431 Class 14

thomaselove.github.io/431

2021-10-07

Today's R Setup

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

library(broom)
library(Hmisc) # for smean.cl.boot(), mostly
library(janitor)
library(knitr)
library(magrittr)
library(naniar)
library(tidyverse)

source("data/Love-boost.R") # for bootdif() function

theme_set(theme_bw())
```

Today's Agenda

- Comparing Two Population Means
  - In a Study using Independent Samples
    - T tests (Pooled and Welch) and Bootstrap Approaches
  - In a Study using Matched (Paired) Samples
    - T test and Bootstrap Approaches
- New Examples: Two Studies from the Cleveland Clinic

Paired vs. Independent Samples

Suppose you can afford to measure n = 400 outcome values and want to compare the outcome's mean under exposure A to the outcome's mean under exposure B. Consider these two designs:

1. Select a random sample from the population of interest containing 200 people, each of whom provide you with an outcome under exposure A, and then provide you with an outcome under exposure B.
2. Select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B.

- What are the main differences between the studies?

Paired vs. Independent Samples

Suppose you can afford to measure n = 400 outcome values and want to compare the outcome's mean under exposure A to the outcome's mean under exposure B. Consider these two designs:

1. Select a random sample from the population of interest containing 200 people, each of whom provide you with an outcome under exposure A, and then provide you with an outcome under exposure B.
2. Select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B.

- What are the main differences between the studies?
- We'll call Study 1 a **paired samples** study, since each result under exposure A is matched to the exposure B result from the same person. Calculating paired B - A differences by person makes sense.

Paired vs. Independent Samples

Suppose you can afford to measure n = 400 outcome values and want to compare the outcome's mean under exposure A to the outcome's mean under exposure B. Consider these two designs:

1. Select a random sample from the population of interest containing 200 people, each of whom provide you with an outcome under exposure A, and then provide you with an outcome under exposure B.
2. Select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B.

- What are the main differences between the studies?
- We'll call Study 1 a **paired samples** study, since each result under exposure A is matched to the exposure B result from the same person. Calculating paired B - A differences by person makes sense.
- We'll call Study 2 an **independent samples** study, where there is no pairing or matching of individual observations across exposure groups.

A Study Involving Two Independent Samples

The Supraclavicular Data

These data come from the Cleveland Clinic's Statistical Education Dataset Repository, which is a great source of examples for me, but the data there cannot be used for Project B (just to let you know in advance.)

The Supraclavicular data come from Roberman et al. "Combined Versus Sequential Injection of Mepivacaine and Ropivacaine for Supraclavicular Nerve Blocks". *Reg Anesth Pain Med* 2011; 36: 145-50.

```
supra_raw <- read_csv("data/Supraclavicular.csv") %>%
  clean_names()

dim(supra_raw)
```

```
[1] 103  17
```

Supraclavicular Study Objective (in brief)

> *This study consisted of 103 patients, aged 18 to 70 years, who were scheduled to undergo an upper extremity procedure suitable for supraclavicular anesthesia. These procedures were expected to be associated with considerable postoperative pain.*
>
> *We tested the hypothesis that sequential supraclavicular injection of 1.5% mepivacaine followed 90 seconds later by 0.5% ropivacaine provides a quicker onset and a longer duration of analgesia than an equidose combination of the 2 local anesthetics.*
>
> *Patients were randomly assigned to either (1) combined group-ropivacaine and mepivacaine mixture; or (2) sequential group-mepivacaine followed by ropivacaine. The primary outcome was time to 4-nerve sensory block onset.*

All quotes here are from the Supraclavicular study description

Supraclavicular Variables We'll Study Today

Variable	Description
subject	subject identifier (1-103)
group	1 = mixture, 2 = sequential (randomly assigned)
onset_sensory	Time to 4 nerve sensory block onset (min.)

```r
supra <- supra_raw %>%
  mutate(trt = fct_recode(factor(group),
                              "mixture" = "1",
                              "sequential" = "2")) %>%
  rename(onset = onset_sensory) %>%
  select(subject, trt, onset, group)
```

The supra data

```
supra
```

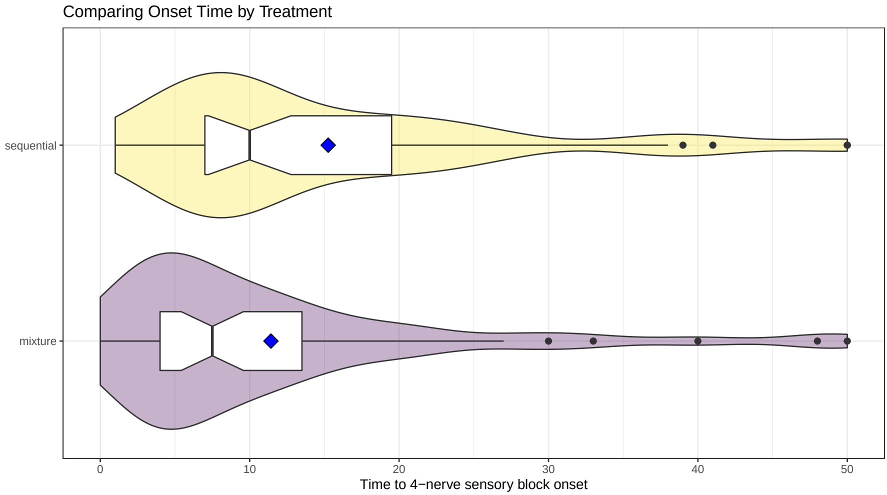
```
# A tibble: 103 x 4
   subject trt         onset group
     <dbl> <fct>       <dbl> <dbl>
 1       1 mixture         0     1
 2       2 sequential      7     2
 3       3 sequential     24     2
 4       4 mixture         4     1
 5       5 mixture        30     1
 6       6 sequential      4     2
 7       7 mixture        12     1
 8       8 sequential     13     2
 9       9 mixture        27     1
10      10 mixture         4     1
# ... with 93 more rows
```

DTDP: Compare onset by treatment

We'll add a blue diamond to indicate the means in each group, too.

```
ggplot(supra, aes(x = trt, y = onset)) +
  geom_violin(aes(fill = trt)) +
  geom_boxplot(width = 0.3, outlier.size = 2, notch = T) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  guides(fill = "none") +
  scale_fill_viridis_d(alpha = 0.3) +
  coord_flip() +
  labs(y = "Time to 4-nerve sensory block onset",
       x = "",
       title = "Comparing Onset Time by Treatment")
```

DTDP: Compare onset by treatment (Result)



Numerical Summaries of Onset Time by Treatment

```r
mosaic::favstats(onset ~ trt, data = supra)
```

```
         trt min Q1 median   Q3 max     mean       sd
1    mixture   0  4    7.5 13.5  50 11.42308 11.45553
2 sequential   1  7   10.0 19.5  50 15.25490 12.08113
   n missing
1 52       0
2 51       0
```

```r
supra %>% group_by(trt) %>%
  summarize(n = n(), mean(onset), sd(onset), var(onset)) %>%
  kable(digits = 3)
```

trt	n	mean(onset)	sd(onset)	var(onset)
mixture	52	11.423	11.456	131.229
sequential	51	15.255	12.081	145.954

Study Description

- We selected 103 subjects from the population of all people:
  - ages 18-70 years
  - scheduled to undergo an upper extremity procedure suitable for supraclavicular anesthesia
  - who would have been eligible to participate in the study (details are fuzzy)
- We have randomly allocated subjects to one of two treatments (sequential or mixture.)
- For each subject, we have an outcome (onset time) associated with the treatment they received.
- The subjects were sampled from the population of interest independently of each other, so that the outcomes we see are not matched (or paired) in any way.

Key Question

Does the (true population) mean onset time differ between the two treatments?

Formal Language of Hypothesis Testing

- Null hypothesis H_0
    - H_0: population mean onset time with sequential $=$ population mean onset time with mixture
    - H_0: difference in population means (sequential - mixture) $= 0$
- Alternative (research) hypothesis H_A or H_1
    - H_1: population mean onset time with sequential $\neq$ population mean onset time with mixture
    - H_1: difference in population means (sequential - mixture) $\neq 0$

Two (related) next steps

1. Given the data, we can then calculate an appropriate test statistic, then compare that test statistic to an appropriate probability distribution to obtain a p value. Small p values favor H_1 over H_0.
2. More usefully, we can use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the difference in population means.

Comparing Two Population Means

If we have independent samples (as we do in this scenario) where the data in the two treatment groups aren't matched or paired in any way, then we have at least four alternatives.

1. Compare population means using a pooled t test or confidence interval

   - This assumes equal population variances of the outcome in the two treatment groups.
   - This also assumes Normality of the outcome in each of the two treatment groups.
   - This is the result of a linear model of outcome $\sim$ treatment.

2. Compare the population means using a Welch's t test or confidence interval

   - This does not assume equal population variances of the outcome.
   - This does assume Normality of the outcome in each of the two treatment groups.

Comparing Two Population Means (continued)

Additional alternatives when working with independent samples:

3. Compare the population means using a bootstrap approach to generate a confidence interval.

   - This does not assume either equal population variances or Normality.

4. Compare the population pseudo-medians (whatever those are) using a Wilcoxon signed rank test or confidence interval

   - This does not assume either equal population variances or Normality, but describes something other than population means, so we'll hold off on this for now.

Using a linear model to obtain pooled t-test results

- Let's start our comparison process with a pooled t test and associated 90% confidence interval, as we can obtain from a linear model.

