431 Class 17

thomaselove.github.io/431

2021-10-21

Today's R Setup

```
knitr::opts_chunk$set(comment=NA) # as always
options(width = 55) # to fit things on the slides

library(readxl) # to read in an .xlsx file
library(ggrepel) # to help label residual plots
library(patchwork)
library(broom)
library(pwr)
library(janitor)
library(knitr)
library(magrittr)
library(tidyverse)

source("data/Love-boost.R")

theme_set(theme_bw())
```

Today's Agenda

- Power Calculations for Comparing Two Proportions
  - With `power.prop.test` for balanced designs
  - With the `pwr` package for unbalanced designs
- A Few Thoughts on Project A
- The Analysis of Variance
  - Using Regression to Develop an ANOVA model
  - Methods for pairwise multiple comparisons
  - Three Examples Using the `ohio_20` data

Note: The material introduced today will appear in Quiz 3, rather than Quiz 2.

Comparing Two Proportions

Tuberculosis Prevalence Among IV Drug Users

Suppose now that we are investigating factors affecting tuberculosis prevalence among intravenous drug users.

We collect the following information:

- Among 97 individuals who admit to sharing needles,
  - 24 (24.7%) had a positive tuberculin skin test result.
- Among 161 drug users who deny sharing needles,
  - 28 (17.4%) had a positive test result.

What does the 2x2 table look like?

Tuberculosis Prevalence Among IV Drug Users

Among 97 individuals who admit to sharing needles, 24 (24.7%) had a positive tuberculin skin test result; among 161 drug users who deny sharing needles, 28 (17.4%) had a positive test result.

The 2x2 Table is. . .

	TB+	TB-
share	24	73
don't	28	133

- rows describe needle sharing, columns describe TB test result
- row 1 people who share needles: 24 TB+, and 97-24 = 73 TB-
- row 2 people who don't share: 28 TB+ and 161-28 = 133 TB-

`twobytwo` (with Bayesian Augmentation)

To start, we'll test the null hypothesis that the population proportions of intravenous drug users who have a positive tuberculin skin test result are identical for those who share needles and those who do not.

$$H_0 : \pi_{share} = \pi_{donotshare}$$

$$H_A : \pi_{share} \neq \pi_{donotshare}$$

We'll use the Bayesian augmentation.

```
twobytwo(24+2, 73+2, 28+2, 133+2,
         "Sharing", "Not Sharing",
         "TB test+", "TB test-")
```

Two-by-Two Table Result

```
Outcome    : TB test+
Comparing : Sharing vs. Not Sharing

             TB test+ TB test- P(TB test+) 95% conf. int.
Sharing            26       75     0.2574  0.1816 0.3513
Not Sharing        30      135     0.1818  0.1301 0.2482

                                        95% conf. interval
                Relative Risk: 1.4158    0.8910   2.2498
            Sample Odds Ratio: 1.5600    0.8594   2.8318
Conditional MLE Odds Ratio: 1.5572    0.8189   2.9511
    Probability difference: 0.0756   -0.0244   0.1819

Exact P-value: 0.1638      Asymptotic P-value: 0.1438
```

What conclusions should we draw?

Designing a New TB Study

PI:

- OK. That's a nice pilot.
- We saw $p_{nonshare} = 0.18$ and $p_{share} = 0.26$ after your augmentation.
- Help me design a new study.
  - This time, let's have as many needle-sharers as non-sharers.
  - We should have 90% power to detect a difference almost as large as what we saw in the pilot, or larger, so a difference of 6 percentage points.
  - We'll use a two-sided test, and $\alpha = 0.05$, of course.

What sample size would be required to accomplish these aims?

How `power.prop.test` works

`power.prop.test` works much like the `power.t.test` we saw for means.

Again, we specify 4 of the following 5 elements of the comparison, and R calculates the fifth.

- The sample size (interpreted as the # in each group, so half the total sample size)
- The true probability in group 1
- The true probability in group 2
- The significance level (α)
- The power (1 - β)

The big weakness with the `power.prop.test` tool is that it doesn't allow you to work with unbalanced designs.

Using `power.prop.test` for Balanced Designs

To find the sample size for a two-sample comparison of proportions using a balanced design:

- we will use a two-sided test, with $\alpha = .05$, and power $= .90$,
- we estimate that non-sharers have probability .18 of positive tests,
- and we will try to detect a difference between this group and the needle sharers, who we estimate will have a probability of .24

Finding the required sample size in R

```
power.prop.test(p1 = .18, p2  = .24,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

Any guess as to needed sample size?

Results: power.prop.test for Balanced Design

```
power.prop.test(p1 = .18, p2  = .24,
                alternative = "two.sided",
                sig.level = 0.05, power = 0.90)
```

```
Two-sample comparison of proportions power calculation
n = 966.3554
p1 = 0.18, p2 = 0.24
sig.level = 0.05, power = 0.9, alternative = two.sided
NOTE: n is number in *each* group
```

So, we'd need at least 967 non-sharing subjects, and 967 more who share needles to accomplish the aims of the study, or a total of 1934 subjects.

Another Scenario

Suppose we can get 400 sharing and 400 non-sharing subjects. How much power would we have to detect a difference in the proportion of positive skin test results between the two groups that was identical to the pilot data above or larger, using a *one-sided* test, with $\alpha = .10$?

```
power.prop.test(n=400, p1=.18, p2=.26, sig.level = 0.10,
                alternative="one.sided")
```

```
Two-sample comparison of proportions power calculation
n = 400, p1 = 0.18, p2 = 0.26
sig.level = 0.1, power = 0.9273602
alternative = one.sided
NOTE: n is number in *each* group
```

We would have just over 92.7% power to detect such an effect.

Using the `pwr` package to assess sample size for Unbalanced Designs

The `pwr.2p2n.test` function in the `pwr` package can help assess the power of a test to determine a particular effect size using an unbalanced design, where n_1 is not equal to n_2.

As before, we specify four of the following five elements of the comparison, and R calculates the fifth.

- `n1` = The sample size in group 1
- `n2` = The sample size in group 2
- `sig.level` = The significance level (α)
- `power` = The power (1 - β)
- `h` = the effect size h, which can be calculated separately in R based on the two proportions being compared: p_1 and p_2.

Calculating the Effect Size h

To calculate the effect size for a given set of proportions, use `ES.h(p1, p2)` which is available in the `pwr` package.

For instance, comparing .18 to .25, we have the following effect size.

```
ES.h(p1 = .18, p2 = .25)
```

```
[1] -0.1708995
```

Using `pwr.2p2n.test` in R

Suppose we can have 700 samples in group 1 (the not sharing group) but only 400 in group 2 (the group of users who share needles).

How much power would we have to detect the distinction between p1 = .18, p2 = .25 with a 5% significance level in a two-sided test?

R Command to find the resulting power

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 700, n2 = 400, sig.level = 0.05)
```

Results of using pwr.2p2n.test

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 700, n2 = 400, sig.level = 0.05)
```

```
difference of proportion power calculation
for binomial distribution (arcsine transformation)

h = 0.1708995, n1 = 700, n2 = 400
sig.level = 0.05, power = 0.7783562
alternative = two.sided
NOTE: different sample sizes
```

We will have just under 78% power under these circumstances.

Comparison to Balanced Design

How does this compare to the results with a balanced design using 1100 drug users in total, i.e. with 550 patients in each group?

```
pwr::pwr.2p2n.test(h = ES.h(p1 = .18, p2 = .25),
                   n1 = 550, n2 = 550, sig.level = 0.05)
```

which yields a power estimate of 0.809. Or we could instead have used...

