431 Class 19

github.com/THOMASELOVE/2019-431

2019-11-05

Today's Agenda

- Comparing 3 or more Population Means with the Analysis of Variance
- Indicator Variable Regression Analysis
- Interpreting the ANOVA table
- ANOVA assumptions and the Kruskal-Wallis test
- The Problem of Multiple Comparisons
    - Bonferroni pairwise testing
    - Tukey HSD pairwise comparisons

Today's Setup and Data

```r
library(readxl) # to read in an .xlsx file
library(magrittr); library(janitor)
library(broom); library(here)
library(tidyverse)

source(here("R", "Love-boost.R"))
```

County Health Rankings Data for Ohio, 2018

Data Source:
http://www.countyhealthrankings.org/app/ohio/2018/downloads

In the `ohio_2018.xlsx` file I have provided to you, each row describes one of Ohio's 88 counties in terms of:

- `FIPS` code (basically an identifier for mapping)
- `state` and `county` name
- health outcomes (standardized - more positive means better outcomes)
- health behavior ranking (1-88, we'll divide into 4 groups)
- clinical care ranking (1-88, we'll split into 3 groups)
- population density (urban or rural)
- median income, in dollars

Importing the Data / Creating some Factors

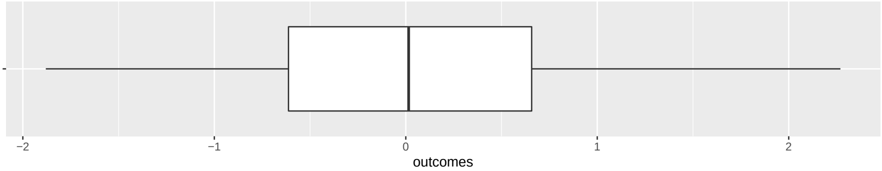
```r
ohio18 <- read_xlsx(here("data", "ohio_2018_rankings.xlsx")) %
  mutate(behavior = Hmisc::cut2(rk_behavior, g = 4),
         clin_care = Hmisc::cut2(rk_clin_care, g = 3)) %>%
  mutate(behavior = fct_recode(behavior,
            "Best" = "[ 1,23)", "High" = "[23,45)",
            "Low" = "[45,67)", "Worst" = "[67,88]")) %>%
  mutate(clin_care = fct_recode(clin_care,
            "Strong" = "[ 1,31)", "Middle" = "[31,60)",
            "Weak" = "[60,88]")) %>%
  mutate(density = factor(density)) %>%
  select(FIPS, state, county, outcomes,
         behavior, clin_care, density, income)
```

A Quick Look at the Data

```
ohio18 %>% filter(county == "Cuyahoga") %>%
  select(FIPS, county, outcomes, behavior, clin_care)
```

```
# A tibble: 1 x 5
  FIPS  county   outcomes behavior clin_care
  <chr> <chr>       <dbl> <fct>    <fct>
1 39035 Cuyahoga    -0.38 Low      Strong
```

```
ggplot(ohio18, aes(x = "", y = outcomes)) +
  geom_boxplot() + coord_flip() + labs(x = "")
```



Use inspect to inspect the data frame?

```r
ohio18 %>% select(outcomes, behavior, clin_care) %>%
  mosaic::inspect()
```

```
categorical variables:
       name  class levels  n missing
1  behavior factor      4 88       0
2 clin_care factor      3 88       0
                                   distribution
1 Best (25%), High (25%) ...
2 Strong (34.1%), Middle (33%) ...

quantitative variables:
      name   class   min      Q1 median     Q3  max
1 outcomes numeric -1.88 -0.6125  0.015 0.6575 2.27
           mean        sd  n missing
1 -0.0001136364 0.8940885 88       0
```