```
m1 <- lm(onset ~ trt, data = supra)

tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.423	1.632	6.999	0.000	8.714	14.133
trtsequential	3.832	2.319	1.652	0.102	-0.019	7.682

What can we conclude about the difference in means?

Using a Two-Sample `t.test()` approach

We can obtain the same results for the t test comparing two independent samples, and assuming equal variances, with…

```
t.test(onset ~ trt, data = supra,
       var.equal = TRUE, conf.level = 0.90)
```

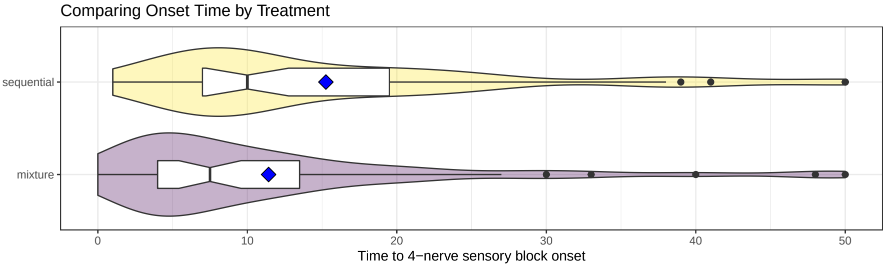
```

    Two Sample t-test

data:  onset by trt
t = -1.652, df = 101, p-value = 0.1016
alternative hypothesis: true difference in means between grou
90 percent confidence interval:
 -7.68230689  0.01865682
sample estimates:
   mean in group mixture mean in group sequential
                11.42308                 15.25490
```

Assessing Pooled T test Assumptions

In preparing a t test with equal variances, we assume that:

- each of the samples (sequential and mixture) are drawn from a Normally distributed population
- each of those populations have the same variance

Do these seem like reasonable assumptions in this case?



Let's first consider dropping the "equal variances" assumption.

The Welch's t test approach

Here is the Welch's t test comparing two independent samples, without assuming equal variances…

```
t.test(onset ~ trt, data = supra, conf.level = 0.90)
```

```
    Welch Two Sample t-test

data:  onset by trt
t = -1.6512, df = 100.47, p-value = 0.1018
alternative hypothesis: true difference in means between group
90 percent confidence interval:
 -7.6845015  0.0208514
sample estimates:
   mean in group mixture mean in group sequential
                11.42308                 15.25490
```

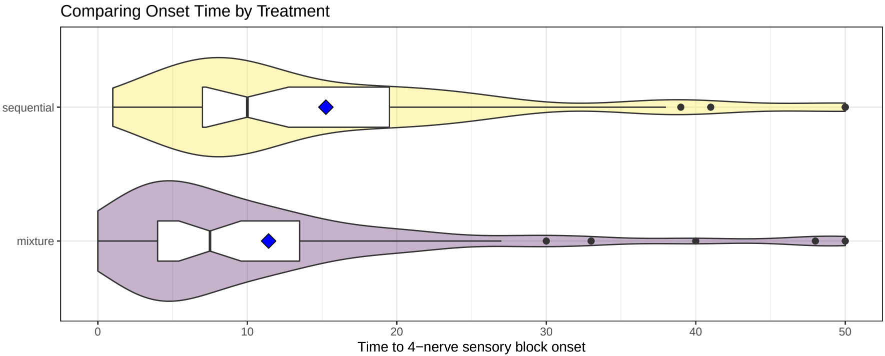
Comparing the two "T tests"

```
t1 <- t.test(onset ~ trt, data = supra, conf.level = 0.90,
             var.equal = TRUE)
w1 <- t.test(onset ~ trt, data = supra, conf.level = 0.90)

bind_rows(tidy(t1), tidy(w1)) %>%
  select(method, estimate, conf.low, conf.high, p.value) %>%
  kable(digits = 3)
```

method	estimate	conf.low	conf.high	p.value
Two Sample t-test	-3.832	-7.682	0.019	0.102
Welch Two Sample t-test	-3.832	-7.685	0.021	0.102

- Note: If we have a **balanced design** with equal sample sizes in the two groups, then these two approaches will yield essentially the same results. Here we have 51 sequential and 52 mixture subjects.

What about the Normality assumption?



- Does it seem reasonable to assume that the onset times are Normally distributed across the populations of sequential and mixed subjects, based on these samples of data?

Using a bootstrap approach

Consider the **bootstrap**, without assuming the population distributions are Normal, or have the same variance, at the expense of requiring some random sampling, which can lead to some conflicts.

- We'll use the `bootdif()` function I've provided in the Love-boost.R script.

```r
set.seed(20211006)
supra %$% bootdif(y = onset, g = trt, conf.level = 0.90,
                  B.reps = 2000)
```

```
Mean Difference            0.05            0.95
     3.83182504      0.07592383      7.52083333
```

Using a bootstrap approach

- If we'd set a different seed or selected a different number of bootstrap replications, we'd get a different result.

```
set.seed(431)
supra %$% bootdif(y = onset, g = trt, conf.level = 0.90,
                  B.reps = 2000)
```

```
Mean Difference            0.05            0.95
     3.83182504     -0.08301282      7.55837104
```

```
set.seed(431)
supra %$% bootdif(y = onset, g = trt, conf.level = 0.90,
                  B.reps = 10000)
```

```
Mean Difference            0.05            0.95
     3.83182504      0.04935897      7.61973982
```

- This doesn't mean to suggest that we "shop around" until we find an appealing result, of course.

The Gathered Set of Estimates

Method	Estimate and 90% CI for $\mu_{Seq} - \mu_{Mix}$
Pooled Two-Sample T	3.83 (-0.02, 7.68)
Welch Two-Sample T	3.83 (-0.02, 7.69)
Bootstrap	3.83 (0.08, 7.52)

All of these results are in minutes (recall 0.08 minutes = 4.8 seconds) so are these **clinically meaningful** differences in this context?

- Do these data involve random sampling?
- What population(s) do these data represent?
- What can we say about the *p* values associated with these approaches?

A Study Involving Two Matched (Paired) Samples (to be discussed in Class 15)

The Hypoxia MAP Data

These data also come from the Cleveland Clinic's Statistical Education Dataset Repository.

Source: Turan et al. "Relationship between Chronic Intermittent Hypoxia and Intraoperative Mean Arterial Pressure in Obstructive Sleep Apnea Patients Having Laparoscopic Bariatric Surgery" *Anesthesiology* 2015; 122: 64-71.

```
hypox_raw <- read_csv("data/HypoxiaMAP.csv") %>%
  clean_names() %>%
  mutate(subject = row_number())

dim(hypox_raw)
```

```
[1] 281  37
```

Background and Study Description

[The Hypoxia MAP study] retrospectively examined the intraoperative blood pressures in 281 patients who had laparoscopic bariatric surgery between June 2005 and December 2009 and had a diagnosis of OSA within two preoperative years.

Time-weighted average (TWA) intraoperative MAP was the main outcome in the study. MAP (or mean arterial pressure) is a term used to describe an average blood pressure in a subject.

MAP is normally between 65 and 110 mmHg, and it is believed that a MAP $>$ 70 mmHg is enough to sustain the organs of the average person. If the MAP falls below this number for an appreciable time, vital organs will not get enough oxygen perfusion, and will become hypoxic, a condition called ischemia.

Our Objective with these Data

We will focus today on two measurements of MAP for each subject (outside of some missing data).

