ANOVA Post-hoc Testing

BIOS 6611

CU Anschutz

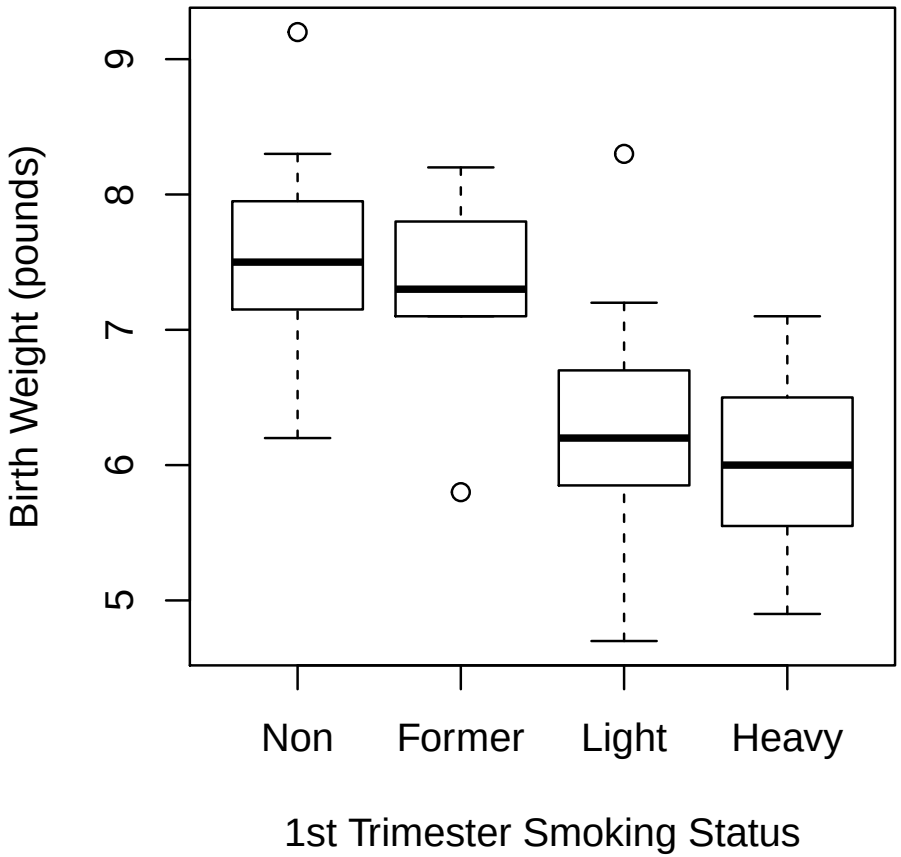
Week 11

Motivation

Motivating Example

Our motivating example will be infant birthweight (pounds) and smoking status of mother during the first trimester.

	Smoking Status			
i	**Non**	**Former**	**Light**	**Heavy**
1	7.5	5.8	5.9	6.2
2	6.2	7.3	6.2	6.8
3	6.9	8.2	5.8	5.7
4	7.4	7.1	4.7	4.9
5	9.2	7.8	8.3	6.2
6	8.3		7.2	7.1
7	7.6		6.2	5.8
8				5.4
$\bar{Y}_j$	7.59	7.24	6.33	6.01
s_j^2	0.93	0.83	1.3	0.52



Motivation

Our one-way ANOVA model tests the *global* hypothesis that all group means are equal versus *at least one* group mean is different:

$H_0 : \mu_1 = \mu_2 = ... = \mu_J$

H_1 : at least one of the means is different

If we reject our null hypothesis, a natural follow-up question is what group or groups are different?

One challenge with addressing this question is how to handle multiple comparisons without inflating our family-wise (overall) type I error rate.

We will introduce some approaches to **post-hoc testing** when the global null hypothesis is reject for our one-way ANOVA that assumes equal variances.

Post-Hoc Strategy 1: No Formal Correction for Multiple Comparisons

Pairwise Two-Sample t-tests

One simple approach to test what groups have significantly different means would be to conduct pairwise t-tests:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ where } s = \sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}}$$

For our birthweight data, this would result in p-values (from `t.test` in `R`) of:

	Former	**Light**	**Heavy**
Non	0.543	0.046	0.005
Former		0.156	0.038
Light			0.542

Incorporating the Equal Variance Assumption

However, if the variances from all groups are assumed to be equal, then a more accurate estimate of σ could be obtained. The pooled estimate of the variance for one-way ANOVA:

$$s^2 = \frac{\sum_{j=1}^{J} (n_j - 1) s_j^2}{\sum_{j=1}^{J} (n_j - 1)}$$

The t-test then becomes:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{N-J}$$

This test statistic is used for the **least significant difference (LSD)** method.

LSD Considerations

The LSD post-hoc test does not really correct for multiple comparisons. Instead, it uses a *pooled* estimate of the standard deviation, which provides more degrees of freedom and power.

In the special case when we only have 3 groups (i.e., $J = 3$), the family-wise error rate is controlled. As J gets larger, our desired overall type I error rate is no longer controlled.

Therefore, in practice, we may still want to apply a Bonferroni or false discovery rate (FDR) correction if $J > 3$, or use other post-hoc testing methods.