Key Measure Details

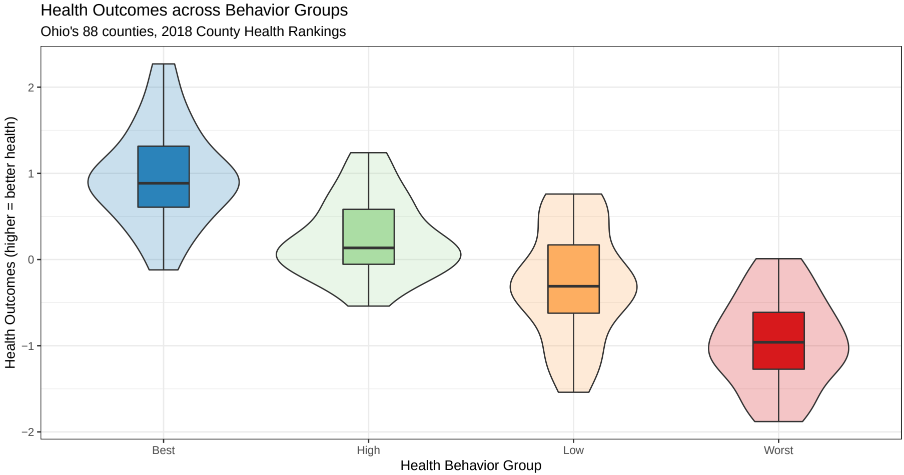
- **outcomes** = quantity that describes the county's premature death and quality of life results, weighted equally and standardized (z scores).
  - Higher (more positive) values indicate better outcomes in this county.
- **behavior** = (Best/High/Low/Worst) reflecting adult smoking, obesity, food environment, inactivity, exercise, drinking, alcohol-related driving deaths, sexually tranmitted infections and teen births.
  - Counties in the Best group had the best behavior results.
- **clin_care** = (Strong/Middle/Weak) reflects rates of uninsured, care providers, preventable hospital stays, diabetes monitoring and mammography screening.
  - Strong means that clinical care is strong in this county.

Our Questions

1. Do average health outcomes vary significantly across groups of counties defined by health behavior?
2. Do groups of counties defined by clinical care show meaningful differences in average health outcomes?

Question 1

Do average health outcomes differ by health behavior?



Source: http://www.countyhealthrankings.org/app/ohio/2018/downloads

Question 1 Numerical Summaries

Do average health outcomes vary significantly across groups of counties defined by health behavior?

```
mosaic::favstats(outcomes ~ behavior, data = ohio18) %>%
  knitr::kable(digits = 2)
```

behavior	min	Q1	median	Q3	max	mean	sd	n	missing
Best	-0.12	0.61	0.88	1.31	2.27	0.97	0.57	22	0
High	-0.54	-0.06	0.14	0.58	1.24	0.25	0.45	22	0
Low	-1.54	-0.62	-0.31	0.17	0.76	-0.28	0.65	22	0
Worst	-1.88	-1.27	-0.96	-0.61	0.01	-0.95	0.52	22	0

Note that there is no missing data here.

Analysis of Variance (ANOVA) testing: Question 1

Does the mean `outcomes` result differ across the `behavior` groups?

$$H_0 : \mu_{Best} = \mu_{High} = \mu_{Low} = \mu_{Worst} \text{ vs. } H_A : \text{At least one } \mu \text{ is different.}$$

To test this set of hypotheses, we will build a linear model to predict each county's outcome based on what behavior group the county is in.

- We then look at whether the `behavior` group effect has a statistically significant impact on the model's predictions of `outcomes`.
- If `behavior` has a significant effect in that model, it means that we reject H_0 in favor of H_A.

Building the Linear Model: Question 1

Are there statistically significant differences in mean outcome across the behavior group means?

```r
model_one <- lm(outcomes ~ behavior, data = ohio18)
model_one
```

```

Call:
lm(formula = outcomes ~ behavior, data = ohio18)

Coefficients:
  (Intercept)   behaviorHigh    behaviorLow  
       0.9718        -0.7186        -1.2495  
behaviorWorst  
      -1.9195  
```

How do we interpret this model?

Interpreting the Indicator Variables

The regression model (`model_one`) equation is

```
outcomes = 0.97 - 0.72 behaviorHigh
                    - 1.25 behaviorLow
                      - 1.92 behaviorWorst
```

What do the indicator variables mean?

group	behaviorHigh	behaviorLow	behaviorWorst
Best	0	0	0
High	1	0	0
Low	0	1	0
Worst	0	0	1

- So what is the predicted `outcomes` score for a county in the High behavior group, according to this model?

Interpreting the Indicator Variables

The regression model (`model_one`) equation is

```
outcomes = 0.97 - 0.72 behaviorHigh
                    - 1.25 behaviorLow
                      - 1.92 behaviorWorst
```

What predictions does the model make?

group	`High`	`Low`	`Worst`	Prediction
Best	0	0	0	0.97
High	1	0	0	0.97 - 0.72 = 0.25
Low	0	1	0	0.97 - 1.25 = -0.28
Worst	0	0	1	0.97 - 1.92 = -0.95

Do these predictions make sense?