- MAP1 = time-weighted average mean arterial pressure from ET intubation to trocar insertion, in mm Hg.
- MAP2 = time-weighted average mean arterial pressure from trocar insertion to the end of the surgery, in mm Hg.

We are interested in estimating the **difference** between the two MAP levels, across a population of subjects like those enrolled in this study.

Our Key Variables

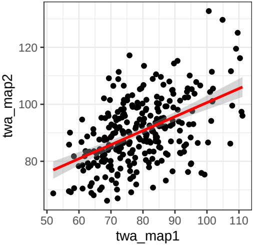
- For each subject, we have two outcomes to compare: their MAP1 and their MAP2.

```
hypox <- hypox_raw %>%
  select(subject, twa_map1, twa_map2) %>%
  mutate(map_diff = twa_map2 - twa_map1)

hypox %>% head(., 4)
```

```
# A tibble: 4 x 4
  subject twa_map1 twa_map2 map_diff
    <int>    <dbl>    <dbl>    <dbl>
1       1     67.9     87.4    19.5
2       2     67.0     83.3    16.3
3       3     91.6     83.0    -8.59
4       4     67.1     79.9    12.8
```

We have Paired Samples in this setting

- Every MAP1 value is connected to the MAP2 value for the same subject. We say that the MAP1 and MAP2 are paired by subject.



Each subject provides a MAP1 and a MAP2

- The pairing is fairly strong here. The Pearson correlation of MAP1 and MAP2 across the subjects with complete data is 0.494.
- It makes sense to calculate the (paired) difference in MAP values for each subject, so long as there aren't any missing data.

Are there any missing values?

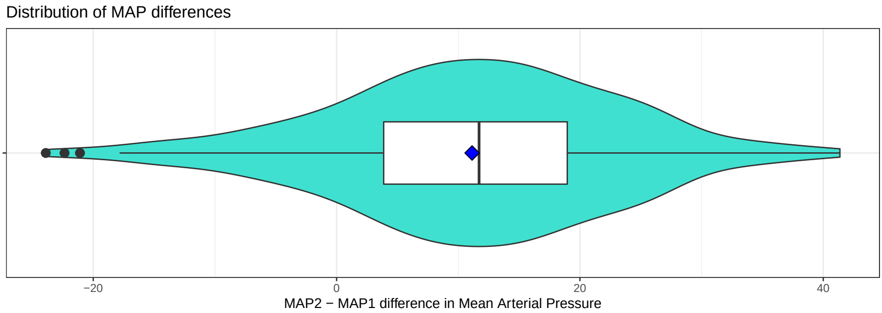
```
miss_var_summary(hypox)
```

```
# A tibble: 4 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 twa_map1      4     1.42
2 map_diff      4     1.42
3 subject       0     0
4 twa_map2      0     0
```

```
hypox <- hypox %>% filter(complete.cases(map_diff))
```

Boxplot of the MAP differences

```
ggplot(data = hypox, aes(x = map_diff, y = "")) +
  geom_violin(fill = "turquoise") +
  geom_boxplot(width = 0.3, outlier.size = 3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  labs(x = "MAP2 - MAP1 difference in Mean Arterial Pressure",
       y = "", title = "Distribution of MAP differences")
```



Numerical Summaries

```
res1 <- as_tibble(bind_rows(
  mosaic::favstats(~ twa_map1, data = hypox),
  mosaic::favstats(~ twa_map2, data = hypox),
  mosaic::favstats(~ map_diff, data = hypox))) %>%
  mutate(item = c("map1", "map2", "map_diff")) %>%
  select(item, n, mean, sd, min, median, max)

res1 %>% kable()
```

item	n	mean	sd	min	median	max
map1	277	79.24274	11.73903	51.96	78.02	111.10
map2	277	90.37921	11.69104	66.17	89.74	132.71
map_diff	277	11.13646	11.78064	-23.90	11.71	41.37

- Is the mean of `map_diff` equal to the difference between the mean of `map2` and the mean of `map1`? Other summaries?

Hypothesis Testing Comparing Paired Samples

- Null hypothesis H_0
  - H_0: population mean of paired differences (MAP2 - MAP1) $= 0$
- Alternative (research) hypothesis H_A or H_1
  - H_1: population mean of paired differences (MAP2 - MAP1) $\neq 0$

Two (related) next steps

1. Given the data, we can then calculate the paired differences, then an appropriate test statistic based on those differences, which we compare to an appropriate probability distribution to obtain a p value. Again, small p values favor H_1 over H_0.
2. More usefully, we can calculate the paired differences, and then use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the population of those differences.

Paired T test via Linear Model

```r
m3 <- lm(map_diff ~ 1, data = hypox)

summary(m3)$coef
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.13646  0.7078298 15.73325 2.971357e-40
```

```r
confint(m3, conf.level = 0.90)
```

```
               2.5 %   97.5 %
(Intercept) 9.743031 12.52989
```

```r
summary(m3)$r.squared
```

```
[1] 0
```

Tidied Regression Model

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	11.136	9.968	12.305

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, statistic, p.value) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.136	0.708	15.733	0

Paired T test via t.test

```r
hypox %$% t.test(map_diff, conf.level = 0.90)
```

```

	One Sample t-test

data:  map_diff
t = 15.733, df = 276, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  9.968265 12.304660
sample estimates:
mean of x 
 11.13646 
```

Paired T Confidence Interval yet another way

```
hypox %$%
  smean.cl.normal(map_diff, conf = 0.90)
```

```
     Mean     Lower     Upper 
11.136462  9.968265 12.304660 
```

The function `smean.cl.normal` (and that's an L, not a 1 after C) comes from the `Hmisc` package.

So does the `smean.cl.boot` function we'll see on the next slide, which will let us avoid the key assumption of Normality for the population of paired differences.

Bootstrap for Comparing Paired Means

```
set.seed(20211006)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 1000)
```

```
     Mean     Lower     Upper 
11.136462  9.932226 12.323684 
```

```
set.seed(123431)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 5000)
```

```
    Mean    Lower    Upper 
11.13646 10.00769 12.30044 
```

Gathered Estimates from our Paired Samples

Method	Estimate and 90% CI	Assumes Normality?
Paired t	11.14 (9.97, 12.30)	Yes
Bootstrap	11.14 (9.93, 12.32)	No

We estimate that the time-weighted average mean arterial pressure is 11.14 mm Hg higher (90% CIs shown above) after trocar insertion than it is during the period from ET intubation to trocar insertion, based on our sample of 277 subjects with complete data in this study.

- Does it matter much whether we assume Normality here?
- What can we say about the *p* values here?
- Is this a random sample of subjects?
- What population do these data represent?

Paired vs. Independent Samples

If you can afford to obtain n = 400 observations to compare means under exposure A to means under exposure B, and you could either:

1. select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B (thus doing an independent samples study), or
2. select a random sample from the population of interest containing 200 people and then randomly assign 100 of them to get exposure A first, and then, a little later, when the effects have worn off, to then receive exposure B, while the other 100 people are assigned to receive B first, then A (thus doing a paired samples study)

Which do you think would be the more powerful study design?

431 Class 15

This is a subtitle.

thomaselove.github.io/431

2021-10-11

Today's R Setup

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

library(broom)
library(Hmisc) # for smean.cl.boot(), mostly
library(janitor)
library(knitr)
library(magrittr)
library(naniar)
library(pwr) # for specialized power functions
library(tidyverse)

source("data/Love-boost.R") # for bootdif() function

theme_set(theme_bw())
```

Today's Agenda

- Comparing Two Population Means Using Paired Samples
    - T test and Bootstrap Approaches
- Power and Sample Size When Comparing Means

The Hypoxia MAP Data

These data also come from the Cleveland Clinic's Statistical Education Dataset Repository.

Source: Turan et al. "Relationship between Chronic Intermittent Hypoxia and Intraoperative Mean Arterial Pressure in Obstructive Sleep Apnea Patients Having Laparoscopic Bariatric Surgery" *Anesthesiology* 2015; 122: 64-71.