```
power.prop.test(p1 = .18, p2 = .25, sig.level = 0.05,
                n = 550)
```

which yields an estimated power of 0.808.

Each approach uses approximations, and slightly different ones, so it's not surprising that the answers are similar, but not identical.

What haven't I included here?

1. Some people will drop out.
2. What am I going to do about missing data?
3. What if I want to do my comparison while adjusting for covariates?

How Big A Sample Size Do I need?

1. What is the budget?
2. What are you trying to compare?
3. What is the study design?
4. How big an effect size do you expect (hope) to see?
5. What was that budget again?
6. OK, tell me the maximum allowable rates of Type I and Type II error that you want to control for. Or, if you like, tell me the confidence level and power you want to have.
7. And what sort of statistical inference do you want to plan for?

On Project A

Deliverables on 2021-11-01 at 9 PM

- To Canvas: R Markdown and HTML report (please don't use PDF)
- To Canvas: Video (has outsized importance)
- Google Form (self-evaluation) submitted after the Canvas stuff is in

Working with a Partner?

- The submitting investigator submits Rmd, HTML and video to Canvas, and then submits the Google Form.
- The partner submits the one-sentence text document to Canvas (yes, again), and then submits the Google Form.

On the R Markdown and HTML Report

Please review the 40-item checklist for the final report, listing things the TAs will be looking for in evaluating your project. Details matter.

- Be 100% certain that your sections are numbered automatically using number_sections: TRUE in your YAML.
- The most important part of your analyses are the research questions, and your paragraphs about conclusions and limitations. This is the only part Dr. Love will read before he reviews your video, although he'll come back and read other things after grading the videos.
    - We are NOT looking for separate training and testing samples. We want you to use your whole sample for all elements of this Project.

Dealing with Missing Data, I

You can absolutely use complete cases in each of the three analyses, but these should have explicitly specified and different sample sizes if you have missing data in something other than your outcome.

- Be certain to specify what you are assuming (MCAR, MAR or MNAR) about the missing data mechanism in developing your models. (Refer to Chapter 8 of the Course Notes.)

Dealing with Missing Data, II

If you decide instead to use single imputation (with the `simputation` package), as described in Chapter 8 of the Course Notes, great.

- Obviously, this would mean that you are making a different assumption about the missing data mechanism, and this should be explicitly specified.
- It's fine to use `rlm` or `pmm` to impute quantitative predictors, and to use `pmm` or `cart` for categorical ones.
- Use (at least) all other variables included in your planned model to help with the imputation, and you are also welcome to use other variables from your tidy data set, if you like. Be sure to state explicitly what you are doing.
- Be sure to set a seed just before you do any imputation.
- For Project A, you can impute predictors, but not the outcome. Use complete cases for your outcome, regardless of any other decisions you make.

Your Job in the R Markdown / HTML report

Your job is to provide reasonable research questions (one question per analysis) and reasoned conclusions and clear, compelling logical motivations for those conclulsions.

- All of the material from your proposal should be included in your final report, of course. Some of that material needs to be augmented, and, of course, may have changed in light of what you've done since your proposal was improved.
- Don't fight our example. Use it to make sure you've covered everything we need to see, and so that you make it as easy for us to review as possible.

The Analyses

Create a section called Analysis 1, another called Analysis 2, and another called Analysis 3, **using the outline (including subsections) we've provided** for all of your headings.

1. Analysis 1 - predict your outcome using one of your quantitative predictors
2. Analysis 2 - predict your outcome using one of your categorical predictors
3. Analysis 3 - predict your outcome using one of your quantitative predictors (can be same as Analysis 1, or different) as well as `state`.

Many of you will only use four variables: your outcome, one quantitative predictor, one categorical one, and the `state`, in these analyses. Some will use five.

Don't use language to suggest causality if your data and model cannot justify that.

Analysis 2 with a binary predictor

If you're using a binary predictor here, then this regression model boils down to a pooled t test. You should complete that regression model (with whatever transformation you like) and obtain an appropriate regression-based analysis.

- If you are unsatisfied with the adherence of the situation to regression assumptions, and want to ALSO develop an alternative confidence interval based on a Welch's t procedure, or a bootstrap comparison, that would be a good idea.
- Be sure to justify your final choice of approach (pooled t or other) with results from the output, and a description of what you're doing. Draw clear conclusions.
- Be certain that you address the issue of what population is being described by the confidence interval you develop in this setting.

Analysis 2 with a multi-categorical predictor

If you're using a multi-categorical predictor here, then this regression model boils down to an analysis of variance (ANOVA).

- We will show a detailed ANOVA in our last example today. Use that example, and Chapter 25 in the Course Notes to provide indications about what we're looking for.
- A set of formal pairwise comparisons should be a part of your analysis should your multi-categorical predictor demonstrate meaningful predictive value for your outcome.

Analysis 3

Does `state` have a large impact on the model between your outcome and your quantitative predictor, and what does this imply about that outcome-quant predictor relationship?

- I suppose you could fit a model using `state` only to affect the intercept of the relationship between your outcome and your quantitative predictor, but . . .
- We'd far prefer that you fit a model where the `state` also can affect the slope of that relationship.
  - This implies that you should be fitting a model with an interaction term between `state` and your quantitative predictor, and that model requires careful interpretation.
  - A part of that interpretation should be a clear and appropriate visualization explaining what the impact of the `state` is on the regression line describing the outcome based on the quantitative predictor.

On the Video, I

The video has more weight on your project grade than you might think, despite the fact that it is short ($<$ 3 minutes if working alone, $<$ 5 minutes if working with a partner.)

- Do what we ask you to do in the video. (See the Final Report instructions.)
- Dr. Love will review all videos in detail **before** he looks at (most of) your report.

All videos should include a clear statement of the research questions for both analyses you present, and justify the responses to those questions with results from the analyses.

- The video must stand on its own, in the sense that it must be completely understandable to someone who has not read your report, but who is generally familiar with County Health Rankings and its measurements.

On the Video, II

You need to tell us everything we need to know to evaluate your claims, and no more.

- Don't use causal or sloppy language in your video unless you can back that up.
- Make sure we can clearly see everything you want us to see in the video.
- Building the video is going to take more than the time to record it. Leave that time in your planning. It's a very bad idea to try to toss this together in the last 30 minutes.
- Make smart choices about what to present in the video. You cannot possibly include everything you put in your main report. What are the conclusion-driving things to show us?

The Self-Evaluation

The Google Form for the Project A Self-Evaluation is available now at https://bit.ly/431-2021-projA-self-evaluation. Fill it out AFTER you have submitted the other materials to Canvas. This way, you'll have completed that work before you submit the self-evaluation, which is what we want you to do.

- Again, if you're working with a partner for the rest of the project, you still do this part completely on your own.

The Google Form has the same deadline as everything else: November 1 at 9 PM.

Analysis of Variance for Comparing Multiple Means

Today's ANOVA Data (`ohio_2020.xlsx`)

`ohio_2020.xlsx` rows describe 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, because we've taken the negative of the Z score CHR provides)
- health behavior ranking (1-88, we'll divide into 4 groups)
- clinical care ranking (1-88, we'll split into 3 groups)
- proportion of county residents who live in rural areas
- median income, in dollars
- proportion of votes in the 2016 Presidential Election for Donald Trump

Sources (these bullets are links)

- County Health Rankings (2020 Ohio Data)
- Wikipedia for 2016 Election Results

Importing the Data / Creating some Factors

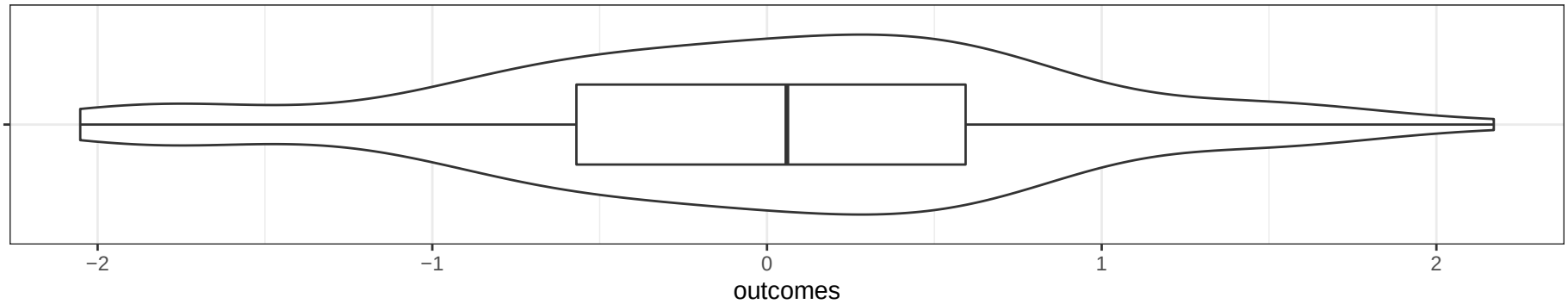
```
ohio20 <- read_xlsx("data/ohio_2020.xlsx") %>%
  mutate(behavior = Hmisc::cut2(rk_behavior, g = 4),
         clin_care = Hmisc::cut2(rk_clin_care, g = 3)) %>%
  mutate(behavior = fct_recode(behavior,
           "Best" = "[ 1,23)", "High" = "[23,45)",
           "Low" = "[45,67)", "Worst" = "[67,88]")) %>%
  mutate(clin_care = fct_recode(clin_care,
           "Strong" = "[ 1,31)", "Middle" = "[31,60)",
           "Weak" = "[60,88]")) %>%
  select(FIPS, state, county, outcomes, behavior, clin_care,
         everything())
```

A Quick Look at the Data

```
ohio20 %>% 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    -0.807 Worst    Strong
```

```
ggplot(ohio20, aes(x = "", y = outcomes)) + geom_violin() +
  geom_boxplot(width = 0.4) + coord_flip() + labs(x = "")
```



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.

Analytic Questions for Today's ANOVA

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

Question 1

Do average health outcomes differ by health behavior?

Question 1 Numerical Summaries

How do average health outcomes vary across groups of counties defined by health behavior?

```
mosaic::favstats(outcomes ~ behavior, data = ohio20) %>%
  rename(na = missing) %>% knitr::kable(digits = 2)
```

behavior	min	Q1	median	Q3	max	mean	sd	n	na
Best	-0.33	0.60	0.86	1.46	2.17	0.96	0.57	22	0
High	-0.35	0.00	0.30	0.55	0.77	0.25	0.35	22	0
Low	-1.15	-0.52	-0.09	0.16	0.73	-0.18	0.47	22	0
Worst	-2.05	-1.75	-0.87	-0.59	-0.08	-1.04	0.63	22	0

Note that there is no missing data here.

Analysis of Variance (ANOVA) testing: Question 1

Does the mean `outcomes` result differ detectably 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 detectable impact on the model's predictions of `outcomes`.

Building the Linear Model: Question 1

Can we detect differences in the population means of `outcomes` across the four `behavior` groups, using a 10% significance level?

```
model_one <- lm(outcomes ~ behavior, data = ohio20)
tidy(model_one, conf.int = 0.90) %>%
  select(term, estimate, std.error,
         conf.low, conf.high, p.value) %>% kable(dig = 2)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	0.96	0.11	0.75	1.18	0
behaviorHigh	-0.71	0.16	-1.02	-0.40	0
behaviorLow	-1.14	0.16	-1.45	-0.83	0
behaviorWorst	-2.01	0.16	-2.32	-1.70	0

How do we interpret this result?

Interpreting the Indicator Variables

The regression model (`model_one`) equation is

```
outcomes = 0.96 - 0.71 behaviorHigh
                - 1.14 behaviorLow
                  - 2.01 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 = 0.96 - 0.71 behaviorHigh
                - 1.14 behaviorLow
                  - 2.01 behaviorWorst
```

What predictions does the model make?

group	High	Low	Worst	Prediction
Best	0	0	0	0.96
High	1	0	0	0.96 - 0.71 = 0.25
Low	0	1	0	0.96 - 1.14 = -0.18
Worst	0	0	1	0.96 - 2.01 = -1.05

Do these predictions make sense?

Interpreting the Indicator Variables

The regression model (`model_one`) equation is

```
outcomes = 0.96 - 0.71 behaviorHigh
                - 1.14 behaviorLow
                  - 2.01 behaviorWorst
```

Sample means are. . .

```
ohio20 %>% group_by(behavior) %>%
  summarize(n = n(),
            mean = round_half_up(mean(outcomes),2)) %>%
  kable()
```

behavior	n	mean
Best	22	0.96
High	22	0.25
Low	22	-0.18
Worst	22	-1.04

ANOVA for the Linear Model: Question 1

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

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

```
anova(model_one)
```

```
Analysis of Variance Table

Response: outcomes
          Df Sum Sq Mean Sq F value    Pr(>F)
behavior   3 46.421 15.4736  57.718 < 2.2e-16 ***
Residuals 84 22.519  0.2681
---
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**.

