Experiment-wise Type I error rate
  - Tukey-Kramer Honest Significant Difference (HSD)
  - Scheffe's
  - Bonferroni

Least Significant Difference (LSD)

- Conduct overall F-test of $H_0 : \mu_1 = ... = \mu_r$ at the α level
- If H_0 is not rejected then declare all means the same (chance of any false declarations of significant differences is less than α)
- If H_0 is rejected then calculate confidence intervals or conduct hypothesis tests
- Commonly used, but substantial loss of power when only a few groups have different means

LSD: Donut Example

$$LSD = t_{20,.975}\sqrt{MS_{error}\left(\frac{2}{n}\right)} = (2.086)\sqrt{100.9\left(\frac{2}{6}\right)} = 12.1$$

Declare a significant difference if $|\bar{Y}_i - \bar{Y}_j| \geq$LSD.

Order sample means from smallest to largest:

Oil 4	Oil 1	Oil 3	Oil 2
12	22	26	35

Tukey-Kramer Honest Significant Difference (HSD)

- Used to compare all pairs of treatment means
- Experiment-wise error rate is α for the entire set of the $\binom{r}{2}$ possible comparisons
- An exact solution for all pairwise comparisons with equal sample sizes (Tukey)
- Conservative for unequal sample sizes (Kramer modification)

Tukey-Kramer HSD

- Based on the distribution of studentized range

$$q_{(r,N-r)} = \left(\max_i \bar{Y}_i - \min_i \bar{Y}_i\right) / (S_p/\sqrt{n})$$

- Use $\frac{1}{\sqrt{2}} q_{(r,N-r,1-\alpha)}$ in CIs

- For tests, declare a significant difference if

$$|\bar{Y}_i - \bar{Y}_j| \geq \frac{1}{\sqrt{2}} q_{(r,N-r,1-\alpha)} \sqrt{MS_{error}\left(\frac{1}{n} + \frac{1}{n}\right)}$$

HSD: Donut Example

$$HSD = \frac{1}{\sqrt{2}} q_{(4,20,.95)} \sqrt{MS_{error}\left(\frac{2}{n}\right)}$$

$$= \frac{1}{\sqrt{2}} (3.958) \sqrt{100.9\left(\frac{2}{6}\right)} = 16.23$$

Declare a significant difference if $|\bar{Y}_i - \bar{Y}_j| \geq$HSD.
Order sample means from smallest to largest:

Oil 4	Oil 1	Oil 3	Oil 2
12	22	26	35

Scheffe'

- Works for any number of (actually all possible) linear contrasts
- Most conservative procedure, but relatively easy to apply
- use $\sqrt{(r-1)F_{r-1,N-r,1-\alpha}}$ in place of $t_{N-r,1-\alpha/2}$
- Declare a significant difference if

$$|\bar{Y}_i - \bar{Y}_j| \geq \sqrt{(r-1)F_{r-1,N-r,1-\alpha}}\sqrt{MS_{error}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Scheffe': Donut Example

$$Scheffe = \sqrt{3F_{3,20,.975}}\sqrt{MS_{error}\left(\frac{1}{6}+\frac{1}{6}\right)}$$

$$= \sqrt{(3)(3.098)}\sqrt{100.9\left(\frac{2}{6}\right)} = 17.68$$

(Table B.4 in KNNL)

Declare a significant difference if $|\bar{Y}_i - \bar{Y}_j| \geq$ Scheffe. Order sample means from smallest to largest:

Oil 4	Oil 1	Oil 3	Oil 2
12	22	26	35

Bonferroni

- If we have m tests (or confidence intervals), use α/m instead of α in each test (or confidence interval)
- Easy to implement
- Declare a significant difference if

$$|\bar{Y}_i - \bar{Y}_j| \geq t_{N-r,1-\frac{\alpha}{2m}} \sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Conservative, especially if m is large and tests are not independent (experiment-wide type I error rate $< \alpha$)
- Need to pre-specify the number of comparisons m

