ANOVA continued

Today's lecture

- Worked example of ANOVA testing
- What happens after a significant ANOVA
  - pairwise comparisons
  - adjustment for multiple testing

The test statistic (ANOVA F Statistic)

$$F = \frac{\text{variation among group means}}{\text{variation among individuals in the same group}}$$

Example: Cannabis to treat brain cancer

Example: Cannabis to treat brain cancer (in mice)

High-grade glioma is an aggressive type of brain cancer with a low long-term survival rate. Cannabinoids, which are chemical compounds found in cannabis, are thought to inhibit glioma cell growth. Researchers transplanted glioma cells in otherwise-healthy mice, and then randomly assigned these mice to 4 cancer treatments: irradiation alone, cannabinoids, alone, irradiation combined with cannabinoids, or no cancer treatment. The treatments were administered for 21 days, after which the glioma tumor volume (in cubic millimeters) was assessed in each mouse using brain imaging.

The data

```
head(cancer_data)
```

```
##      treatment tumor_volume
## 1  Irradiation           30
## 2  Irradiation           46
## 3  Irradiation           46
## 4  Irradiation           95
## 5 Cannabinoids           12
## 6 Cannabinoids           14
```

Organize the data

- Think about how you want the data to look.
- I want to plot the raw data points and display the mean for each treatment group
- I also want to specify the order that the treatment groups show up in the graph

```
# specify the order of the treatment groups for plotting
library(forcats)
cancer_data <- cancer_data %>%
  mutate(trt_order = fct_relevel(treatment, c("Neither", "Irradiation",
                                              "Cannabinoids", "Both")))
```

Look at summary statistics

```
# calculate the means and SD for each group
summary_stats <- cancer_data %>%
  group_by(trt_order) %>%
  summarise(mean_vol = mean(tumor_volume),
            sd_vol = sd(tumor_volume),
            samp_size = n())

summary_stats
```

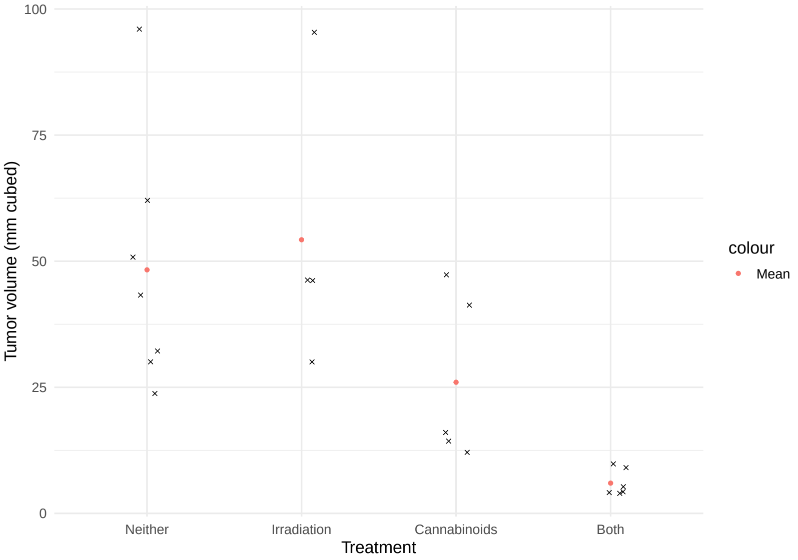
```
## # A tibble: 4 x 4
##   trt_order    mean_vol sd_vol samp_size
##   <fct>           <dbl>  <dbl>     <int>
## 1 Neither          48.3  24.8          7
## 2 Irradiation      54.2  28.2          4
## 3 Cannabinoids     26    16.6          5
## 4 Both              6     2.76         6
```

Graph the data

```
ggplot(cancer_data, aes(x = trt_order, y = tumor_volume)) +
geom_jitter(pch = 4, width = 0.1) +
geom_point(data = summary_stats, aes(y = mean_vol, col = "Mean"), pch = 19) +
labs(y = "Tumor volume (mm cubed)", x = "Treatment") +
theme_minimal(base_size = 15)
```

note: geom_jitter() with width = 0.1 randomly "jitters" the location of the points along the x axis so that we can see each of them since some have the exact same values.

Graph the data



ANOVA in R: use aov(), then tidy() it up!

```
library(broom)
cancer_anova <- aov(formula = tumor_volume ~ treatment, data = cancer_data)
tidy(cancer_anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic  p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>    <dbl>
## 1 treatment     3 8060.  2687.      6.70  0.00313
## 2 Residuals    18 7218.   401.     NA    NA
```

This F says that the variation between the means is nearly 7 times as large as the variation within the groups.

This p-value is equal to 0.3%. There is a 0.3% chance of observing the F statistic we observed (or more extreme) under the null hypothesis that all the means are the same.