```
hypox_raw <- read_csv("data/HypoxiaMAP.csv") %>%
  clean_names() %>%
  mutate(subject = row_number())

dim(hypox_raw)
```

```
[1] 281  37
```

Background and Study Description

[The Hypoxia MAP study] retrospectively examined the intraoperative blood pressures in 281 patients who had laparoscopic bariatric surgery between June 2005 and December 2009 and had a diagnosis of OSA within two preoperative years.

Time-weighted average (TWA) intraoperative MAP was the main outcome in the study. MAP (or mean arterial pressure) is a term used to describe an average blood pressure in a subject.

MAP is normally between 65 and 110 mmHg, and it is believed that a MAP $>$ 70 mmHg is enough to sustain the organs of the average person. If the MAP falls below this number for an appreciable time, vital organs will not get enough oxygen perfusion, and will become hypoxic, a condition called ischemia.

Our Objective with these Data

We will focus today on two measurements of MAP for each subject (outside of some missing data).

- MAP1 = time-weighted average mean arterial pressure from ET intubation to trocar insertion, in mm Hg.
- MAP2 = time-weighted average mean arterial pressure from trocar insertion to the end of the surgery, in mm Hg.

We are interested in estimating the **difference** between the two MAP levels, across a population of subjects like those enrolled in this study.

Our Key Variables

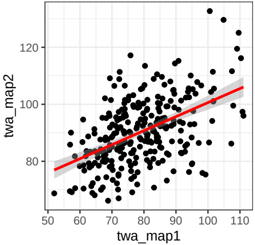
- For each subject, we have two outcomes to compare: their MAP1 and their MAP2.

```
hypox <- hypox_raw %>%
  select(subject, twa_map1, twa_map2) %>%
  mutate(map_diff = twa_map2 - twa_map1)

hypox %>% head(., 4)
```

```
# A tibble: 4 x 4
  subject twa_map1 twa_map2 map_diff
    <int>    <dbl>    <dbl>    <dbl>
1       1     67.9     87.4    19.5
2       2     67.0     83.3    16.3
3       3     91.6     83.0    -8.59
4       4     67.1     79.9    12.8
```

We have Paired Samples in this setting

- Every MAP1 value is connected to the MAP2 value for the same subject. We say that the MAP1 and MAP2 are paired by subject.



Each subject provides a MAP1 and a MAP2

- The pairing is fairly strong here. The Pearson correlation of MAP1 and MAP2 across the subjects with complete data is 0.494.
- It makes sense to calculate the (paired) difference in MAP values for each subject, so long as there aren't any missing data.

Are there any missing values?

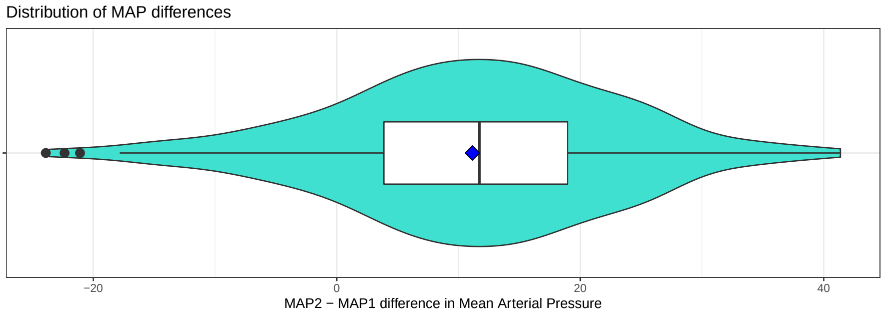
```
miss_var_summary(hypox)
```

```
# A tibble: 4 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 twa_map1      4     1.42
2 map_diff      4     1.42
3 subject       0     0
4 twa_map2      0     0
```

```
hypox <- hypox %>% filter(complete.cases(map_diff))
```

Boxplot of the MAP differences

```
ggplot(data = hypox, aes(x = map_diff, y = "")) +
  geom_violin(fill = "turquoise") +
  geom_boxplot(width = 0.3, outlier.size = 3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  labs(x = "MAP2 - MAP1 difference in Mean Arterial Pressure",
       y = "", title = "Distribution of MAP differences")
```



Numerical Summaries

```
res1 <- as_tibble(bind_rows(
  mosaic::favstats(~ twa_map1, data = hypox),
  mosaic::favstats(~ twa_map2, data = hypox),
  mosaic::favstats(~ map_diff, data = hypox))) %>%
  mutate(item = c("map1", "map2", "map_diff")) %>%
  select(item, n, mean, sd, min, median, max)

res1 %>% kable()
```

item	n	mean	sd	min	median	max
map1	277	79.24274	11.73903	51.96	78.02	111.10
map2	277	90.37921	11.69104	66.17	89.74	132.71
map_diff	277	11.13646	11.78064	-23.90	11.71	41.37

- Is the mean of `map_diff` equal to the difference between the mean of `map2` and the mean of `map1`? Other summaries?

Hypothesis Testing Comparing Paired Samples

- Null hypothesis H_0
  - H_0: population mean of paired differences (MAP2 - MAP1) $= 0$
- Alternative (research) hypothesis H_A or H_1
  - H_1: population mean of paired differences (MAP2 - MAP1) $\neq 0$

Two (related) next steps

1. Given the data, we can then calculate the paired differences, then an appropriate test statistic based on those differences, which we compare to an appropriate probability distribution to obtain a p value. Again, small p values favor H_1 over H_0.
2. More usefully, we can calculate the paired differences, and then use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the population of those differences.

Paired T test via Linear Model

```
m3 <- lm(map_diff ~ 1, data = hypox)

summary(m3)$coef
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.13646  0.7078298 15.73325 2.971357e-40
```

```
confint(m3, conf.level = 0.90)
```

```
               2.5 %   97.5 %
(Intercept) 9.743031 12.52989
```

```
summary(m3)$r.squared
```

```
[1] 0
```

Tidied Regression Model

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  kable(digits = 3)
```

term	estimate	conf.low	conf.high
(Intercept)	11.136	9.968	12.305

```
tidy(m3, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, statistic, p.value) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.136	0.708	15.733	0

Paired T test via t.test

```r
hypox %$% t.test(map_diff, conf.level = 0.90)
```

```
    One Sample t-test

data:  map_diff
t = 15.733, df = 276, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  9.968265 12.304660
sample estimates:
mean of x
 11.13646
```

Paired T Confidence Interval yet another way

```
hypox %$%
  smean.cl.normal(map_diff, conf = 0.90)
```

```
     Mean     Lower     Upper
11.136462  9.968265 12.304660
```

The function `smean.cl.normal` (and that's an L, not a 1 after C) comes from the `Hmisc` package.

So does the `smean.cl.boot` function we'll see on the next slide, which will let us avoid the key assumption of Normality for the population of paired differences.

Bootstrap for Comparing Paired Means

```
set.seed(20211006)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 1000)
```

```
     Mean     Lower     Upper 
11.136462  9.932226 12.323684 
```

```
set.seed(123431)
hypox %$%
  Hmisc::smean.cl.boot(map_diff, conf = 0.90, B = 5000)
```

```
    Mean    Lower    Upper 
11.13646 10.00769 12.30044 
```

Gathered Estimates from our Paired Samples

Method	Estimate and 90% CI	Assumes Normality?
Paired t	11.14 (9.97, 12.30)	Yes
Bootstrap	11.14 (9.93, 12.32)	No

We estimate that the time-weighted average mean arterial pressure is 11.14 mm Hg higher (90% CIs shown above) after trocar insertion than it is during the period from ET intubation to trocar insertion, based on our sample of 277 subjects with complete data in this study.

- Does it matter much whether we assume Normality here?
- What can we say about the *p* values here?
- Is this a random sample of subjects?
- What population do these data represent?

Key Things to Remember in Hypothesis Testing

- Findings can be both statistically significant and practically significant or either or neither.
    - A statistically significant effect is not the same thing as a scientifically interesting effect.
    - A non-significant effect is not the same thing as "no difference".
- When we have large samples, we will regularly find small differences that have a small p value even though they have no practical importance.
- At the other extreme, with small samples, even large differences will often not be large enough to create a small p value.

Errors in Hypothesis Testing

In testing hypotheses, there are two potential decisions and each one brings with it the possibility that a mistake has been made.

–	H_A is true	H_0 is true
Test rejects H_0	Correct	Type I error (False positive
Test retains H_0	Type II error (False negative)	Correct

- A Type I error can only be made if H_0 is actually true.
- A Type II error can only be made if H_A is actually true.

Specifying Error Probabilities (Type I)

If we say we are using 90% confidence, this means:

- we have a 10% significance level
- α, the probability of Type I error, is set to 0.10
- In general, confidence level $=$ 100(1-α).
- The probability of correctly retaining H_0 is designed to be 0.90.

Specifying Error Probabilities (Type II)

- A Type II error is made if the alternative hypothesis is true, but you fail to choose it.
  - The probability depends on exactly which part of the alternative hypothesis is true, so that computing the probability of making a Type II error is not feasible.
- The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true.
- β is defined as the probability of concluding that there was no difference, when in fact there was one (a Type II error).
- A more powerful test will have a lower Type II error rate, β.

Trading off significance (α) and power (β).

In many sample size decisions,

- we find that people set α, the tolerable rate of Type I error, to be 0.05.
- they then often try to set the sample size and other parameters so that the power (1 - β) is at least 0.80.

We'll advocate for thinking hard about the relative costs of Type I and Type II errors.

- The underlying framework that assumes a power of 80% with a significance level of 5% is sufficient for most studies is pretty silly.

Power and Sample Size Calculations

A power calculation is likely the most common element of an scientific grant proposal on which a statistician is consulted.

- The tests that have power calculations worked out in intensive detail using R are mostly those with more substantial assumptions.
    - t tests that assume population normality, common population variance and balanced designs in the independent samples setting
    - paired t tests that assume population normality
- These power calculations are also usually based on tests rather than confidence intervals. Simulation is your friend here.
- This process of doing power and related calculations is far more of an art than a science.

Paired vs. Independent Samples

If you can afford to obtain n = 400 observations to compare means under exposure A to means under exposure B, and you could either:

1. select a random sample from the population of interest containing 400 people and then randomly assign 200 people to receive exposure A and the remaining 200 people to receive exposure B (thus doing an independent samples study), or
2. select a random sample from the population of interest containing 200 people and then randomly assign 100 of them to get exposure A first, and then, a little later, when the effects have worn off, to then receive exposure B, while the other 100 people are assigned to receive B first, then A (thus doing a paired samples study)

Assuming the effect size is unchanged, which seems as though it would be the more powerful study design?

Power of an Independent Samples t test