Bonferroni: Donut Example

Donut cooking oils:

There are $r = 4$ treatments and $m = 6$ pairs of means to compare.

$$Bonferroni = t_{20,1-\frac{0.05}{12}}\sqrt{MS_{error}\left(\frac{1}{6}+\frac{1}{6}\right)}$$

$$= (2.927)\sqrt{100.9\left(\frac{2}{6}\right)} = 16.975$$

Declare a significant difference if $|\bar{Y}_i - \bar{Y}_j| \geq$ Bonferroni. Order sample means from smallest to largest:

Oil 4	Oil 1	Oil 3	Oil 2
12	22	26	35

```
proc glm data=donut;
  class oil;
  model y = oil / p;
  estimate 'o4-(o1+o2+o3)/3' oil -1 -1 -1 3 / divisor=3;
  estimate 'o2-(o1+o3)/2' oil -0.5 1 -0.5 0;
  estimate 'o1-o3' oil 1 0 -1 0 ;
  contrast 'o4-(o1+o2+o3)/3' oil -1 -1 -1 3 ;
  contrast 'o2-(o1+o3)/2' oil -0.5 1 -0.5 0;
  contrast 'o1-o3' oil 1 0 -1 0 ;
  means oil /alpha=.05 bon lsd scheffe tukey snk;
  output out=set2 residual=r predicted=yhat;
  run;
```

The GLM Procedure

t Tests (LSD) for y

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	100.9
Critical Value of t	2.08596
Least Significant Difference	12.097

Means with the same letter are not significantly different.

t Grouping		Mean	N	Oil
	A	35.000	6	2
	A			
B	A	26.000	6	3
B				
B	C	22.000	6	1
	C			
	C	12.000	6	4

The GLM Procedure

Student-Newman-Keuls Test for y

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

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	100.9

Number of Means	**2**	**3**	**4**
Critical Range	12.097171	14.672442	16.232038

Means with the same letter are not significantly different.

SNK Grouping		**Mean**	**N**	**Oil**
	A	35.000	6	2
	A			
B	A	26.000	6	3
B	A			
B	A	22.000	6	1
B				
B		12.000	6	4

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

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
Error Degrees of Freedom	20
Error Mean Square	100.9
Critical Value of Studentized Range	3.95825
Minimum Significant Difference	16.232

Means with the same letter are not significantly different.

Tukey Grouping		Mean	N	oil
	A	35.000	6	2
	A			
B	A	26.000	6	3
B	A			
B	A	22.000	6	1
B				
B		12.000	6	4

The GLM Procedure

Bonferroni (Dunn) t Tests for y

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
Error Degrees of Freedom	20
Error Mean Square	100.9
Critical Value of t	2.92712
Minimum Significant Difference	16.976

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	oil
	A	35.000	6	2
	A			
B	A	26.000	6	3
B	A			
B	A	22.000	6	1
B				
B		12.000	6	4

The GLM Procedure

Scheffe's Test for y

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

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	100.9
Critical Value of F	3.09839
Minimum Significant Difference	17.681

Means with the same letter are not significantly different.				
Scheffe Grouping		**Mean**	**N**	**oil**
	A	35.000	6	2
	A			
B	A	26.000	6	3
B	A			
B	A	22.000	6	1
B				
B		12.000	6	4

Multiple Comparison Procedures

Many, many other multiple comparison techniques

- Dunnet's procedure to compare each of r-1 treatment means to the mean for a control group
- Step down procedures, like the Student-Newman-Kuels (SNK) procedure, increase power
- Decision theory inspired procedures like Duncan's multiple range procedure
- Methods to control false discovery rates in genomic experiments

Multiple Comparison Procedures

- Set n=10 observations per group.
  Consider: r=3, 3 comparisons, error df=27
  r=10, 45 comparisons, error df=90
- Compare critical values, $\alpha = 0.05$

Method	r=3	r=10
LSD (unadjusted t)	2.05	1.99
Tukey-Kramer (q/$\sqrt{2}$)	2.48	3.25
Bonferroni	2.55	3.37
Scheffe'	2.59	4.23