```
          Df Sum Sq Mean Sq F value
behavior   3 46.421 15.4736  57.718
Residuals 84 22.519  0.2681
```

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)

```
          Df Sum Sq Mean Sq F value
behavior   3 46.421 15.4736  57.718
Residuals 84 22.519  0.2681
```

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)

$$\eta^2 = \frac{SS(\text{behavior})}{SS(\text{Total})} = \frac{46.421}{46.421 + 22.519} = \frac{46.421}{68.94} \approx 0.673$$

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

```
          Df Sum Sq Mean Sq F value
behavior   3 46.421 15.4736  57.718
Residuals 84 22.519  0.2681
```

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

$$MS(\text{behavior}) = \frac{SS(\text{behavior})}{df(\text{behavior})} = \frac{46.421}{3} \approx 15.4736$$

- 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**.

$$\text{ANOVA } F = \frac{MS(\text{behavior})}{MS(\text{Residuals})} = \frac{15.4736}{0.2681} \approx 57.718$$

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

```
tidy(anova(model_one)) %>% kable(dig = 3)
```

term	df	sumsq	meansq	statistic	p.value
behavior	3	46.421	15.474	57.718	0
Residuals	84	22.519	0.268	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(57.718, df1 = 3, df2 = 84, lower.tail = FALSE)
```

```
[1] 2.377323e-20
```

Alternative ways to show ANOVA results

```
glance(model_one) %>% select(r.squared, statistic, df, df.resi
```

```
# A tibble: 1 x 5
  r.squared statistic    df df.residual  p.value
      <dbl>     <dbl> <dbl>       <int>    <dbl>
1     0.673      57.7     3          84 2.38e-20
```

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

```
             Df Sum Sq Mean Sq F value Pr(>F)
behavior      3  46.42  15.474   57.72 <2e-16 ***
Residuals    84  22.52   0.268
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Multiple Comparisons

What's Left to do? (Multiple Comparisons)

9. If an overall test rejects the null, can we identify pairwise comparisons of means that show detectable 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 detectable differences, while dealing with the problem of multiple comparisons.

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

Compare behavior group means of outcomes?

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

```
anova(lm(outcomes ~ behavior, data = ohio20))
```

```
Analysis of Variance Table

Response: outcomes
          Df Sum Sq Mean Sq F value    Pr(>F)
behavior   3 46.421 15.4736  57.718 < 2.2e-16 ***
Residuals 84 22.519  0.2681
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Is, for example, Best detectably 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.

```
pairwise.t.test(ohio20$outcomes, ohio20$behavior,
                p.adjust.method = "none")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  ohio20$outcomes and ohio20$behavior

      Best    High    Low
High  1.8e-05 -       -
Low   1.4e-10 0.007   -
Worst < 2e-16 1.6e-12 3.6e-07

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 reduce the necessary α to 0.05 / 6 = 0.0083 and that maintains an error rate no higher than $\alpha = 0.05$ across the 6 tests.

- Or, R can adjust the *p* values directly. . .

```
pairwise.t.test(ohio20$outcomes, ohio20$behavior,
                p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  ohio20$outcomes and ohio20$behavior

      Best    High    Low
High  0.00011 -       -
Low   8.3e-10 0.04224 -
Worst < 2e-16 9.4e-12 2.1e-06

P value adjustment method: bonferroni
```

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 = ohio20)),
         conf.level = 0.99, ordered = TRUE)
```

Output on the next slide...

Tukey HSD Output

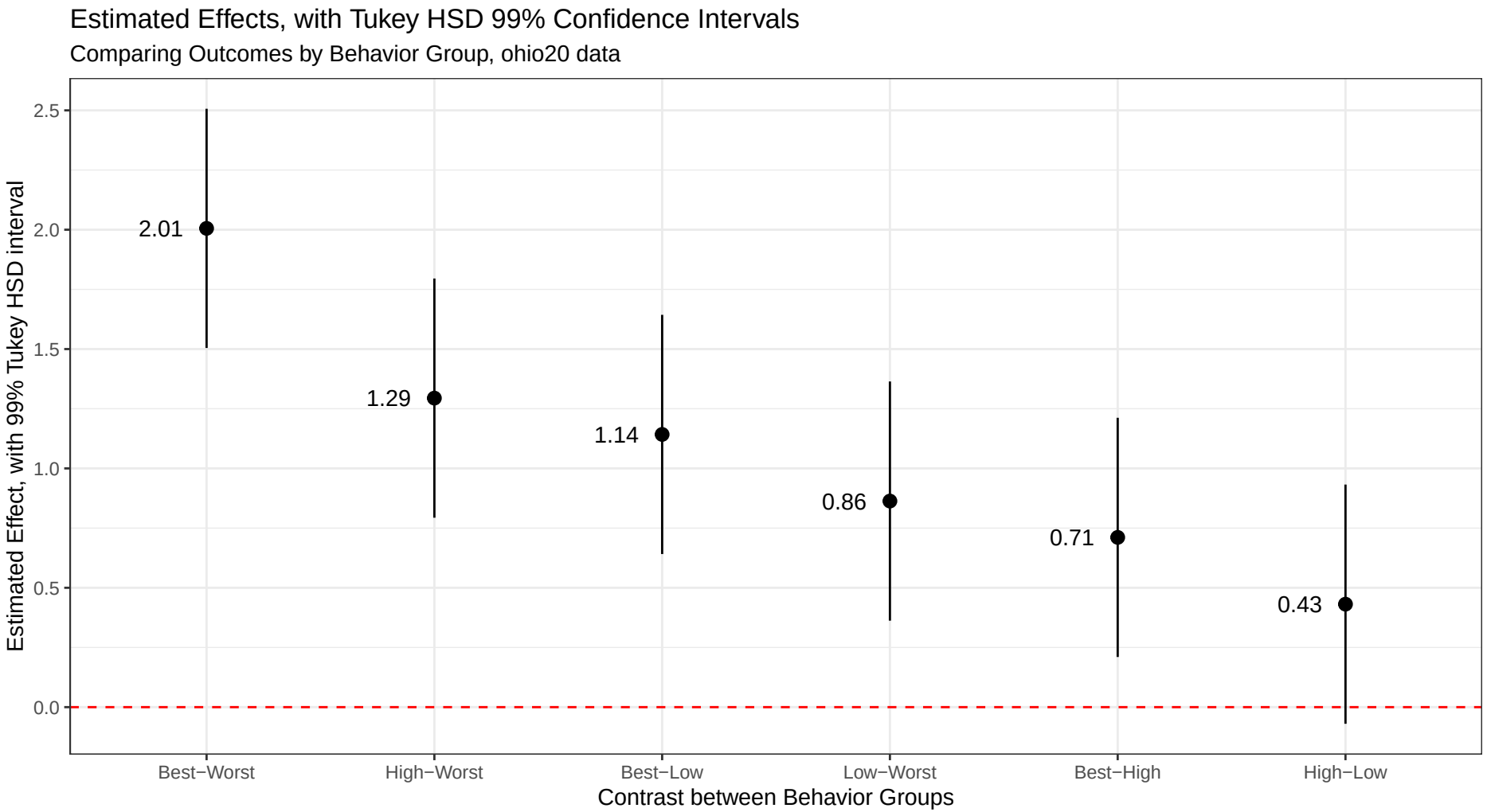
```
  Tukey multiple comparisons of means
    99% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = lm(outcomes ~ behavior, data = ohio20))

$behavior
                  diff          lwr       upr     p adj
Low-Worst   0.8632211   0.36223069 1.3642115 0.0000021
High-Worst  1.2945256   0.79353515 1.7955159 0.0000000
Best-Worst  2.0056105   1.50462011 2.5066009 0.0000000
High-Low    0.4313045 -0.06968593 0.9322949 0.0348350
Best-Low    1.1423894   0.64139903 1.6433798 0.0000000
Best-High   0.7110850   0.21009456 1.2120753 0.0001023
```

Tidying the Tukey HSD confidence intervals

```
model_one <- lm(outcomes ~ behavior, data = ohio20)
tukey_one <- tidy(TukeyHSD(aov(model_one),
                           ordered = TRUE,
                           conf.level = 0.99))
tukey_one %>% rename(null = null.value) %>% kable(dig = 3)
```

term	contrast	null	estimate	conf.low	conf.high	adj.p.value
behavior	Low-Worst	0	0.863	0.362	1.364	0.000
behavior	High-Worst	0	1.295	0.794	1.796	0.000
behavior	Best-Worst	0	2.006	1.505	2.507	0.000
behavior	High-Low	0	0.431	-0.070	0.932	0.035
behavior	Best-Low	0	1.142	0.641	1.643	0.000
behavior	Best-High	0	0.711	0.210	1.212	0.000

Plotting Your Tukey HSD intervals, Approach 1



Code for Plot on Previous Slide

```
ggplot(tukey_one, aes(x = reorder(contrast, -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) +
  labs(x = "Contrast between Behavior Groups",
       y = "Estimated Effect, with 99% Tukey HSD interval",
       title = "Estimated Effects, with Tukey HSD 99% Confider
       subtitle = "Comparing Outcomes by Behavior Group, ohio2
```