Interpreting the Indicator Variables

The regression model (`model_one`) equation is

```
outcomes = 0.97 - 0.72 behaviorHigh
                    - 1.25 behaviorLow
                      - 1.92 behaviorWorst
```

Recall that the sample data shows...

```
ohio18 %>% group_by(behavior) %>%
  summarize(n = n(), mean = round(mean(outcomes),2))
```

```
# A tibble: 4 x 3
  behavior      n  mean
  <fct>     <int> <dbl>
1 Best         22  0.97
2 High         22  0.25
3 Low          22 -0.28
4 Worst        22 -0.95
```

ANOVA for the Linear Model: Question 1

Are there statistically significant differences in mean outcome across the behavior group means?

$$H_0 : \mu_{Best} = \mu_{High} = \mu_{Low} = \mu_{Worst} \text{ vs. } H_A : \text{At least one } \mu \text{ is different.}$$

```r
anova(model_one)
```

```
Analysis of Variance Table

Response: outcomes
          Df Sum Sq Mean Sq F value    Pr(>F)
behavior   3 43.645 14.5482  47.179 < 2.2e-16 ***
Residuals 84 25.903  0.3084
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, what's in the ANOVA table? (df)

The ANOVA table reports here on a single **factor** (behavior group) with 4 levels, and on the residual variation in health **outcomes** not accounted for by that factor.

```r
anova(model_one)[1:4]
```

```
          Df Sum Sq Mean Sq F value
behavior   3 43.645 14.5482  47.179
Residuals 84 25.903  0.3084
```

Degrees of Freedom (df) is an index of sample size...

- df for our factor (behavior) is one less than the number of categories. We have four behavior groups, so 3 degrees of freedom.
- Adding df(behavior) + df(Residuals) = 3 + 84 = 87 = df(Total), one less than the number of observations (counties) in Ohio.
- n observations and g groups yield $n - g$ residual df in a one-factor ANOVA table.

So, what's in the ANOVA table? (Sum of Squares)

```
anova(model_one)[1:4]
```

```
          Df Sum Sq Mean Sq F value
behavior   3 43.645 14.5482  47.179
Residuals 84 25.903  0.3084
```

Sum of Squares (`Sum Sq`, or SS) is an index of variation...

- SS(factor), here SS(`behavior`) measures the amount of variation accounted for by the `behavior` groups in our `model_one`.
- The total variation in `outcomes` to be explained by the model is SS(factor) + SS(Residuals) = SS(Total) in a one-factor ANOVA table.
- We describe the proportion of variation explained by a one-factor ANOVA model with η^2 ("eta-squared": same as Multiple R^2)

$$\eta^2 = \frac{SS(\text{behavior})}{SS(\text{Total})} = \frac{43.645}{43.645 + 25.903} = \frac{43.645}{69.548} \approx 0.628$$

So, what's in the ANOVA table? (MS and F)

```
anova(model_one)[1:4]
```

```
          Df Sum Sq Mean Sq F value
behavior   3 43.645 14.5482  47.179
Residuals 84 25.903  0.3084
```

Mean Square (`Mean Sq`, or MS) = Sum of Squares / df

$$MS(\text{behavior}) = \frac{SS(\text{behavior})}{df(\text{behavior})} = \frac{43.645}{3} \approx 14.55$$

- MS(Residuals) estimates the **residual variance**, the square of the residual standard deviation (residual standard error in earlier work).
- The ratio of MS values is the ANOVA **F value**.

$$\text{ANOVA } F = \frac{MS(\text{behavior})}{MS(\text{Residuals})} = \frac{14.5482}{0.3084} \approx 47.18$$

So, what's in the ANOVA table? (p value)

```
tidy(anova(model_one))
```

```
# A tibble: 2 x 6
  term          df sumsq meansq statistic   p.value
  <chr>      <int> <dbl>  <dbl>     <dbl>     <dbl>
1 behavior       3  43.6 14.5        47.2  5.68e-18
2 Residuals     84  25.9  0.308      NA   NA
```

- The *p* value is derived from the ANOVA F statistic, as compared to the F distribution.
- Which F distribution is specified by the two degrees of freedom values, as the F table is indexed by both a numerator and a denominator df.