```
power.t.test(n = 200, delta = 0.25, sd = 1,
             sig.level = 0.10)
```

```
     Two-sample t test power calculation 

              n = 200
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.8025858
    alternative = two.sided

NOTE: n is number in *each* group
```

Power of an Paired Samples t test

```r
power.t.test(n = 200, delta = 0.25, sd = 1,
             sig.level = 0.10, type = "paired")
```

```
     Paired t test power calculation 

              n = 200
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.9698521
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

What sample size do we need?

How many pairs of observations would we need to maintain 80% power?

```
power.t.test(delta = 0.25, sd = 1, sig.level = 0.10,
             power = 0.80, type = "paired")
```

```

     Paired t test power calculation 

              n = 100.2877
          delta = 0.25
             sd = 1
      sig.level = 0.1
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

Note that we'd need 101 pairs of measurements.

What happens when you change assumptions?

In our independent-samples test, we chose

- `n = 200` (per group)
- `delta = 0.25` (the minimum clinically important difference in means that we want to detect)
- `sd = 1` (assumed population standard deviation in each group)
- `sig.level = 0.10` (since we want 90% confidence)

Changing any of these will change the power from the calculated 0.802 to something else.

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change *n* from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?

Which direction will power move in?

Original Setup yielded power = 0.802

- If we change *n* from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change *n* from 200 to 400, power = 0.970

Which direction will power move in?

Original Setup yielded power $= 0.802$

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change n from 200 to 400, power $= 0.970$
- What if we change n from 200 to 100?

Which direction will power move in?

Original Setup yielded power $= 0.802$

- If we change n from 200 to 400, leaving everything else untouched, do you think the power will increase or decrease?
- If we change n from 200 to 400, power $= 0.970$
- What if we change n from 200 to 100?
- If we change n from 200 to 100, power $= 0.546$

Changing the other parameters

Original Setup yielded power = 0.802

What do you think will happen?

New Setup	Resulting Power
a. δ from 0.25 to 0.5	Higher or Lower than 0.802?
b. δ from 0.25 to 0.1	?
c. `sd` from 1 to 2	?
d. `sd` from 1 to 0.5	?
e. α from 0.1 to 0.05	?
f. α from 0.1 to 0.2	?

- Which of these six setups will lead to **larger** power estimates than the original 0.802?

Results of Parameter Changes

Original Setup yielded power $= 0.802$

Change	Resulting power
a. δ from 0.25 to 0.5	0.9996
b. δ from 0.25 to 0.1	0.259
c. `sd` from 1 to 2	0.345
d. `sd` from 1 to 0.5	0.9996
e. α from 0.1 to 0.05	0.703
f. α from 0.1 to 0.2	0.888

- Setups **a** (larger δ), **d** (smaller `sd`) and **f** (larger α, i.e. smaller confidence level) led to larger power estimates than the original setup.

What if you have an unbalanced design?

The most efficient design for an independent samples comparison will be balanced.

- What if we used our original setup for δ, `sd` and α, (which with n $=$ 200 in each group, yielded power $=$ 0.802) but instead we placed
  - 150 subjects into one exposure group, and
  - planned to recruit some number X larger than 150 into the other.
- How many people would we have to recruit into the second exposure group to yield the same power as our original 200 in each group result?

Using `pwr.t2n.test` from the `pwr` package

- Note the use here of d $= \delta / \texttt{sd}$.

```
pwr.t2n.test(n1 = 150, d = 0.25/1, sig.level = 0.10,
             power = 0.802)
```

```
     t test power calculation

             n1 = 150
             n2 = 298.1132
              d = 0.25
      sig.level = 0.1
          power = 0.802
    alternative = two.sided
```

So we can either have 200 and 200, or we can have 150 and 299 to maintain the same power.

Assessing Unbalanced Designs

The power is always stronger for a balanced design than for an unbalanced design with the same overall sample size.

See chapter 19 of the Course Notes for additional examples using the `pwr.t2n.test()` function within the `pwr` package.

One-Sided or Two-Sided

Note that I used a two-sided test to establish my power calculation - in general, this is the most conservative and defensible approach for any such calculation, unless there is a strong and specific reason to use a one-sided approach in building a power calculation, don't.

Coming Up

Class 16:

- Confidence Intervals for a Population Proportion
- Comparing Two Proportions with `twobytwo`
- Power Calculations for Studying Proportions
- Chi-Square Tests for Independence in Larger Contingency Tables

Classes 17-18:

- Analysis of Variance for Comparing $>$ 2 Population Means
- Methods of Dealing with Multiple Comparisons
- More on Statistical Significance and the trouble with p values
- Making Sure You Can Complete Project A's Analyses

and then we return to Multiple Regression

431 Class 16

thomaselove.github.io/431

2021-10-14

Today's R Setup

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

library(Epi) # for twoby2() function
library(mosaic) # not usually something we load
library(readxl) # to import an Excel file
library(pwr) # specialized power functions
library(broom)
library(janitor)
library(knitr)
library(magrittr)
library(tidyverse)

source("data/Love-boost.R") # for twobytwo() function

theme_set(theme_bw())
```

Today's Agenda

- Confidence Intervals for a Population Proportion
  - Five Methods to Accomplish This Task
- Comparing Two Proportions using Independent Samples
  - Standard Epidemiological Format
  - Working with 2x2 Tables
  - Note: we'll discuss paired samples comparisons in 432
- Power Calculations for Comparing Two Proportions
  - With `power.prop.test` for balanced designs
  - With the `pwr` package for unbalanced designs

Section 1

Confidence Intervals for a Population Proportion

Moving on from Means to Proportions

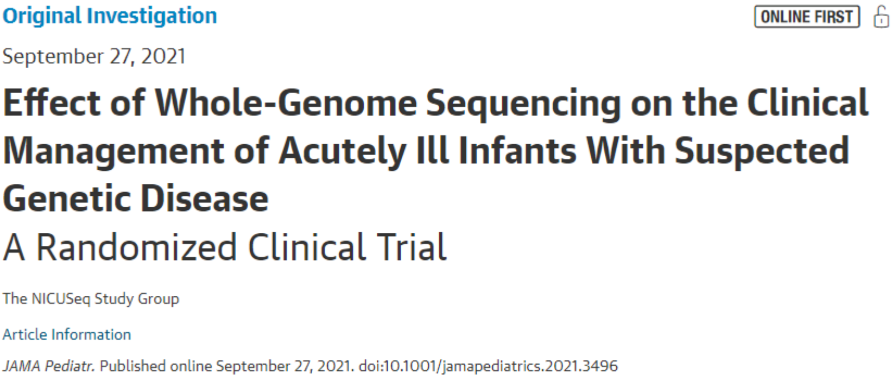
We've focused on creating statistical inferences about a population mean when we have a quantitative outcome. Now, we'll tackle a **categorical** outcome.