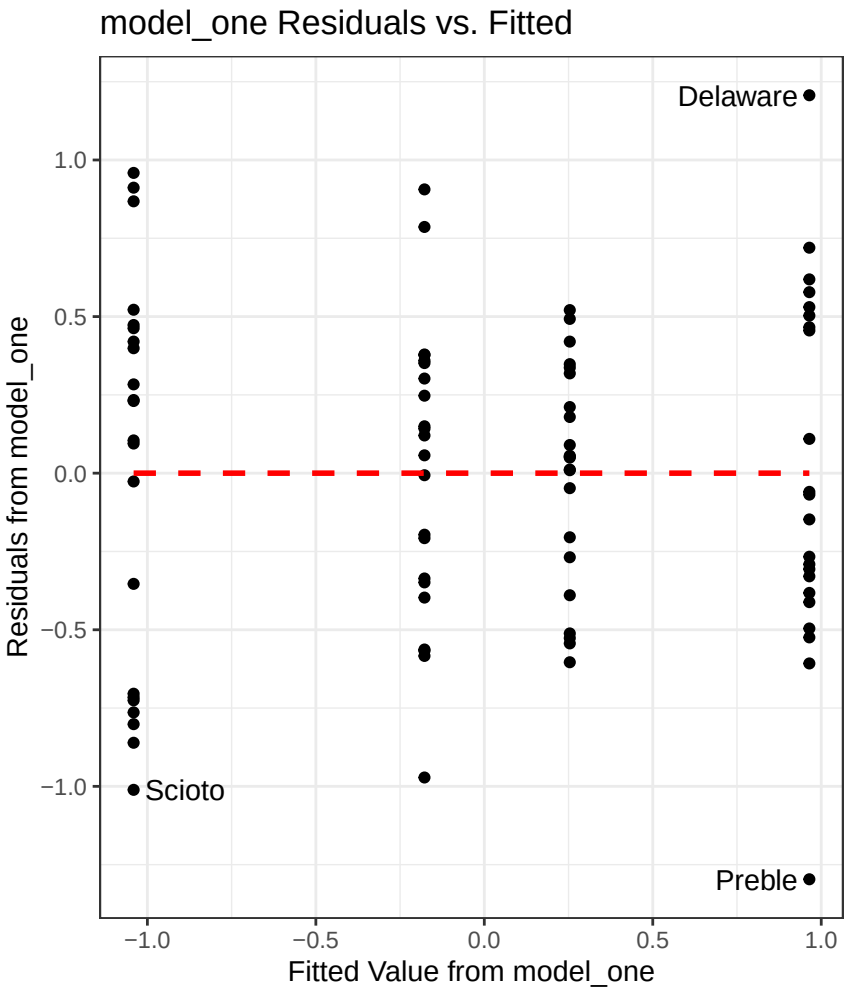
ANOVA Assumptions

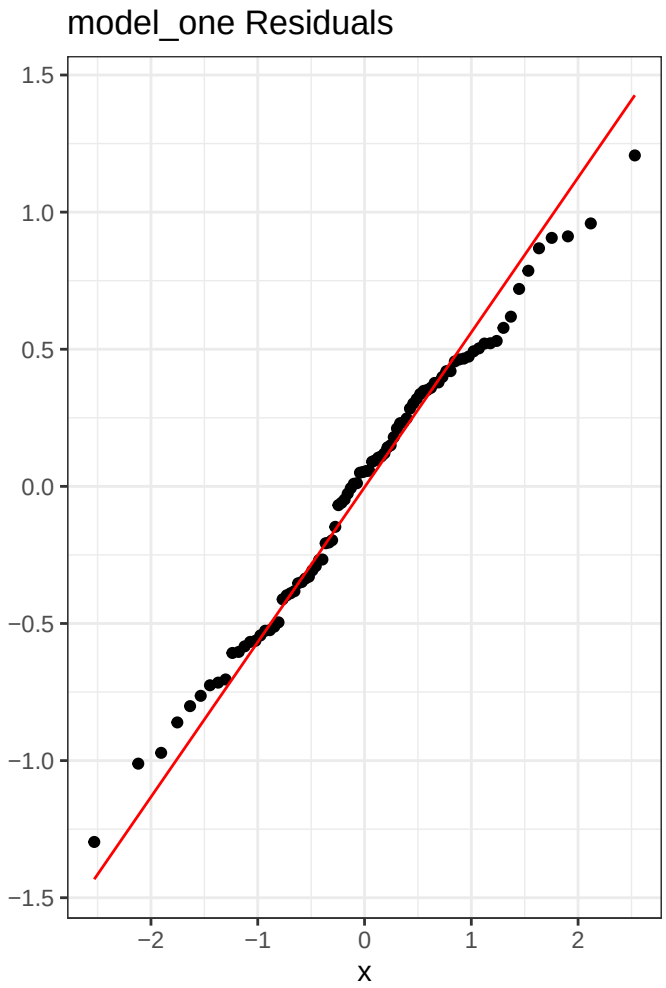
The assumptions behind analysis of variance are those of a linear model. Of specific interest are:

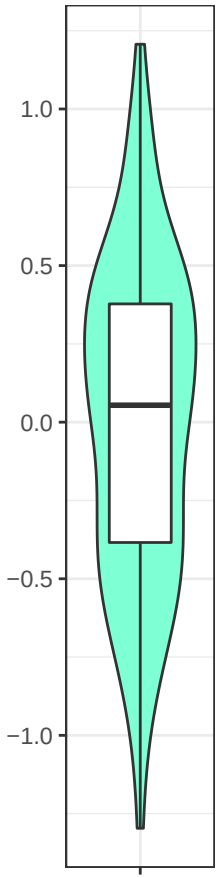
- The samples obtained from each group are independent.
- Ideally, the samples from each group are a random sample from the population described by that group.
- In the population, the variance of the outcome in each group is equal. (This is less of an issue if our study involves a balanced design.)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

Residual Plots for `model_one`







Can we avoid assuming equal population variances?

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

```
oneway.test(outcomes ~ behavior, data = ohio20)
```

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

data:  outcomes and behavior
F = 43.145, num df = 3.000, denom df = 45.494,
p-value = 2.349e-13
```

- 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 try:

```
kruskal.test(outcomes ~ behavior, data = ohio20)
```

```
    Kruskal-Wallis rank sum test

data:  outcomes by behavior
Kruskal-Wallis chi-squared = 61.596, df = 3,
p-value = 2.681e-13
```

- 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 here?

K-Sample Study Design, Comparing Means

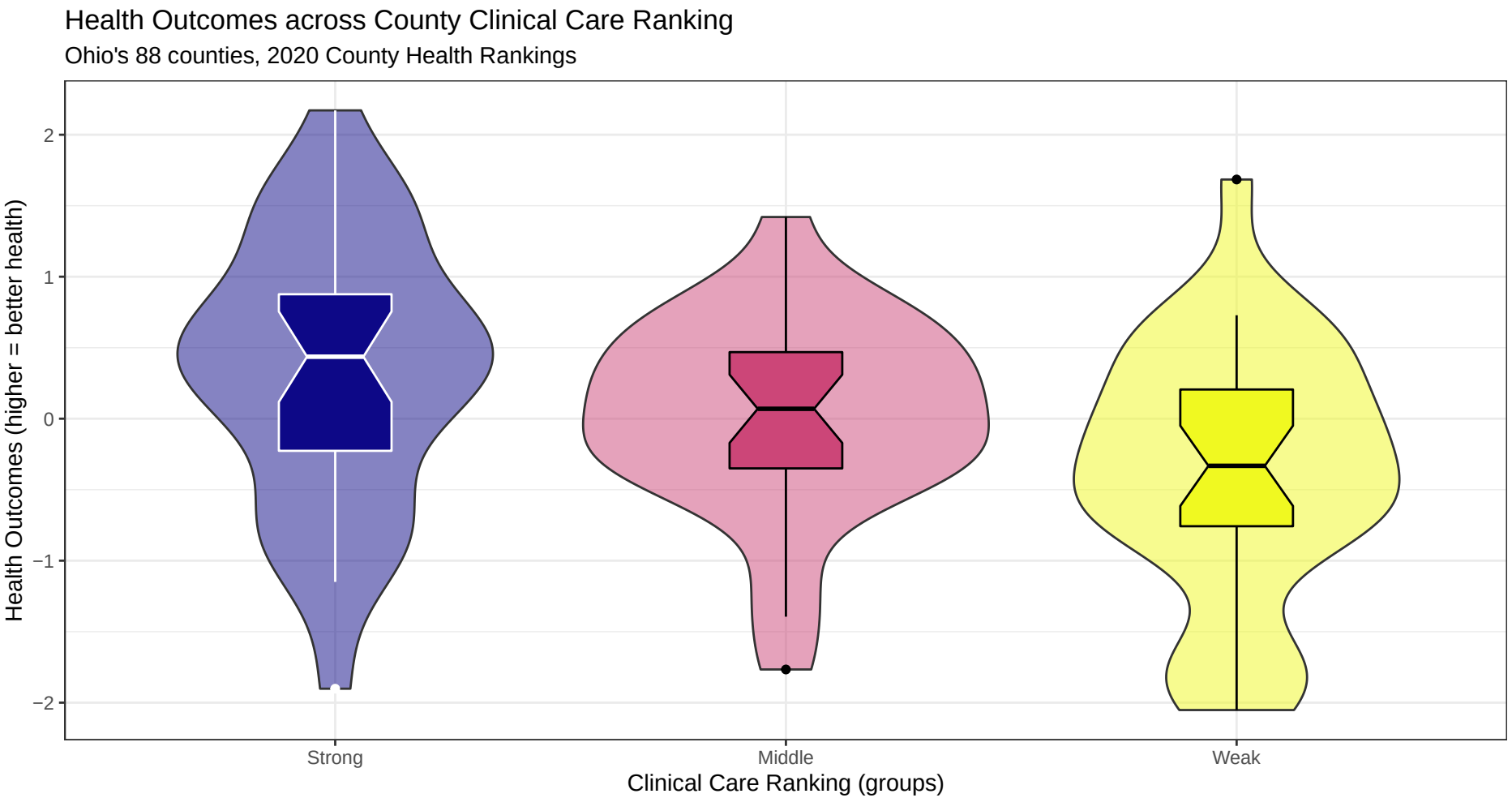
1. What is the outcome under study?
2. What are the (in this case, $K \geq 2$) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the 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 detectable differences between population means?
9. If an overall test rejects the null, can we identify pairwise comparisons of means that show detectable differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

This is the last of today's slides that will be discussed live. The remaining slides provide three more examples (using the same data) designed for self-study. Use these and the example in Chapter 25 of the Course Notes to guide your work.

For Self-Study: Health Outcomes compared across Clinical Care Groups

Question 2

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



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 = ohio20) %>%
  rename(na = missing) %>% knitr::kable(digits = 2)
```

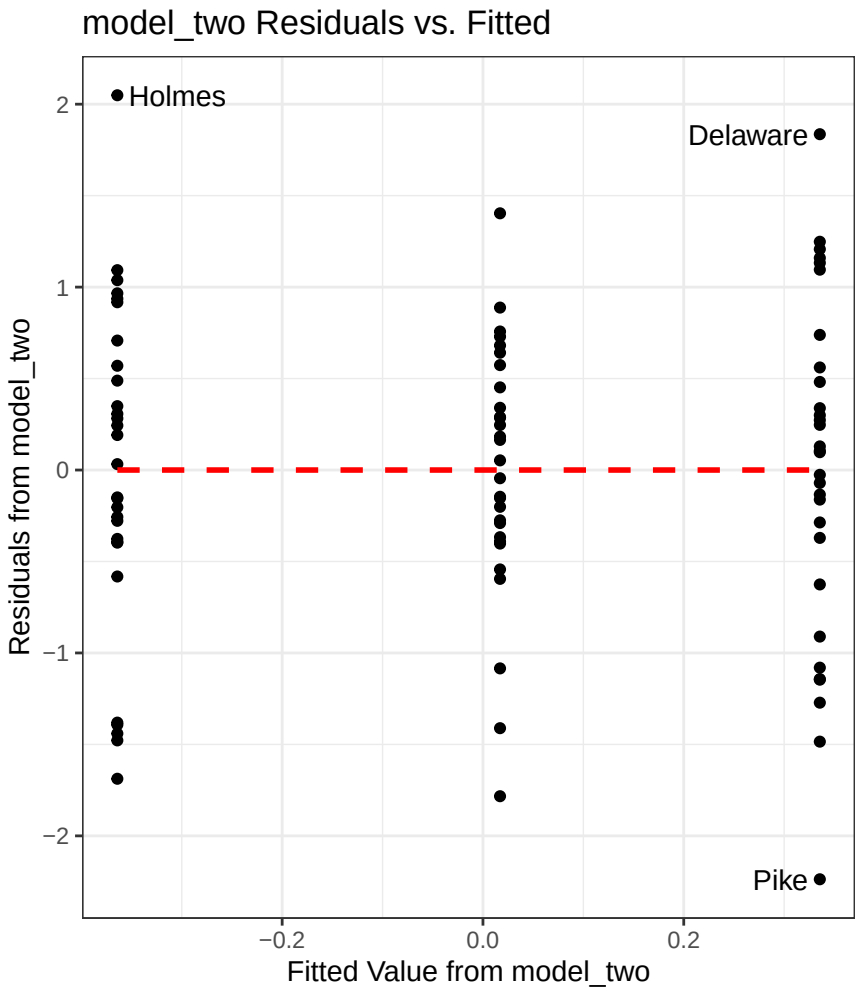
clin_care	min	Q1	median	Q3	max	mean	sd	n	na
Strong	-1.90	-0.23	0.44	0.88	2.17	0.34	0.94	30	0
Middle	-1.77	-0.35	0.07	0.47	1.42	0.02	0.69	29	0
Weak	-2.05	-0.76	-0.33	0.21	1.68	-0.36	0.90	29	0