```
pf(47.17879, df1 = 3, df2 = 84, lower.tail = FALSE)
```

```
[1] 5.680062e-18
```

We could also have used. . .

Are there statistically significant differences in mean outcome across the behavior group means?

$$H_0 : \mu_{Best} = \mu_{High} = \mu_{Low} = \mu_{Worst} \text{ vs. } H_A : \text{At least one } \mu \text{ is different.}$$

```r
summary(aov(model_one))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)    
behavior     3  43.64  14.548   47.18 <2e-16 ***
Residuals   84  25.90   0.308                   
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, what's the conclusion? Is this a surprise?

Another identical approach

Are there statistically significant differences in mean outcome across the behavior group means?

$$H_0 : \mu_{Best} = \mu_{High} = \mu_{Low} = \mu_{Worst} \text{ vs. } H_A : \text{At least one } \mu \text{ is different.}$$

```r
oneway.test(outcomes ~ behavior, data = ohio18,
            var.equal = TRUE)
```

```
    One-way analysis of means

data:  outcomes and behavior
F = 47.179, num df = 3, denom df = 84, p-value <
2.2e-16
```

ANOVA Assumptions

The assumptions behind analysis of variance are the same as those behind a linear model. Of specific interest are:

- The samples obtained from each group are independent.
- Ideally, the samples from each group are a random sample from the population described by that group.
- In the population, the variance of the outcome in each group is equal. (This is less of an issue if our study involves a balanced design.)
- In the population, we have Normal distributions of the outcome in each group.

Happily, the ANOVA F test is fairly robust to violations of the Normality assumption.

Is there an approach that doesn't assume equal variances?

Yes, but this isn't exciting if we have a balanced design.

```
oneway.test(outcomes ~ behavior, data = ohio18)
```

```

    One-way analysis of means (not assuming equal
    variances)

data:  outcomes and behavior
F = 47.322, num df = 3.000, denom df = 46.314,
p-value = 3.788e-14
```

- Note that this approach uses a fractional degrees of freedom calculation in the denominator.

The Kruskal-Wallis Test

If you thought the data were severely skewed, you might avoid the ANOVA and instead try:

```
kruskal.test(outcomes ~ behavior, data = ohio18)
```

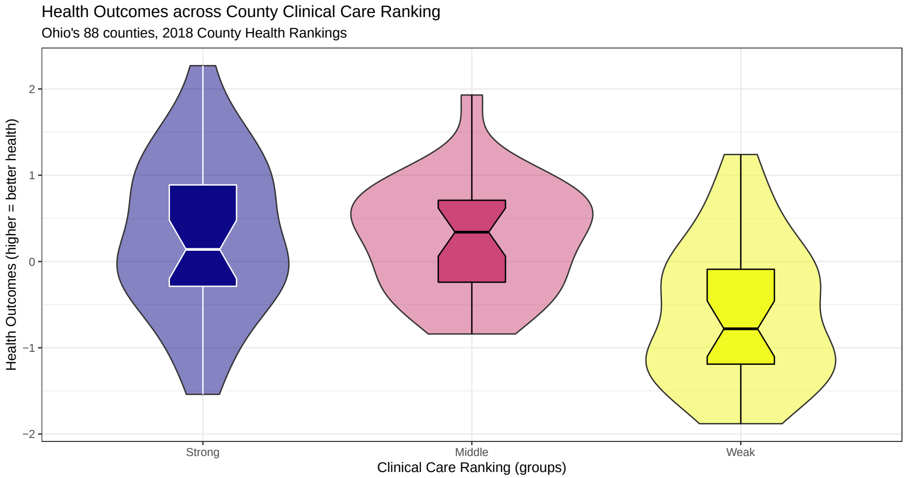
```

    Kruskal-Wallis rank sum test

data:  outcomes by behavior
Kruskal-Wallis chi-squared = 57.049, df = 3,
p-value = 2.508e-12
```

- H_0: The four `behavior` groups have the same center to their `outcomes` distributions.
- H_A: At least one group has a shifted distribution, with a different center to its `outcomes`.

What would be the conclusion in this case?

Question 2

Do groups of counties defined by clinical care show meaningful differences in average health outcomes?