We'll estimate a confidence interval around an unknown population proportion, or rate, symbolized with π, on the basis of a random sample of n observations from the population of interest.

The sample proportion is called $\hat{p}$, which is sometimes, unfortunately, symbolized as p.

- This $\hat{p}$ is the sample proportion - not a p value.

An Example from *JAMA Pediatrics*



Original Investigation

ONLINE FIRST

September 27, 2021

Effect of Whole-Genome Sequencing on the Clinical Management of Acutely Ill Infants With Suspected Genetic Disease

A Randomized Clinical Trial

The NICUSeq Study Group

Article Information

JAMA Pediatr. Published online September 27, 2021. doi:10.1001/jamapediatrics.2021.3496

Outcome: Change in Management (COM)

The study involved infants ages 0-120 days admitted to an intensive care unit with a suspected genetic disease.

- For our first example, we focus on a sample of 326 subjects who received whole-genome sequencing testing at some point in the first 60 days after they were enrolled in the study.
- The outcome of interest is whether or not the subject received a change of management (COM) 60 days after their enrollment.

What can we conclude about the true proportion in the population of infants who meet our study criteria who would have a COM?

Loading the Data

```
nicu <- read_excel("data/nicu_seq.xls") %>%
  clean_names()

head(nicu, 3)
```

```
# A tibble: 3 x 3
  subject interv        outcome
    <dbl> <chr>         <chr>
1       1 Early (15)    No_COM
2       2 Early (15)    COM
3       3 Delayed (60) COM
```

```
nicu %>% tabyl(outcome) %>% adorn_totals()
```

```
 outcome   n   percent
     COM  51 0.1564417
  No_COM 275 0.8435583
   Total 326 1.0000000
```

A Confidence Interval for a Proportion

Our first inferential goal will be to produce a **confidence interval for the true (population) proportion** receiving a COM, across all infants who meet study criteria, based on this sample of 326 infants.

A 100(1-α)% confidence interval for the population proportion π can be created by using the standard normal distribution, the sample proportion, $\hat{p}$, and the standard error of a sample proportion, which is defined as the square root of $\hat{p}$ multiplied by $(1 - \hat{p})$ divided by the sample size, n.

Specifically, that confidence interval estimate is $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

where $Z_{\alpha/2} =$ the value from a standard Normal distribution cutting off the top $\alpha/2$ of the distribution, obtained in R by substituting the desired $\alpha/2$ value into: `qnorm(alpha/2, lower.tail=FALSE)`.

- *Note*: This interval is reasonably accurate so long as $n\hat{p}$ and $n(1 - \hat{p})$ are each at least 5.

Estimating π in the NICU data

- We'll build a 95% confidence interval for the true population proportion, so $\alpha = 0.05$
- We have n = 326 subjects
- Sample proportion is $\hat{p} = .156$, since $51/326 = 0.156$.

The standard error of that sample proportion will be

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.156(1-0.156)}{326}} = 0.020$$

Confidence Interval for π = Pr(COM)

Our 95% confidence interval for the true population proportion, π, of infants who have a COM within 60 days is:

$$\hat{p} \pm Z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.156 \pm 1.96(0.020) = 0.156 \pm 0.039$$

or (0.117, 0.195).

To verify that $Z_{0.025} = 1.96\ldots$

```r
qnorm(0.025, lower.tail=FALSE)
```

```
[1] 1.959964
```

Likely Accuracy of this Confidence Interval?

Since $n\hat{p} = (326)(0.156) = 51$ and $n(1 - \hat{p}) = (326)(1 - 0.156) = 275$ are substantially greater than 5, the CI should be reasonably accurate.

What can we conclude from this analysis?

- Point estimate of the population proportion with COM is 0.156
- 95% confidence interval for the population proportion is (0.117, 0.195)

What is the "margin of error" in this confidence interval?

- The entire confidence interval has width 0.078 (or 7.8 percentage points.)
- The margin of error (or half-width) is 0.039, or 3.9 percentage points.

Happily, that's our last "by hand" calculation.

Using R to estimate a CI for a Proportion

I'll discuss five procedures for estimating a confidence interval for a population proportion. Each can be obtained using the `binom.test` function from within the `mosaic` package. For a 95% CI, we use:

```r
mosaic::binom.test(x = 51, n = 326,
                   p = 0.5, conf.level = 0.95, # defaults
                   ci.method = "XXX")
```

where the appropriate `ci.method` is obtained from the table below.

Approach	`ci.method` to be used
Wald	"Wald"
Clopper-Pearson	"Clopper-Pearson" or "binom.test"
Score	"Score" or "prop.test"
Agresti-Coull	"agresti-coull"
Plus4	"plus4"

These 5 Approaches (each is approximate)

1. **Wald** is the "basic biostatistics" method we just calculated, where we estimate the standard error using the sample proportion and then use the Normal distribution to set the endpoints. The Wald interval is always symmetric, and can dip below 0 or above 1.
2. **Clopper-Pearson** is used by `stats::binom.test()` in R as well. It guarantees coverage at least as large as the nominal coverage rate, but may produce wider intervals than the other methods.
3. **Score** is used by `stats::prop.test()` and creates CIs by inverting p-values from score tests. It can be applied with a continuity correction (use ci.method = "prop.test") or without.
4. **Agresti-Coull** is the Wald method after adding Z successes and Z failures to the data, where Z is the appropriate quantile for a standard Normal distribution (1.96 for a 95% CI)
5. **Plus4** is the Wald method after adding 2 successes and 2 failures (so 4 observations) to the data.

Formulas? See Wikipedia's entry: Binomial proportion confidence interval

Method 1: The Wald Procedure

```r
method1 <- binom.test(x = 51, n = 326, conf.level = 0.95,
                      ci.method = "Wald")
```

```r
method1
```

```
Exact binomial test (Wald CI)


data:  51 out of 326
number of successes = 51, number of trials = 326,
                                        p-value < 2.2e-16
alternative hypothesis: true probability of success
                                 is not equal to 0.5
95 percent confidence interval: 0.1170075 0.1958759
sample estimates: probability of success  0.1564417
```

Tidying up a binom.test result from mosaic

```
tidy1 <- tidy(method1)

tidy1 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.1564	0.117	0.1959	51	326

Method 2: The Clopper-Pearson Procedure

```
method2 <- binom.test(x = 51, n = 326, conf.level = 0.95,
                         ci.method = "Clopper-Pearson")

tidy2 <- tidy(method2)

tidy2 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.1564	0.1188	0.2005	51	326

Method 3: The Score Procedure

```
method3 <- binom.test(x = 51, n = 326, conf.level = 0.95,
                      ci.method = "Score")

tidy3 <- tidy(method3)

tidy3 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.1564	0.121	0.1999	51	326

Method 4: The Agresti-Coull Procedure

```
method4 <- binom.test(x = 51, n = 326, conf.level = 0.95,
                      ci.method = "agresti-coull")

tidy4 <- tidy(method4)

tidy4 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.1564	0.1208	0.2001	51	326

Method 5: The Plus4 Procedure

```
method5 <- binom.test(x = 51, n = 326, conf.level = 0.95,
                      ci.method = "plus4")

tidy5 <- tidy(method5)

tidy5 %>%
  select(estimate, conf.low, conf.high, statistic, parameter)
  kable(dig = 4)
```

estimate	conf.low	conf.high	statistic	parameter
0.1564	0.121	0.2002	51	326

Comparison of Methods

```r
res1 <- tidy1 %>% select(estimate, conf.low, conf.high)
res2 <- tidy2 %>% select(estimate, conf.low, conf.high)
res3 <- tidy3 %>% select(estimate, conf.low, conf.high)
res4 <- tidy4 %>% select(estimate, conf.low, conf.high)
res5 <- tidy5 %>% select(estimate, conf.low, conf.high)

res <- bind_rows(res1, res2, res3, res4, res5)
res <- res %>% mutate(
  approach = c("Wald", "Clopper-Pearson", "Score",
               "Agresti-Coull", "Plus4"))
```

Results with too many decimal places

95% confidence intervals based on x = 51 successes in n = 326 trials.

```
res %>% kable(dig = 5)
```

estimate	conf.low	conf.high	approach
0.15644	0.11701	0.19588	Wald
0.15644	0.11875	0.20051	Clopper-Pearson
0.15644	0.12104	0.19985	Score
0.15644	0.12084	0.20005	Agresti-Coull
0.15644	0.12099	0.20022	Plus4

This is way more precision than we can really justify, but I just want you to see that the five results are all (slightly) different.