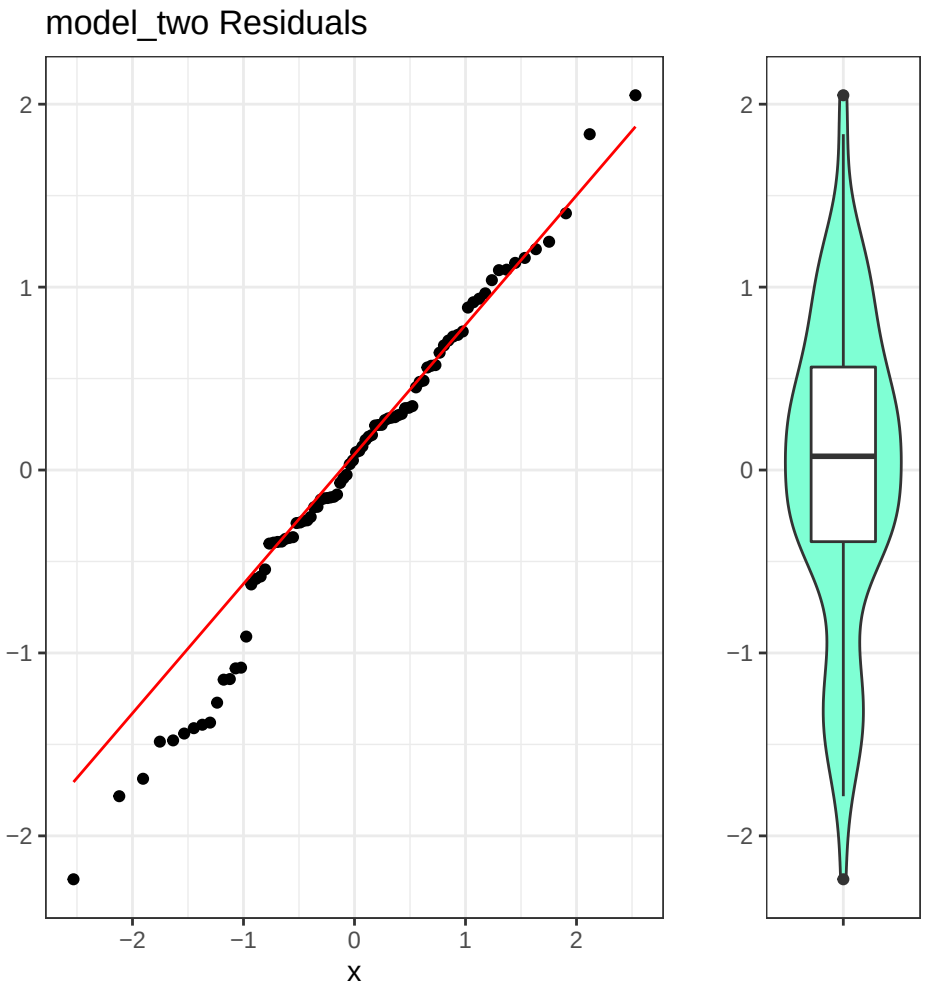
Question 2 Analysis of Variance

```
model_two <- lm(outcomes ~ clin_care, data = ohio20)

anova(model_two)
```

```
Analysis of Variance Table

Response: outcomes
           Df Sum Sq Mean Sq F value   Pr(>F)
clin_care   2  7.232  3.6159  4.9807 0.009007 **
Residuals  85 61.708  0.7260
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Plots for `model_two`





Question 2 Kruskal-Wallis test

```
kruskal.test(outcomes ~ clin_care, data = ohio20)
```

```
    Kruskal-Wallis rank sum test

data:  outcomes by clin_care
Kruskal-Wallis chi-squared = 8.3139, df = 2,
p-value = 0.01566
```

K-Sample Study Design, Comparing Means

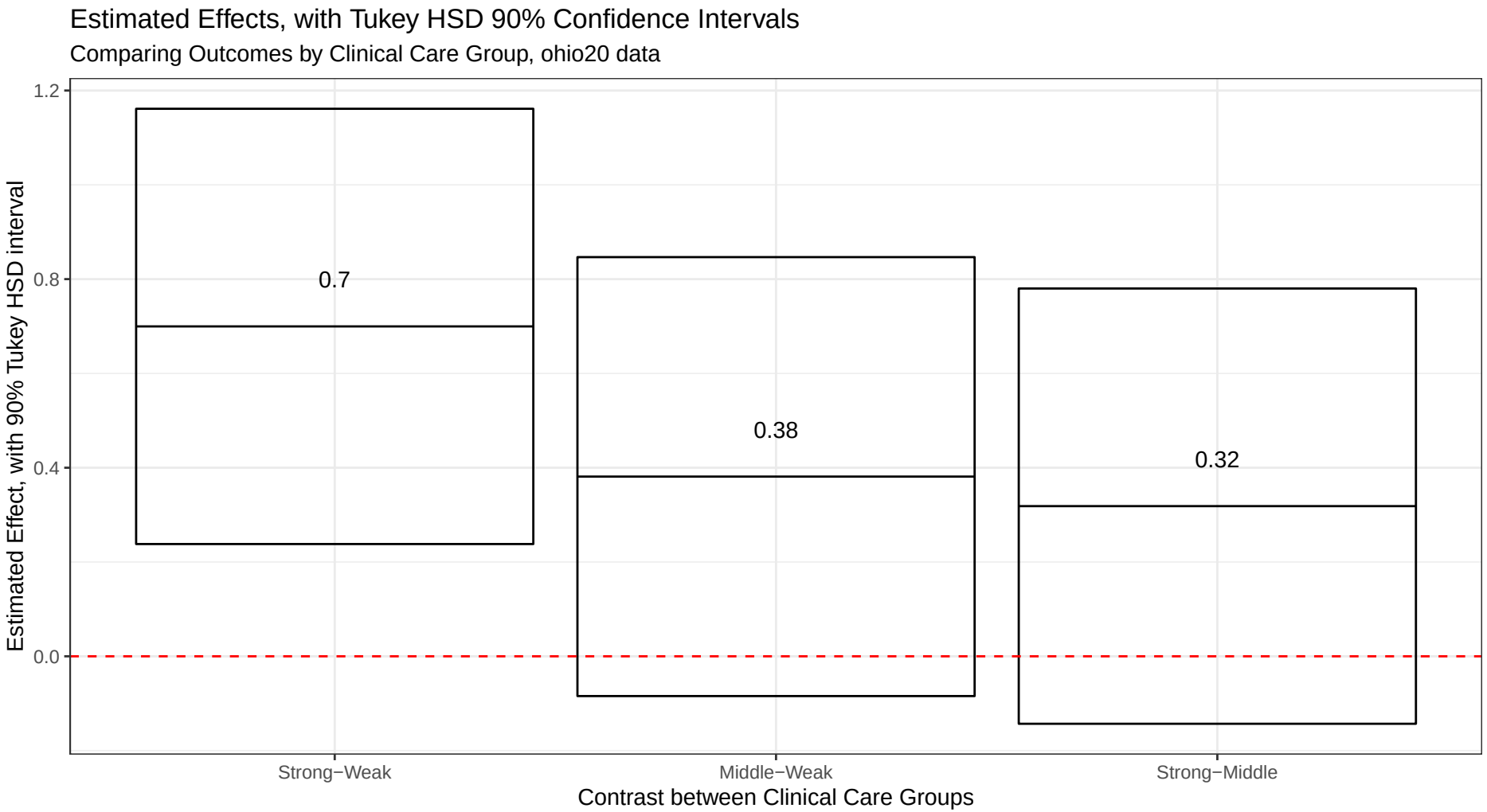
1. What is the outcome under study?
2. What are the (in this case, $K \geq 2$) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the 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 rejects the null, can we identify pairwise comparisons of means that show detectable differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

Question 2: 90% Tukey HSD intervals, tidying

```
model_two <- lm(outcomes ~ clin_care, data = ohio20)
tukey_two <- tidy(TukeyHSD(aov(model_two),
                           ordered = TRUE,
                           conf.level = 0.90))
tukey_two %>% select(-term, -null.value) %>% kable(dig = 3)
```

contrast	estimate	conf.low	conf.high	adj.p.value
Middle-Weak	0.381	-0.084	0.847	0.210
Strong-Weak	0.700	0.238	1.161	0.006
Strong-Middle	0.319	-0.143	0.780	0.327

Plotting Question 2 Tukey HSD intervals



Code for Question 2 Tukey HSD plot

```
ggplot(tukey_two, aes(x = reorder(contrast, -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) +
  labs(x = "Contrast between Clinical Care Groups",
       y = "Estimated Effect, with 90% Tukey HSD interval",
       title = "Estimated Effects, with Tukey HSD 90% Confider
       subtitle = "Comparing Outcomes by Clinical Care Group,
```

For Self-Study: ANOVA Examples about President Trump's 2016 Votes by County

Question 3 (Education)

We have some additional variables in `ohio20`, specifically:

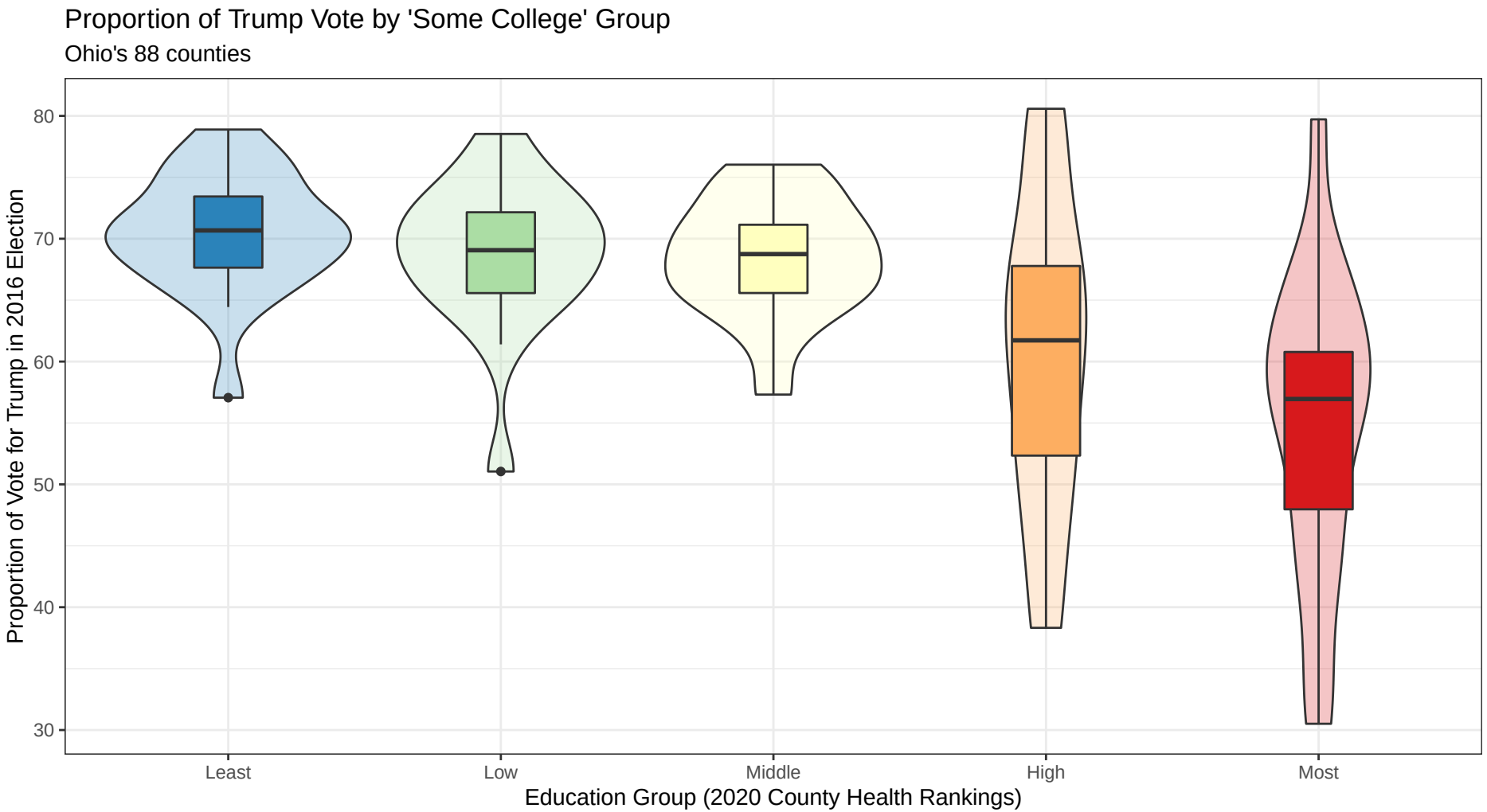
- `trump16` = proportion of the vote cast in 2016 in the county that went to President Trump
- `somecollege` = percentage of adults ages 25-44 with some post-secondary education in the county

Let's break Ohio's counties into 5 groups based on `somecollege`...