LSD Example Code

In SAS we can implement LSD by adding it to PROC ANOVA:

```
proc anova data=BWT;
    class momsmoke;
    model birthwt = momsmoke;
    means momsmoke / LSD;
run;
```

In R we can implement LSD by using either PostHocTest() (DescTools package) or pairwise.t.test() (stats package):

```
BWT <- read.csv('birthweight_smoking_dataset.csv', header=T)

library(DescTools)
aov1 <- aov( birthwt ~ momsmoke , data=BWT)
PostHocTest(aov1, method=c('lsd')) # results on next slide

pairwise.t.test( x=BWT$birthwt, g=BWT$momsmoke, p.adjust.method='none')
```

LSD Example

```
##
##   Posthoc multiple comparisons of means : Fisher LSD
##     95% family-wise confidence level
##
## $momsmoke
##                     diff     lwr.ci     upr.ci   pval
## Heavy-Former -1.2275000 -2.3355337 -0.1194663 0.0314 *
## Light-Former -0.9114286 -2.0494958  0.2266386 0.1112
## Non-Former    0.3457143 -0.7923529  1.4837815 0.5359
## Light-Heavy   0.3160714 -0.6898473  1.3219902 0.5221
## Non-Heavy     1.5732143  0.5672955  2.5791331 0.0037 **
## Non-Light     1.2571429  0.2182344  2.2960513 0.0199 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One way to visually summarize the results is to draw lines between groups without significant differences:

Heavy (6.01 lbs)	Light (6.33 lbs)	Former (7.24 lbs)	Non (7.59 lbs)

Post-Hoc Strategy 2: Formal Correction for Multiple Comparisons

Methods that Correct for Multiple Comparisons

Whereas the LSD method does *not* truly correct for multiple comparisons when we have more than 3 groups, many other methods have been proposed.

We will focus on 3 in the context of post-hoc testing for a one-way ANOVA (in order from most to least conservative):

- **Bonferroni Adjustment:** Can be used for any C independent comparisons. Essentially you conclude that the p-value is significant if it is less than $\frac{0.05}{C}$ instead of 0.05. Also known as *Dunn's Test*.
- **Tukey's Honestly Significant Difference (HSD):** Uses the studentized range distribution to make all pairwise comparisons. The *Games-Howell test* is a similar post-hoc test for Welch's ANOVA.
- **Dunnett's Test:** Used to compare several groups to a single control group; often used in clinical trials.

Multiple Comparisons Example Code

In SAS we can implement these methods by adding them to PROC GLM:

```
PROC GLM DATA = BWT ORDER = internal;
  CLASS momsmoke;
  MODEL birthwt = momsmoke/noint solution;
  MEANS momsmoke/ dunnett('Non') bon tukey;
RUN;
```

In R we can use functions in the DescTools or stats packages:

```
aov1 <- aov( birthwt ~ momsmoke , data=BWT) # fit one-way ANOVA

# Bonferroni/Dunn's Test
DescTools::PostHocTest(aov1, method=c('bonferroni'))
pairwise.t.test(x=BWT$birthwt, g=BWT$momsmoke, p.adjust.method='bonferroni')

# Tukey's HSD
DescTools::PostHocTest(aov1, method=c('hsd'))
TukeyHSD(aov1)

# Dunnett's Test
DescTools::DunnettTest( x=BWT$birthwt, g=BWT$momsmoke, control='Non')
```

Bonferroni/Dunn's Test Example

```
DescTools::PostHocTest(aov1, method=c('bonferroni'))
```

```
##
##   Posthoc multiple comparisons of means : Bonferroni
##     95% family-wise confidence level
##
## $momsmoke
##                     diff     lwr.ci    upr.ci   pval
## Heavy-Former -1.2275000 -2.7734676 0.3184676 0.1885
## Light-Former -0.9114286 -2.4992999 0.6764427 0.6670
## Non-Former    0.3457143 -1.2421570 1.9335856 1.0000
## Light-Heavy   0.3160714 -1.0874218 1.7195646 1.0000
## Non-Heavy     1.5732143  0.1697211 2.9767075 0.0219 *
## Non-Light     1.2571429 -0.1923787 2.7066644 0.1191
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heavy (6.01 lbs)	Light (6.33 lbs)	Former (7.24 lbs)	Non (7.59 lbs)

Tukey's HSD Example

```
DescTools::PostHocTest(aov1, method=c('hsd'))
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##
## $momsmoke
##                     diff     lwr.ci    upr.ci   pval
## Heavy-Former -1.2275000 -2.7097495 0.2547495 0.1293
## Light-Former -0.9114286 -2.4338548 0.6109976 0.3684
## Non-Former    0.3457143 -1.1767119 1.8681405 0.9219
## Light-Heavy   0.3160714 -1.0295759 1.6617188 0.9145
## Non-Heavy     1.5732143  0.2275669 2.9188616 0.0179 *
## Non-Light     1.2571429 -0.1326357 2.6469215 0.0860 .
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heavy (6.01 lbs)	Light (6.33 lbs)	Former (7.24 lbs)	Non (7.59 lbs)

Dunnett's Test Example

```
DescTools::DunnettTest( x=BWT$birthwt, g=BWT$momsmoke, control='Non')
```

```
##
##   Dunnett's test for comparing several treatments with a control :
##     95% family-wise confidence level
##
## $Non
##                  diff    lwr.ci       upr.ci   pval
## Former-Non -0.3457143 -1.730947  1.039518404 0.8671
## Heavy-Non  -1.5732143 -2.797599 -0.348830000 0.0099 **
## Light-Non  -1.2571429 -2.521682  0.007395795 0.0516 .
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Recall, for Dunnett's we are only making pairwise comparisons to our "control" group (here it is the never smokers). So we cannot directly draw a comparison between all possible pairwise group comparisons.

Closing Comments

These different methods allow us to control the family-wise type I error rate to varying degrees. The "best" method will truly be context specific.

One situation to be aware of is when the one-way ANOVA indicates a significant difference, but the post-hoc test does not. This will depend on the test chosen and how strongly it controls the family-wise error rate.

In practice, if we *a priori* know a set of pairwise comparisons are of interest, we should just design our analysis around this specific tests to best control our type I error rate and maximize power.