Results after some rounding

95% confidence intervals based on x = 51 successes in n = 326 trials.

```
res %>% kable(dig = 3)
```

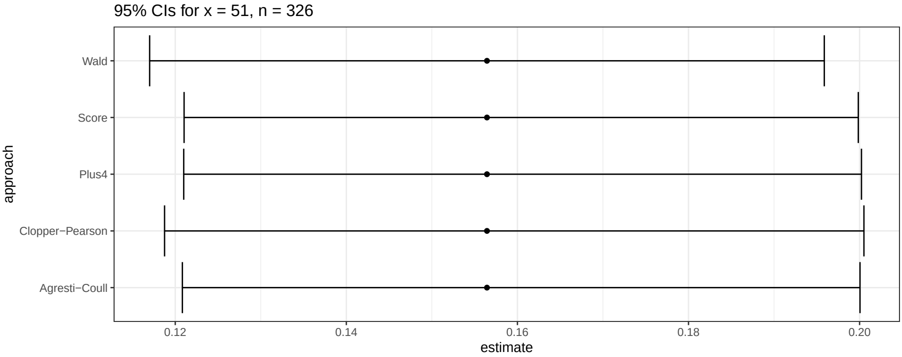
estimate	conf.low	conf.high	approach
0.156	0.117	0.196	Wald
0.156	0.119	0.201	Clopper-Pearson
0.156	0.121	0.200	Score
0.156	0.121	0.200	Agresti-Coull
0.156	0.121	0.200	Plus4

Here's a somewhat more plausible rounding approach.

- Is the distinction between methods important in this scenario?

Plotting the 95% CI Estimates

```
ggplot(res, aes(x = approach, y = estimate)) +
  geom_point() +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high)) +
  coord_flip() +
  labs(title = "95% CIs for x = 51, n = 326")
```



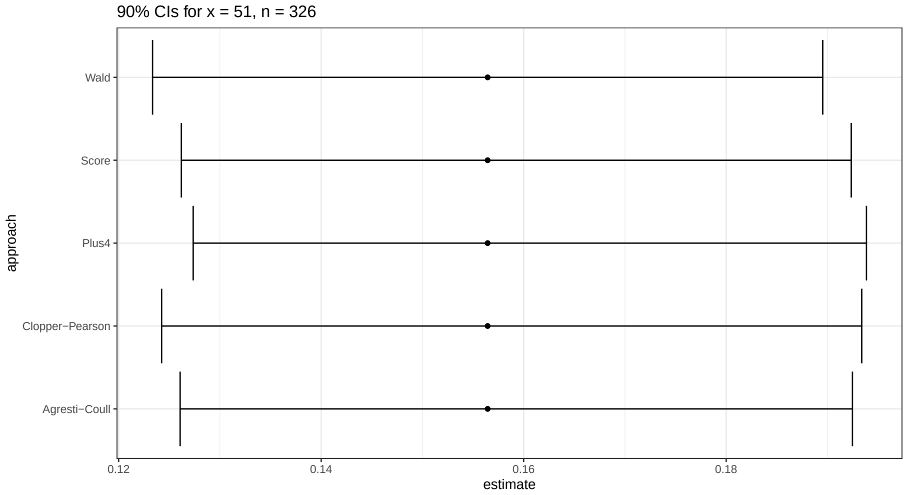
What if we ran 90% Confidence Intervals Instead?

90% confidence intervals based on x = 51 successes in n = 326 trials.

estimate	conf.low	conf.high	approach
0.156	0.123	0.190	Wald
0.156	0.124	0.193	Clopper-Pearson
0.156	0.126	0.192	Score
0.156	0.126	0.192	Agresti-Coull
0.156	0.127	0.194	Plus4

- I've hidden the code here, but it's available in the R Markdown.

Plotting the 90% CI Estimates



Estimating Rates More Accurately

Suppose you have some data involving n independent tries, with x successes. The most natural estimate of the "success rate" in the data is x / n. But, strangely enough, it turns out this isn't an entirely satisfying estimator. Alan Agresti provides substantial motivation for $\frac{x+2}{n+4}$ as an alternative[1]. We'll call this a *Bayesian augmentation*.

Estimates with and without the augmentation will be generally comparable, so long as. . .

a. the sample size is more than, say, 30 subjects, and/or
b. the sample probability of the outcome is between 0.1 and 0.9

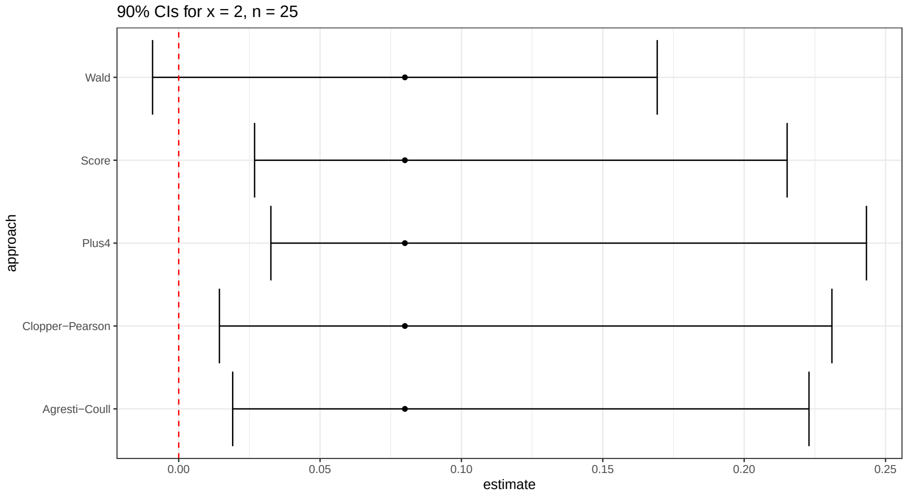
[1] See http://andrewgelman.com/2007/05/15 for instance.

What if we'd had 2 successes in 25 trials instead?

90% confidence intervals based on x = 2 successes in n = 25 trials.

estimate	conf.low	conf.high	approach
0.08	-0.009	0.169	Wald
0.08	0.014	0.231	Clopper-Pearson
0.08	0.027	0.215	Score
0.08	0.019	0.223	Agresti-Coull
0.08	0.033	0.243	Plus4

90% CI Estimates for x = 2, n = 25



What if x = 0 or x = n?

The **Rule of Three** approach is often used.

- An approximate 95% CI for the proportion in a setting where x = 0 in n trials is $(0, \frac{3}{n})$
- An approximate 95% CI for the proportion where x = n in n trials is $(1 - \frac{3}{n}, 1)$

Section 2

Comparing Population Proportions

Comparing Population Proportions

Suppose we want to compare the population proportions π_1 and π_2, based on samples of sizes n_1 and n_2. We can do this using independent samples or using paired samples.

1. The individual observations in exposure group 1 are not linked/matched to individual observations in exposure group 2. (Independent Samples)
2. Each individual observation in exposure group 1 is linked or matched to a specific observation in exposure group 2. (Paired Samples)

The determination as to whether the study design creates paired or independent samples can be determined without summarizing the data. It's a function of the design, not the responses.

A Polling Example

- 200 adult Ohio residents agreed to participate in a poll both two months ago and again today. Each of the 200 people met the polling organization's standards for a "likely voter in the 2020 presidential election". 100 of those polled were under the age of 50 and the rest were 50 or older.
- In between the two polls, a major news event occurred which was relevant to Candidate X.

We asked them the same question at both times: "Are you considering voting for Candidate X?" We are interested in understanding what the data tell us about:

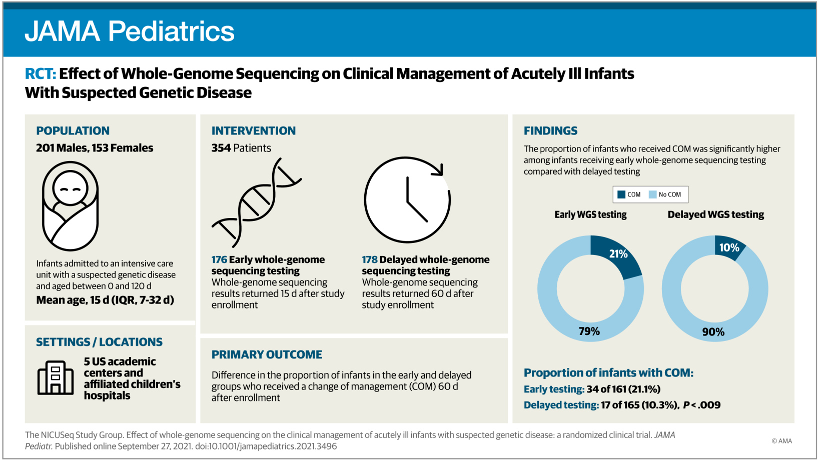
1. Were people under age 50 more likely to be considering Candidate X than people ages 50 and higher?
2. Were people more likely to be considering Candidate X after the news event than before?