```
ohio20 <- ohio20 %>%
  mutate(trump16 = 100*trump16) %>%
  mutate(educ = Hmisc::cut2(somecollege, g = 5)) %>%
  mutate(educ = fct_recode(educ, "Least" = "[20.4,50.3)",
           "Low" = "[50.3,54.3)", "Middle" = "[54.3,59.7)",
           "High" = "[59.7,67.1)", "Most" = "[67.1,85.1]"))
```

Did President Trump's vote percentage in 2016 vary meaningfully across groups of counties defined by educational attainment?

Trump 2016 % by Educational Attainment



Numerical Comparison

```
mosaic::favstats(trump16 ~ educ, data = ohio20) %>%
  rename(na = missing) %>% kable(digits = 2)
```

educ	min	Q1	median	Q3	max	mean	sd	n	na
Least	57.06	67.64	70.67	73.44	78.89	70.34	5.06	18	0
Low	51.05	65.57	69.06	72.16	78.53	68.72	6.17	18	0
Middle	57.31	65.58	68.75	71.14	76.03	68.39	4.89	17	0
High	38.32	52.34	61.72	67.78	80.58	60.42	12.83	18	0
Most	30.51	47.97	56.95	60.78	79.72	55.08	12.51	17	0

Analysis of Variance (ANOVA) testing: Question 3

Does the mean trump16 result differ detectably across the educ groups?

```
model_3 <- lm(trump16 ~ educ, data = ohio20)

tidy(model_3, conf.int = 0.90) %>%
  select(term, estimate, std.error,
         conf.low, conf.high, p.value) %>% kable(dig = 2)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	70.34	2.13	66.11	74.58	0.00
educLow	-1.62	3.01	-7.61	4.37	0.59
educMiddle	-1.95	3.05	-8.02	4.13	0.52
educHigh	-9.92	3.01	-15.91	-3.93	0.00
educMost	-15.26	3.05	-21.33	-9.18	0.00

ANOVA for the Linear Model: Question 3

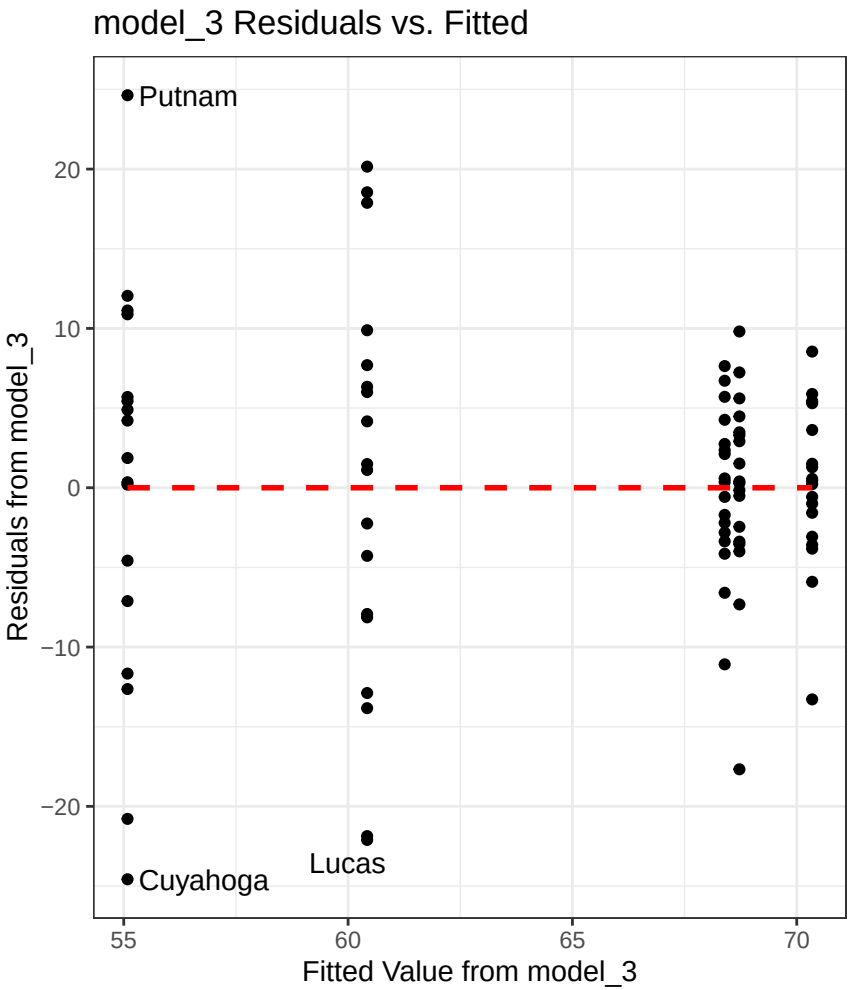
```
anova(model_3)
```

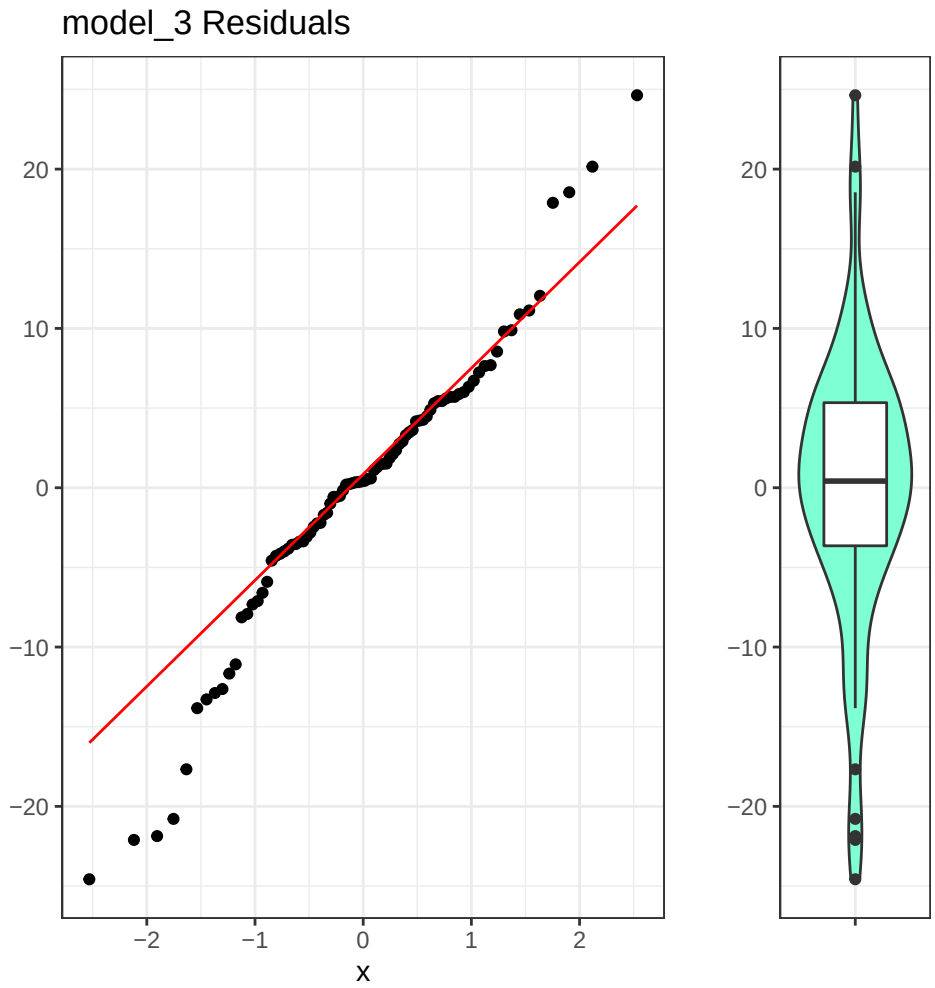
```
Analysis of Variance Table

Response: trump16
           Df Sum Sq Mean Sq F value    Pr(>F)
educ        4 2997.1  749.27  9.1867 3.401e-06 ***
Residuals 83 6769.5   81.56
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glance(model_3) %>%
  select(r.squared, statistic, df, df.residual, p.value)
```

```
# A tibble: 1 x 5
  r.squared statistic    df df.residual    p.value
      <dbl>     <dbl> <dbl>       <int>      <dbl>
```

Residual Plots for `model_3`





Does Kruskal-Wallis give a very different result?

```
kruskal.test(trump16 ~ educ, data = ohio20)
```