- Power to detect $\delta = 1.4$, $\sigma = 1$, n=10

Method	r=3	r=10
LSD (unadjusted t)	84%	84%
Bonferroni	68%	27%

Multiple Comparison Procedures

- Many possible approaches, many different opinions
- My philosophy: Treatment comparisons should be pre-selected to answer specific questions
- When a study has a relatively small number of planned comparisons or contrasts.
  - perform tests or construct confidence intervals with Bonferroni adjustments
  - Prefer but don't require orthogonal contrasts (simple to interpret)
  - Use SNK or HSD to compare all pairs of means.

Multiple Comparison Procedures

- When a large number of unplanned comparisons are examined
  - What is the appropriate family of comparisons (all pairs of means, all possible contrasts)
  - Use the most powerful appropriate multiple comparison procedure (SNK, Bonferroni, Sheffe)
- Confidence intervals
  - Do I need an interval for only one comparison?
  - Or simultaneous intervals for several comparisons? Use multiple comparison adjustment
    - HSD for pairs of means
    - Bonferroni for a few contrasts
    - Scheffe for unlimited contrasts

False discovery rate (FDR)

- In genomic studies, we are simultaneously testing a huge number of hypotheses, each relating to a feature.
  - Gene expression experiments: tens of thousands of genes are measured for each sample
  - GWAS studies: millions of SNPs are tested simultaneously.
  - Microbiome studies: depending on the precision, there could be thousands of OTUs measured for each sample.

False discovery rate (FDR)

- Suppose one test of interest has been conducted for each of the m genes in an RNA-seq experiment to study gene expression.
- Often, the test of interest is that for each gene, whether the mean expression levels change (i.e., the gene is differentially expressed) between different treatments or not.
- Let $H_{01}, H_{02}, \cdots, H_{0m}$ denote the null hypotheses (interpreted as non-differential expression) corresponding to the m tests (genes).
- If we set level 5% for each test, the number of type I error (false positives) is expected to be 5% of m_0, the total number of genes that are not differentially expressed.

False discovery rate (FDR)

- Considering the high dimensionality of RNA-seq data, the number of errors is expected to be really big if we only control error at the level of individual test.

- Considering the high dimensionality of RNA-seq data, multiple comparison adjustments such as Bonferroni's method is impractical.

False discovery rate (FDR)

- FDR (or pFDR = positive FDR) is an alternative error rate that can be useful for RNA-seq experiments or other genomic studies.
- Table of Outcomes for m Tests

Hypothesis	Accept Null	Reject Null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

- FDR (Benjamini and Hochberg, 1995)

$$FDR = E\left(\frac{V}{R}\middle| R > 0\right) Pr(R > 0)$$

A Conceptual Description of FDR

- Suppose a scientist conducts 100 independent RNA-seq experiments.
- For each experiment, the scientist produces a list of genes declared to be differentially expressed by testing a null hypothesis for each gene.
- For each list consider the ratio of the number of false positive results to the total number of genes on the list (set this ratio to 0 if the list contains no genes).
- The FDR is approximated by the average of the ratios described above.

Is Experiment-Wise Error Rate Too Conservative for Genomic Studies?

- Suppose that one of the 100 gene lists consists of 500 genes declared to be differentially expressed.
- Suppose that 1 of those 500 genes is not truly differentially expressed but that the other 499 are.
- This list is considered to be in error and such lists are allowed to make up only a small proportion of the total number of lists if experiment-wise error rate is to be controlled.
- However such a list seems quite useful from the scientific viewpoint.

FDR: The Appropriate Error Rate for Genomic Studies?

- The hypothetical gene list discussed previously with 1 false positive and 499 true positives would be a good list that would help to keep the FDR down.

- Some of the gene lists may contain a high proportion of false positive results and yet the method we are using may still control FDR at a given level because it is the average performance across repeated experiments that matters. (This comment applies to the control of experiment-wise error rate as well.)