Source: http://www.countyhealthrankings.org/app/ohio/2018/downloads

Question 2 Numerical Summaries

Do groups of counties defined by clinical care show meaningful differences in average health outcomes?

```
mosaic::favstats(outcomes ~ clin_care, data = ohio18) %>%
  knitr::kable(digits = 2)
```

clin_care	min	Q1	median	Q3	max	mean	sd	n	missin
Strong	-1.54	-0.29	0.14	0.89	2.27	0.30	0.91	30	
Middle	-0.84	-0.24	0.34	0.71	1.93	0.28	0.65	29	
Weak	-1.88	-1.19	-0.78	-0.09	1.24	-0.59	0.82	29	

Trust me - there's no missing data here. Sorry the table cuts off.

Question 2 Analysis of Variance

```r
model2 <- lm(outcomes ~ clin_care, data = ohio18)

anova(model2)
```

```
Analysis of Variance Table

Response: outcomes
          Df Sum Sq Mean Sq F value    Pr(>F)
clin_care  2 15.221  7.6103  11.907 2.762e-05 ***
Residuals 85 54.327  0.6391
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 2 Kruskal-Wallis test

```
kruskal.test(outcomes ~ clin_care, data = ohio18)
```

```

	Kruskal-Wallis rank sum test

data:  outcomes by clin_care
Kruskal-Wallis chi-squared = 18.54, df = 2,
p-value = 9.422e-05
```

K-Sample Study Design, Comparing Means

1. What is the outcome under study?
2. What are the (in this case, K $>$ 2) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the samples to the population(s)?
5. What is the significance level (or, the confidence level) we require?
6. Are we doing one-sided or two-sided testing? (usually 2-sided)
7. What does the distribution of each individual sample tell us about which inferential procedure to use?
8. Are there statistically meaningful differences between population means?
9. If an overall test is significant, can we identify pairwise comparisons of means that show significant differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

What's Left to do? (Multiple Comparisons)

9. If an overall test is significant, can we identify pairwise comparisons of means that show significant differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

Yes. There are two methods we'll study to identify specific pairs of means where we have statistically significant differences, while dealing with the problem of multiple comparisons.

- Bonferroni pairwise comparisons
- Tukey's HSD (Honestly Significant Differences) approach

We found a significant difference between `behavior` groups

But which ones are different from which? All the ANOVA tells is that there is strong evidence that they aren't all the same.

```r
anova(lm(outcomes ~ behavior, data = ohio18))
```

```
Analysis of Variance Table

Response: outcomes
          Df Sum Sq Mean Sq F value    Pr(>F)
behavior   3 43.645 14.5482  47.179 < 2.2e-16 ***
Residuals 84 25.903  0.3084
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is, for example, Best significantly different from Worst?

Could we just run a bunch of t tests?

This approach assumes that you need to make no adjustment for the fact that you are doing multiple comparisons, simultaneously.

```r
pairwise.t.test(ohio18$outcomes, ohio18$behavior,
                p.adjust.method = "none")
```

```

    Pairwise comparisons using t tests with pooled SD 

data:  ohio18$outcomes and ohio18$behavior 

      Best    High    Low    
High  4.7e-05 -       -      
Low   7.1e-11 0.00212 -      
Worst < 2e-16 2.7e-10 0.00013

P value adjustment method: none 
```

The problem of Multiple Comparisons

- The more comparisons you do simultaneously, the more likely you are to make an error.

In the worst case scenario, suppose you do two tests - first A vs. B and then A vs. C, each at the $\alpha = 0.10$ level.

- What is the combined error rate across those two t tests?

The problem of Multiple Comparisons

In the worst case scenario, suppose you do two tests - first A vs. B and then A vs. C, each at the $\alpha = 0.10$ level.

- What is the combined error rate across those two t tests?

Run the first test. Make a Type I error 10% of the time.

A vs B Type I error	Probability
Yes	0.1
No	0.9

Now, run the second test. Assume (perhaps wrongly) that comparing A to C is independent of your A-B test result. What is the error rate now?

The problem of Multiple Comparisons

In the worst case scenario, suppose you do two tests - first A vs. B and then A vs. C, each at the $\alpha = 0.10$ level.

- What is the combined error rate across those two t tests?

Assuming there is a 10% chance of making an error in either test, independently . . .