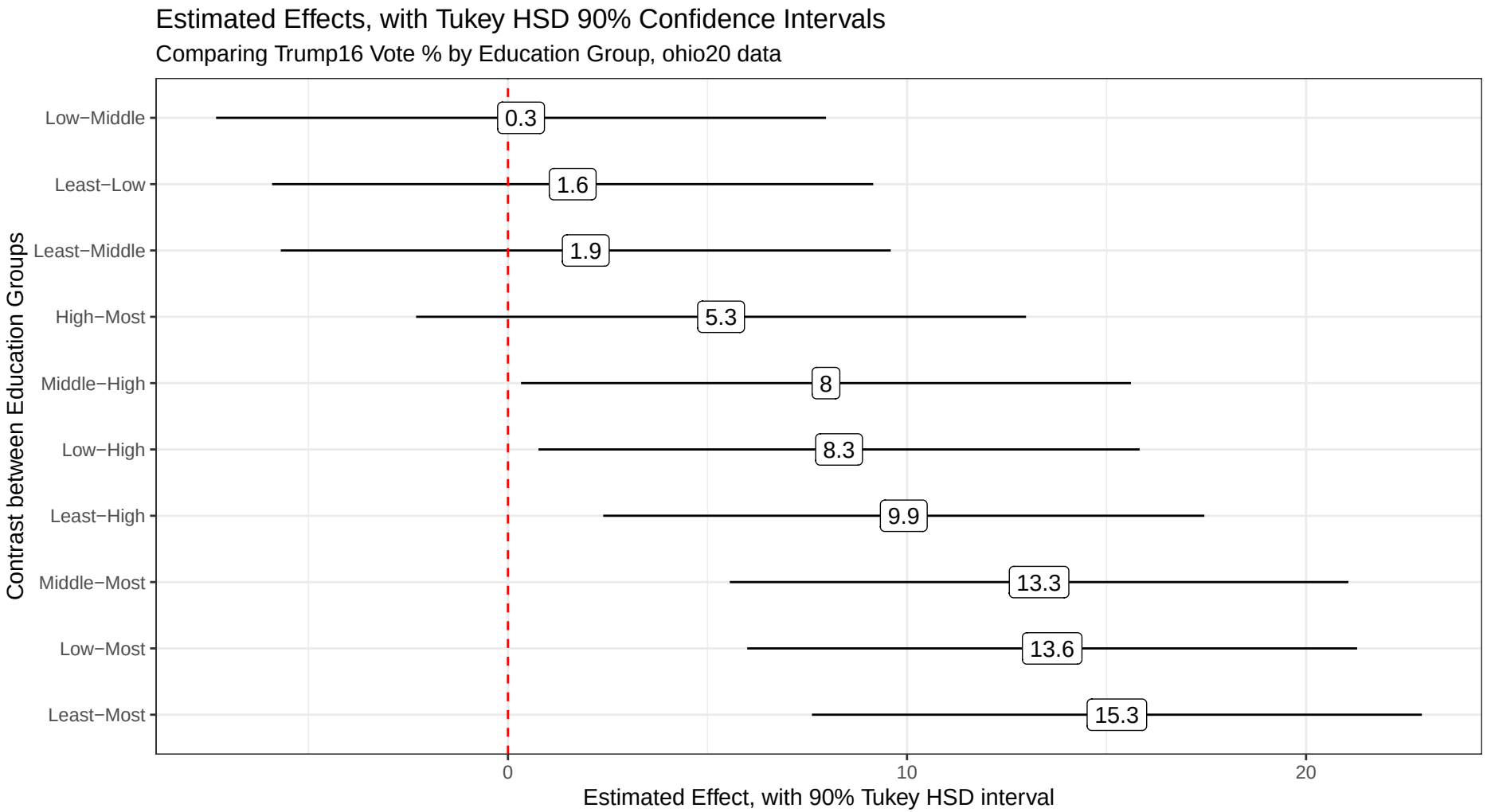
```
    Kruskal-Wallis rank sum test

data:  trump16 by educ
Kruskal-Wallis chi-squared = 25.759, df = 4,
p-value = 3.539e-05
```

Tukey HSD 90% confidence intervals: Example 3

```
tukey_3 <- tidy(TukeyHSD(aov(model_3),
                         ordered = TRUE,
                         conf.level = 0.90))
tukey_3 %>% select(-null.value) %>% kable(dig = 3)
```

term	contrast	estimate	conf.low	conf.high	adj.p.value
educ	High-Most	5.340	-2.302	12.982	0.411
educ	Middle-Most	13.309	5.559	21.060	0.000
educ	Low-Most	13.638	5.995	21.280	0.000
educ	Least-Most	15.259	7.617	22.901	0.000
educ	Middle-High	7.969	0.327	15.611	0.078
educ	Low-High	8.297	0.765	15.829	0.054
educ	Least-High	9.919	2.387	17.451	0.012
educ	Low-Middle	0.328	-7.314	7.970	1.000
educ	Least-Middle	1.950	-5.692	9.592	0.968
educ	Least-Low	1.622	-5.911	9.154	0.983

Plotting Tukey HSD intervals for Example 3



Code for Previous Slide

```
ggplot(tukey_3, aes(x = reorder(contrast, -estimate),
                    y = estimate)) +
  geom_pointrange(aes(ymin = conf.low, ymax = conf.high)) +
  geom_hline(yintercept = 0, col = "red",
             linetype = "dashed") +
  geom_label(aes(label = round_half_up(estimate,1))) +
  coord_flip() +
  labs(x = "Contrast between Education Groups",
   y = "Estimated Effect, with 90% Tukey HSD interval",
   title = "Estimated Effects, with Tukey HSD 90% Confidence I
   subtitle = "Comparing Trump16 Vote % by Education Group, ob
```

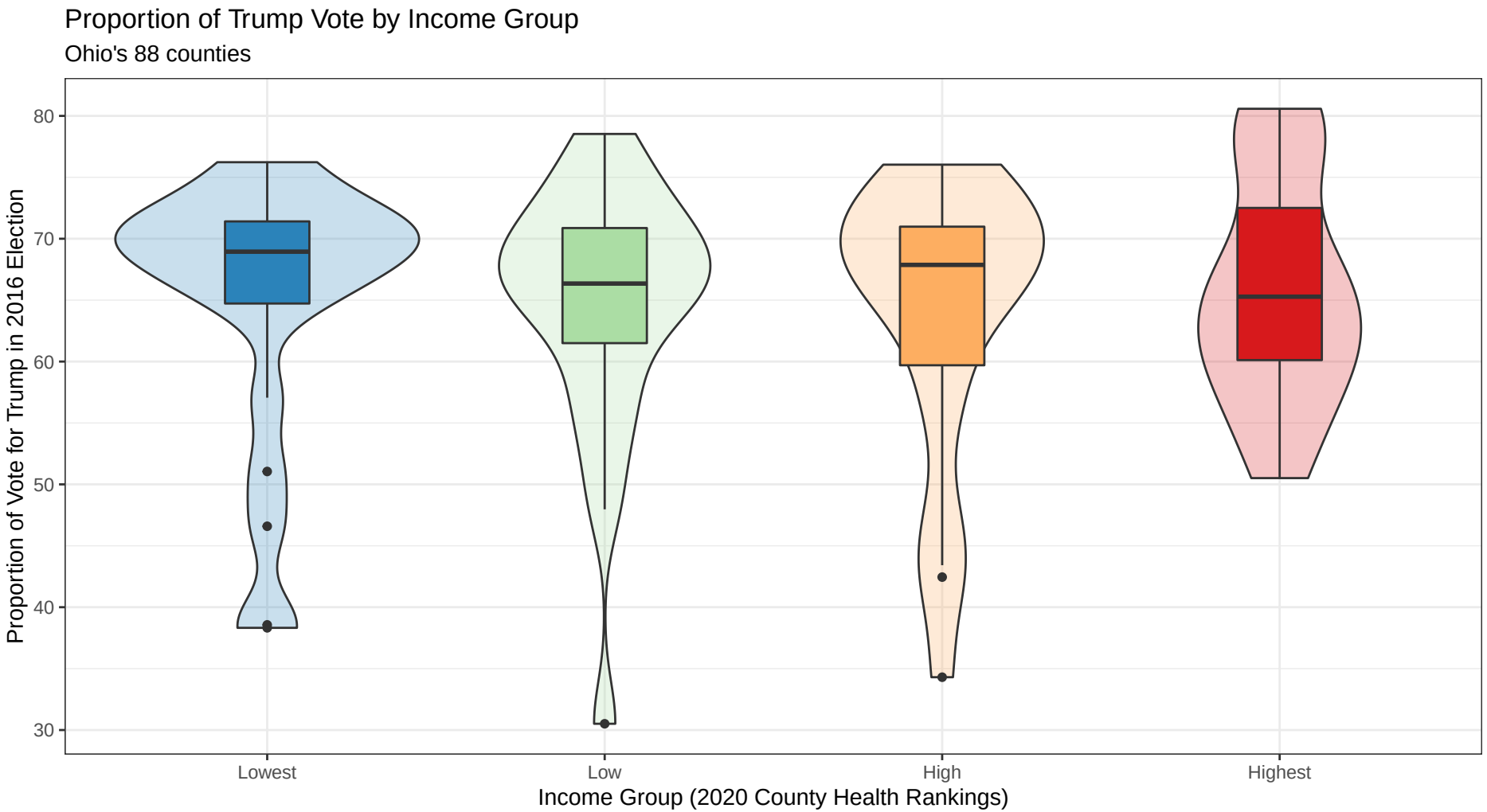
Question 4

Let's break Ohio's counties into 4 groups based on their median `income`...

```
ohio20 <- ohio20 %>%
  mutate(income = Hmisc::cut2(income, g = 4)) %>%
  mutate(income = fct_recode(income, "Lowest" = "[40416, 48792
           "Low" = "[48792, 53904)", "High" = "[53904, 60828)",
           "Highest" = "[60828,103536]"))
```

Did President Trump's vote percentage in 2016 vary meaningfully across income?

Trump 2016 % by Income



Numerical Comparison

```
mosaic::favstats(trump16 ~ income, data = ohio20) %>%
  rename(na = missing) %>% kable(digits = 2)
```

income	min	Q1	median	Q3	max	mean	sd	n	na
Lowest	38.32	64.72	68.94	71.41	76.23	64.71	11.18	22	0
Low	30.51	61.50	66.35	70.87	78.53	64.40	10.71	22	0
High	34.30	59.70	67.87	70.98	76.03	63.73	11.75	22	0
Highest	50.51	60.12	65.28	72.51	80.58	65.80	9.21	22	0

Analysis of Variance (ANOVA) testing

Does the mean `trump16` result differ detectably across the `income` groups?

```
model_4 <- lm(trump16 ~ income, data = ohio20)

tidy(model_4, conf.int = 0.90) %>%
  select(term, estimate, std.error,
         conf.low, conf.high, p.value) %>% kable(dig = 2)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	64.71	2.29	60.15	69.27	0.00
incomeLow	-0.31	3.24	-6.75	6.14	0.93
incomeHigh	-0.98	3.24	-7.42	5.47	0.76
incomeHighest	1.09	3.24	-5.36	7.54	0.74

ANOVA for the Linear Model