P value of the F test

Remember

- k is the number of groups being compared and $N_{Total} = n_1 + n_2 + ... + n_k$ is the total sample size across all the groups.
- The F statistic follows an F distribution with $k - 1$ degrees of freedom in the numerator and $N_{Total} - k$ degrees of freedom in the denominator
- The p-value of the ANOVA F statistic is always the area to the right of the test statistic

```r
pf(6.699489, df1 = 3, df2 = 22 - 4, lower.tail = F)
```

```
## [1] 0.003131703
```

What now?

- The p-value equaled 0.003, indicating a difference. But what groups are actually different?

After a significant ANOVA. . .

next steps. . .

- You could look at all pairwise differences (i.e., comparing each combination of treatments), but we have to be careful because we will find differences "just by chance" if we compare enough groups.

A key example of what not to do?

FOOD FOR THOUGHT

Cornell Food Researcher's Downfall Raises Larger Questions For Science

September 26, 2018 · 3:07 PM ET

BRETT DAHLBERG

FROM WXXI

A brief reminder about p-hacking

Remember, one of the issues with multiple comparisons is that when you repeatedly question the same dataset, you can end up finding "significant" results by chance alone.

We talked about this before as p-hacking or p-fishing or data dredging

This along with other issues that are sometimes unconscious can lead to bias in what is found and what is published.

Ioannidis, John P.A. (August 30, 2005). "Why Most Published Research Findings Are False". PLoS Medicine. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Strategies for multiple comparisons

- Conduct multiple pairwise tests and adjust the critical value for significance using a Bonferroni correction
- Use a Tukey's HSD to generate the pairwise comparisons with adjusted p-values

Bonferroni

Bonferroni Adjustment for multiple comparisons

We could do a series, comparing each combination of groups in pairs or $\binom{k}{2}$ comparisons.

To compensate for making multiple comparisons and set the overall probability of making a type I error at 0.05, we can adjust our α to $\alpha*$ for each comparison by dividing by the number of comparisons we are making.

$$\alpha* = \left(\frac{0.05}{\binom{k}{2}}\right)$$

We then use $\alpha*$ as the significance level for each individual comparison. So for a comparison of 3 groups we would use an α of 0.0167 as the significance level for each comparison.

This modification is known as the Bonferroni correction. Bonferroni is fairly basic and can become unwieldy - what happens if you have a lot of groups?

Bonferroni for our example

This would entail doing a series of two-sample t tests for each pairwise comparison or 4 choose 2 comparisons = 6 two-sample tests so the adjustment would be:

$$\alpha* = \left(\frac{0.05}{\binom{4}{2}}\right) = \left(\frac{0.05}{6}\right) = 0.0083$$

Bonferroni for our example

So we would conduct each pairwise test:

```
both<-cancer_data%>%filter(treatment=="Both")%>%select(tumor_volume)
neither<-cancer_data%>%filter(treatment=="Neither")%>%select(tumor_volume)
can<-cancer_data%>%filter(treatment=="Cannabinoids")%>%select(tumor_volume)
irr<-cancer_data%>%filter(treatment=="Irradiation")%>%select(tumor_volume)
n_b<-t.test(neither, both, alternative="two.sided")
c_b<-t.test(can,both, alternative="two.sided")
i_b<-t.test(irr, both, alternative="two.sided")
n_c<-t.test(neither, can, alternative="two.sided")
n_i<-t.test(neither, irr, alternative="two.sided")
c_i<-t.test(can, irr, alternative="two.sided")
```

Bonferroni for our example

and compare the p-value of this test to an alpha of 0.0083

```
t_statistics<-c(n_b$statistic, c_b$statistic, i_b$statistic, n_c$statistic, n
p_values<-c(n_b$p.value, c_b$p.value, i_b$p.value, n_c$p.value, n_i$p.value)
t_statistics
```

```
##         t         t         t         t         t         t
##  4.479504  2.659192  3.411826  1.862690 -0.352323 -1.772461
```

```
p_values
```

```
## [1] 0.003912559 0.053873393 0.041289736 0.092130403 0.737291103
```

Tukey's HSD

Tukey's honest significant differences (Tukey's HSD)

John Wilder Tukey

Tukey's honestly significant differences (Tukey's HSD)

- Tukey's test maintains a 5% experimentwise or "family" error rate.
- Even if you make many pairwise comparisons, the overall error rate is 5% (at most)
- Overcomes the issue of multiple testing. Recall: If you conducted 100 tests with a 5% error rate (i.e., $\alpha = 0.05$) AND the H_0 was always true, how many p-values would you expect to be < 0.05?
- The Tukey's error rate is 5% overall, no matter how many tests you do. Thus it overcomes the problem of multiple testing

TukeyHSD() to calculate the differences in R

You can think of the TukeyHSD() as a wrap-around for the anova, you can either nest the statements like this:

TukeyHSD(aov(outcome ~ group))

or save the ANOVA as an object and use that in the statement:

modelresult<-aov(outcome ~ group)

TukeyHSD(modelresult, overall_alpha)

TukeyHSD() to calculate the differences in R

Here is the R code and output for the cancer example:

```
diffs <- TukeyHSD(cancer_anova, conf.level = 0.05) %>% tidy()
diffs%>%select(contrast, adj.p.value)
```

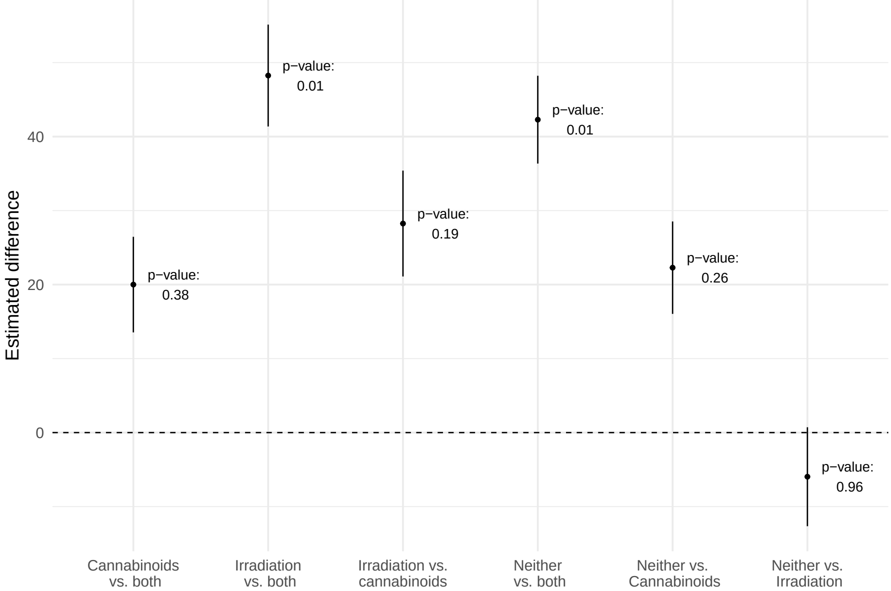
```
## # A tibble: 6 x 2
##   contrast                  adj.p.value
##   <chr>                           <dbl>
## 1 Cannabinoids-Both             0.378
## 2 Irradiation-Both              0.00756
## 3 Neither-Both                  0.00661
## 4 Irradiation-Cannabinoids      0.190
## 5 Neither-Cannabinoids          0.263
## 6 Neither-Irradiation           0.964
```

`TukeyHSD()` to calculate the differences in R

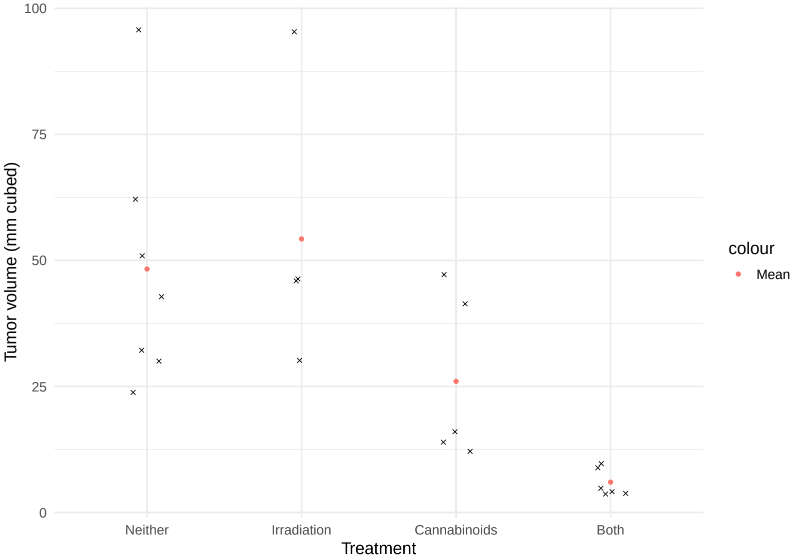
Each row in the table corresponds to a pairwise test. So the first row is looking at the difference between Cannabinoids vs. Both treatments. The estimated difference in means is 20 and the 95% CI is 13.54 to 26.45. The adjusted p-value is 0.38.

- "Adjusted" means that it is adjusted for conducting multiple tests. The unadjusted p-value would be smaller. You can tell the unadjusted p-value would be < 0.05 because the 95% CI doesn't include 0.
- Thus, when you have an adjusted test you can't use the CI to infer the value of the p-value!

Visualize the pairwise differences



Review raw data for comparison



Parting humor XKCD.com