–	Error in A vs. C	No Error	Total
Type I error in A vs. B	0.01	0.09	0.10
No Type I error in A-B	0.09	0.81	0.90
Total	0.10	0.90	1.00

So you will make an error in the A-B or A-C comparison **19%** of the time, rather than the nominal $\alpha = 0.10$ error rate.

But in our case, we're building SIX tests

1. Best vs. High
2. Best vs. Low
3. Best vs. Worst
4. High vs. Low
5. High vs. Worst
6. Low vs. Worst

and if they were independent, and each done at a 5% error rate, we could still wind up with an error rate of

$$.05 + (.95)(.05) + (.95)(.95)(.05) + (.95)^3(.05) + (.95)^4(.05) + (.95)^5(.05) = .265$$

Or worse, if they're not independent.

The Bonferroni Method

If we do 6 tests, we could just reduce the necessary α to 0.05 / 6 = 0.0083 and that would maintain an error rate no higher than $\alpha = 0.05$ across those tests.

- Or we could let R adjust the *p* values directly…

```
pairwise.t.test(ohio18$outcomes, ohio18$behavior,
                p.adjust.method = "bonferroni")
```

```

    Pairwise comparisons using t tests with pooled SD 

data:  ohio18$outcomes and ohio18$behavior 

      Best    High    Low    
High  0.00028 -       -      
Low   4.3e-10 0.01273 -      
Worst < 2e-16 1.6e-09 0.00081
```

Tukey Honestly Significant Differences (HSD)

Tukey's HSD approach is a better choice for pre-planned comparisons with a balanced (or nearly balanced) design. It provides confidence intervals and an adjusted *p* value for each comparison.

- Let's run some confidence intervals to yield an overall 99% confidence level, even with 6 tests...

```
TukeyHSD(aov(lm(outcomes ~ behavior, data = ohio18)),
         conf.level = 0.99, ordered = TRUE)
```

Output on the next slide...

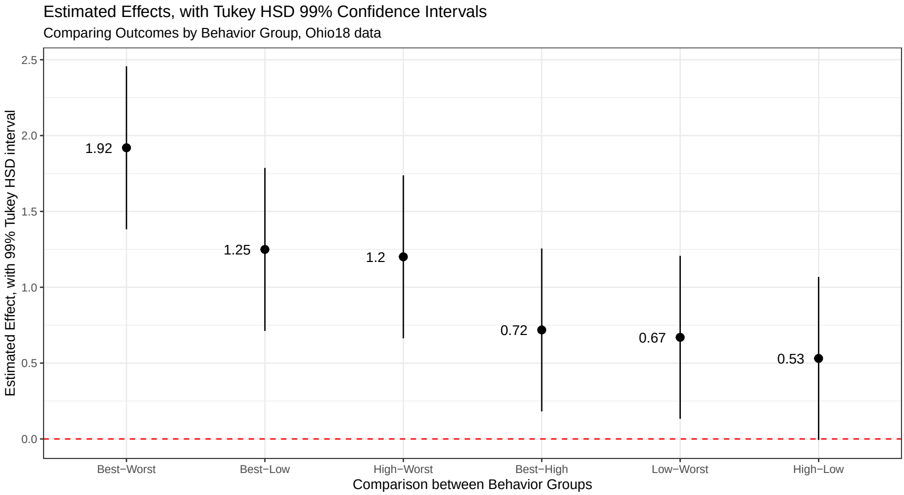
```
  Tukey multiple comparisons of means
    99% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = lm(outcomes ~ behavior, data = ohio18))

$behavior
                diff          lwr      upr     p adj
Low-Worst  0.6700000  0.132693736 1.207306 0.0007665
High-Worst 1.2009091  0.663602827 1.738215 0.0000000
Best-Worst 1.9195455  1.382239190 2.456852 0.0000000
High-Low   0.5309091 -0.006397173 1.068215 0.0111954
Best-Low   1.2495455  0.712239190 1.786852 0.0000000
Best-High  0.7186364  0.181330099 1.255943 0.0002716
```