```
anova(model_4)
```

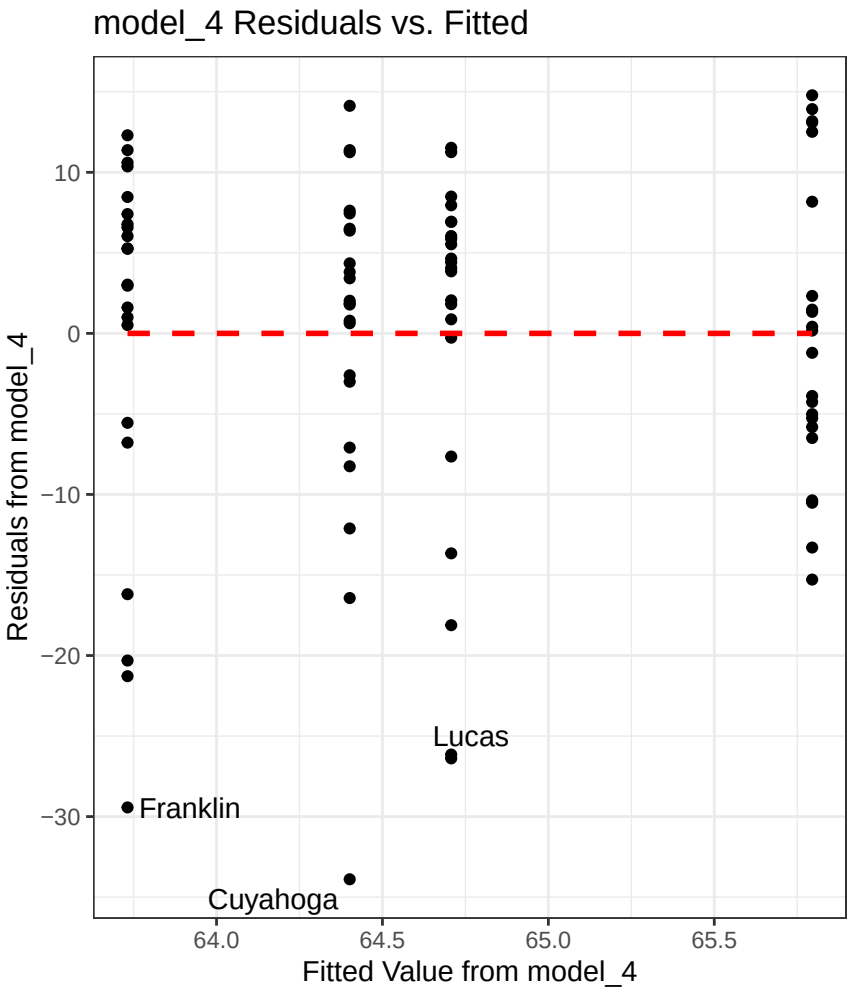
```
Analysis of Variance Table

Response: trump16
          Df Sum Sq Mean Sq F value Pr(>F)
income     3   48.8  16.272  0.1407 0.9354
Residuals 84 9717.8 115.688
```

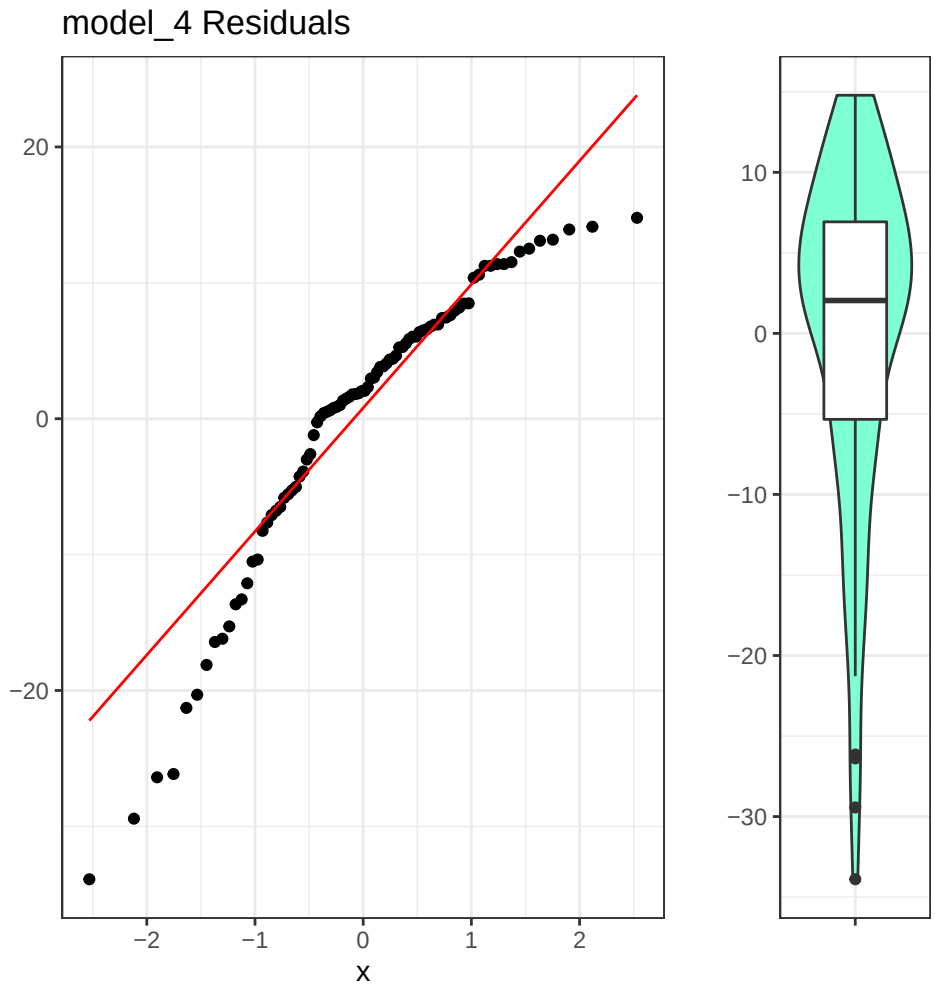
```
glance(model_4) %>%
  select(r.squared, statistic, df, df.residual, p.value)
```

```
# A tibble: 1 x 5
  r.squared statistic    df df.residual p.value
      <dbl>     <dbl> <dbl>       <int>   <dbl>
1   0.00500     0.141     3          84   0.935
```

So, what's the conclusion?

Residual Plots for `model_4`





Does Kruskal-Wallis give a different result?

```
kruskal.test(trump16 ~ income, data = ohio20)
```

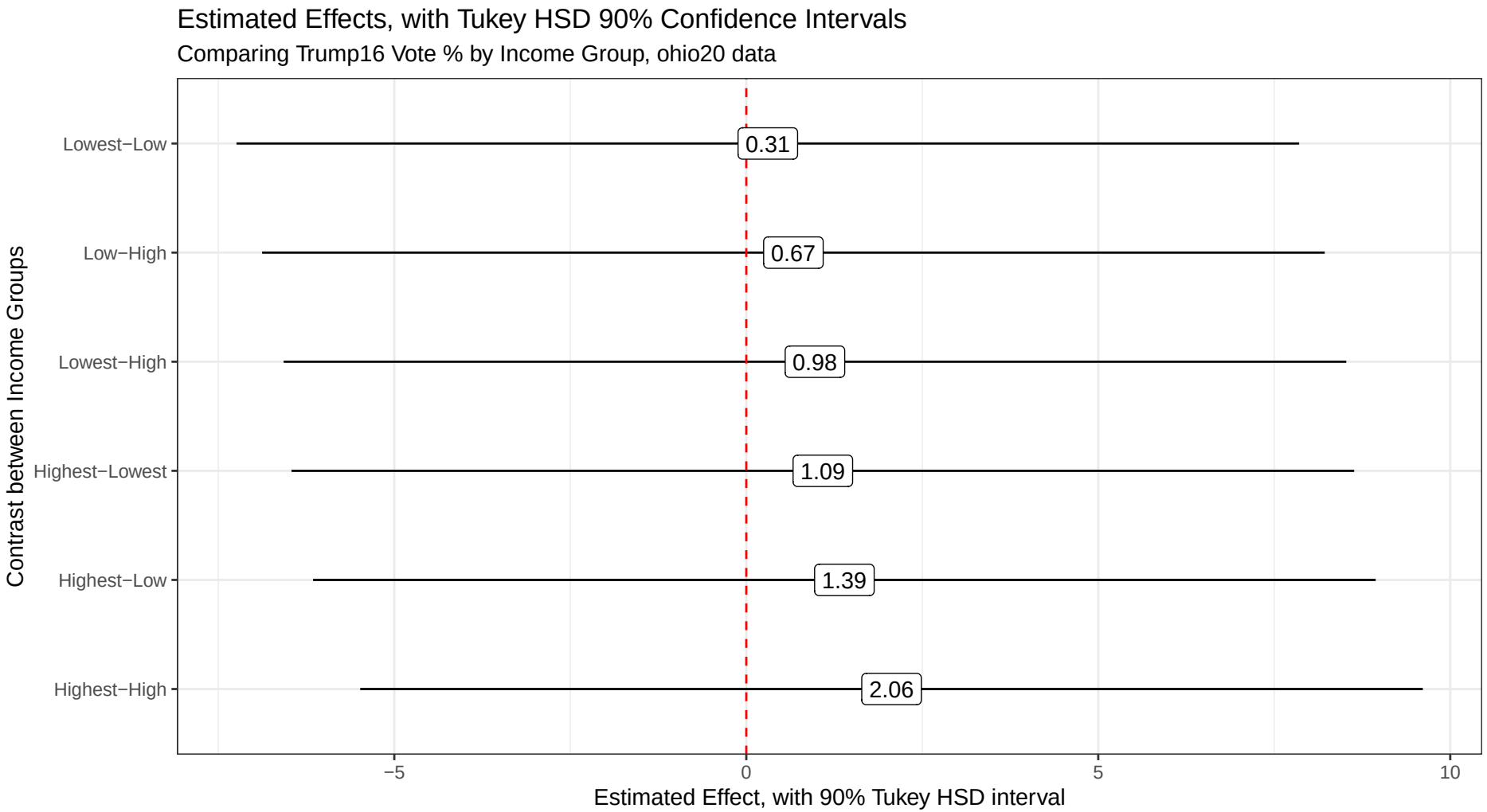
```
    Kruskal-Wallis rank sum test

data:  trump16 by income
Kruskal-Wallis chi-squared = 0.35787, df = 3,
p-value = 0.9488
```

Tukey HSD 90% confidence intervals: Income Groups

```
tukey_4 <- tidy(TukeyHSD(aov(model_4),
                         ordered = TRUE,
                         conf.level = 0.90))
tukey_4 %>% select(-null.value) %>% kable(dig = 3)
```

term	contrast	estimate	conf.low	conf.high	adj.p.value
income	Low-High	0.670	-6.878	8.217	0.997
income	Lowest-High	0.975	-6.572	8.523	0.990
income	Highest-High	2.063	-5.484	9.611	0.920
income	Lowest-Low	0.306	-7.241	7.853	1.000
income	Highest-Low	1.394	-6.154	8.941	0.973
income	Highest-Lowest	1.088	-6.460	8.635	0.987

Plotting Tukey HSD intervals (Income Groups)



K-Sample Study Design, Comparing Means

1. What is the outcome under study?
2. What are the (in this case, $K \geq 2$) 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 detectable differences between population means?
9. If an overall test rejects the null, can we identify pairwise comparisons of means that show detectable differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?