Tidying the Tukey HSD confidence intervals

```
model_one <- lm(outcomes ~ behavior, data = ohio18)
tukey_one <- tidy(TukeyHSD(aov(model_one),
                           ordered = TRUE,
                           conf.level = 0.99))
knitr::kable(tukey_one, digits = 3)
```

term	comparison	estimate	conf.low	conf.high	adj.p.value
behavior	Low-Worst	0.670	0.133	1.207	0.001
behavior	High-Worst	1.201	0.664	1.738	0.000
behavior	Best-Worst	1.920	1.382	2.457	0.000
behavior	High-Low	0.531	-0.006	1.068	0.011
behavior	Best-Low	1.250	0.712	1.787	0.000
behavior	Best-High	0.719	0.181	1.256	0.000

Plotting Your Tukey HSD intervals, Approach 1



Code for Plot on Previous Slide

```
ggplot(tukey_one, aes(x = reorder(comparison, -estimate),
                      y = estimate)) +
  geom_pointrange(aes(ymin = conf.low, ymax = conf.high)) +
  geom_hline(yintercept = 0, col = "red",
             linetype = "dashed") +
  geom_text(aes(label = round(estimate,2)), nudge_x = -0.2) +
  theme_bw() +
  labs(x = "Comparison between Behavior Groups",
       y = "Estimated Effect, with 99% Tukey HSD interval",
       title = "Estimated Effects, with Tukey HSD 99% Confiden
       subtitle = "Comparing Outcomes by Behavior Group, Ohio1
```

Question 2: 90% Tukey HSD intervals, tidying

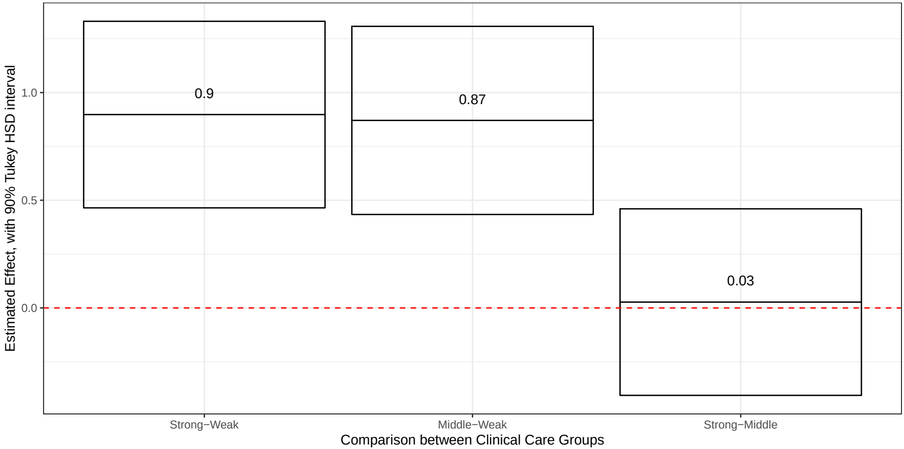
```
model_two <- lm(outcomes ~ clin_care, data = ohio18)
tukey_two <- tidy(TukeyHSD(aov(model_two),
                           ordered = TRUE,
                           conf.level = 0.90))
knitr::kable(tukey_two, digits = 3)
```

term	comparison	estimate	conf.low	conf.high	adj.p.value
clin_care	Middle-Weak	0.871	0.434	1.307	0.000
clin_care	Strong-Weak	0.898	0.465	1.331	0.000
clin_care	Strong-Middle	0.027	-0.406	0.460	0.991

Plotting Question 2 Tukey HSD intervals

Estimated Effects, with Tukey HSD 90% Confidence Intervals

Comparing Outcomes by Clinical Care Group, Ohio18 data



Code for Question 2 Tukey HSD plot

```
ggplot(tukey_two, aes(x = reorder(comparison, -estimate),
                      y = estimate)) +
  geom_crossbar(aes(ymin = conf.low, ymax = conf.high),
                fatten = 1) +
  geom_hline(yintercept = 0, col = "red",
             linetype = "dashed") +
  geom_text(aes(label = round(estimate,2)), nudge_y = 0.1) +
  theme_bw() +
  labs(x = "Comparison between Clinical Care Groups",
       y = "Estimated Effect, with 90% Tukey HSD interval",
       title = "Estimated Effects, with Tukey HSD 90% Confiden
       subtitle = "Comparing Outcomes by Clinical Care Group,
```

Coming Soon

- Power and Sample Size Ideas
- Working with Larger Contingency Tables (Chi-Square Tests of Independence)
- Mantel-Haenszel Procedures for Three-Way Tables