Which of these uses *independent* samples, and which *paired* samples?

Section 3

Comparing Proportions using Independent Samples

Visual Abstract for the NICU Sequencing Paper



JAMA Pediatrics

RCT: Effect of Whole-Genome Sequencing on Clinical Management of Acutely Ill Infants With Suspected Genetic Disease

POPULATION

201 Males, 153 Females

Infants admitted to an intensive care unit with a suspected genetic disease and aged between 0 and 120 d

Mean age, 15 d (IQR, 7-32 d)

SETTINGS / LOCATIONS

5 US academic centers and affiliated children's hospitals

INTERVENTION

354 Patients

176 Early whole-genome sequencing testing
Whole-genome sequencing results returned 15 d after study enrollment

178 Delayed whole-genome sequencing testing
Whole-genome sequencing results returned 60 d after study enrollment

PRIMARY OUTCOME

Difference in the proportion of infants in the early and delayed groups who received a change of management (COM) 60 d after enrollment

FINDINGS

The proportion of infants who received COM was significantly higher among infants receiving early whole-genome sequencing testing compared with delayed testing

■ COM □ No COM

Early WGS testing

21%

79%

Delayed WGS testing

10%

90%

Proportion of infants with COM:

Early testing: 34 of 161 (21.1%)

Delayed testing: 17 of 165 (10.3%), *P* < .009

The NICUSeq Study Group. Effect of whole-genome sequencing on the clinical management of acutely ill infants with suspected genetic disease: a randomized clinical trial. *JAMA Pediatr.* Published online September 27, 2021. doi:10.1001/jamapediatrics.2021.3496

NICU Sequencing Example

Let's compare the proportion who have a COM between:

- Group 1: infants tested early (15 d after enrollment)
- Group 2: infants tested later (60 d after enrollment)

```
nicu %>%
  count(interv, outcome) %>% kable()
```

interv	outcome	n
Delayed (60)	COM	17
Delayed (60)	No_COM	148
Early (15)	COM	34
Early (15)	No_COM	127

- How might we rearrange this information? Exposure? Outcome?

The Table We'd Like To Get To

Let's compare the proportion who have a COM between:

- Group 1: infants tested early (at 15 d)
- Group 2: infants tested later (delayed to 60 d)

Standard Epidemiological Format

- rows are the exposure
- columns are the outcome

What do we want in our setting?

Our Goal: Standard Epidemiological Format

- exposure is *intervention* (15 or 60 days)
- columns are *outcome* category (COM, No COM)

```
                  COM       No COM
Early (15 d)       a          b
Delayed (60 d)     c          d
```

Our 2 x 2 Table

```
nicu %>% tabyl(interv, outcome)
```

```
      interv COM No_COM
 Delayed (60)  17    148
   Early (15)  34    127
```

- Is this in standard epidemiological format, with the rows indicating the exposure, and the columns indicating the outcome, and the correct count in the top left cell?

Switching the Rows

We want Early (15) to come first, before Delayed (60):

```
nicu <- nicu %>%
  mutate(interv = fct_relevel(interv, "Early (15)"))

nicu %>% tabyl(interv, outcome)
```

```
       interv COM No_COM
   Early (15)  34    127
 Delayed (60)  17    148
```

Adding Totals

```
nicu %>% tabyl(interv, outcome) %>%
  adorn_totals(where = c("row", "col"))
```

```
      interv COM No_COM Total
   Early (15)  34    127   161
 Delayed (60)  17    148   165
        Total  51    275   326
```

- How many subjects do we have in each exposure group?
- How many subjects fall into each outcome group?

Can we augment the table to help us understand:

- What is the probability of achieving each of the two possible outcomes?
- How do the outcome probabilities differ by exposure group?

Augmenting the Table

```
nicu %>% tabyl(interv, outcome) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages(denom = "row") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns(position = "front")
```

```
      interv          COM       No_COM         Total
  Early (15) 34 (21.1%) 127 (78.9%) 161 (100.0%)
Delayed (60) 17 (10.3%) 148 (89.7%) 165 (100.0%)
       Total 51 (15.6%) 275 (84.4%) 326 (100.0%)
```

- Why am I using `denom = "row"` here?
  Among these subjects, compare the proportion of early (15 d) tested infants with COM to the proportion of late (60 d) tested infants with COM.

- What are the sample estimates for the two rates I am comparing?

2 x 2 Table: Comparing Probabilities

–	COM	No COM	*Total*
Early (15)	34	127	*161*
Delayed (60)	17	148	*165*
Total	*51*	*275*	*326*

- $\Pr(\text{COM} \mid \text{Early}) = 34/161 = 0.211$
- $\Pr(\text{COM} \mid \text{Delayed}) = 17/165 = 0.103$
- The ratio of those two probabilities (risks) is $0.211/0.103 = 2.05$.

Can we build a confidence interval for the relative risk of COM now in the early tested infants as compared to the delayed tested infants?

- The difference in those risks is $0.211 - 0.103 = 0.108$.

How about a confidence interval for the risk difference, too?

2 x 2 Table for NICU Example, Odds Ratio

–	COM	No COM	*Total*
Early (15)	34	127	*161*
Delayed (60)	17	148	*165*
Total	*51*	*275*	*326*

- Odds = Probability / (1 - Probability)
- Sample Odds of COI if Early $= \frac{34/161}{1-(34/161)} = 0.268$
- Sample Odds of COI if Delayed $= \frac{17/165}{1-(17/165)} = 0.115$
- Ratio of these two Odds are 2.331.

In a 2x2 table, odds ratio = cross-product ratio.

- Here, the cross-product estimate $= \frac{34*148}{17*127} = 2.331$.

Can we build a confidence interval for the population odds ratio for COM given "early" as compared to "delayed" testing?

Using twoby2 from the Epi package

Once we have set up the factors for interv and outcome so that the table we produce is in standard epidemiological format, we can plug it into the twoby2 function from the Epi package.

```
twoby2(nicu %$% table(interv, outcome))
```

Results shown on next slide.

twoby2(nicu %$% table(interv, outcome)) results

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : COM
Comparing : Early (15) vs. Delayed (60)

             COM No_COM    P(COM) 95% conf. interval
Early (15)    34    127    0.2112    0.155   0.2810
Delayed (60)  17    148    0.1030    0.065   0.1595

                                   95% conf. interval
             Relative Risk: 2.0497    1.1942   3.5180
         Sample Odds Ratio: 2.3307    1.2430   4.3701
Conditional MLE Odds Ratio: 2.3247    1.1972   4.6617
    Probability difference: 0.1081    0.0292   0.1871

             Exact P-value: 0.0092
        Asymptotic P-value: 0.0083
```

Using twobytwo from the Love-boost.R script

–	COM	No COM	*Total*
Early (15)	34	127	*161*
Delayed (60)	17	148	*165*
Total	*51*	*275*	*326*

Code we need is:

```r
twobytwo(34, 127, 17, 148,  # note order of counts
      "Early", "Delayed", # names of the rows
      "COM", "NoCOM",   # names of the columns
      conf.level = 0.99)  # default is 95% confidence
```

Complete Output shown on the next slide

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : COM
Comparing : Early vs. Delayed

        COM NoCOM    P(COM) 99% conf. interval
Early    34   127    0.2112    0.1400   0.3057
Delayed  17   148    0.1030    0.0561   0.1818

                                   99% conf. interval
             Relative Risk: 2.0497    1.0078   4.1688
         Sample Odds Ratio: 2.3307    1.0202   5.3245
Conditional MLE Odds Ratio: 2.3247    0.9919   5.7786
    Probability difference: 0.1081    0.0037   0.2125

             Exact P-value: 0.0092
        Asymptotic P-value: 0.0083
------------------------------------------------------
```

Another Way to Create The Table

Suppose we didn't have the data, just the visual abstract.

```
t1 <- matrix(c(34, 127, 17, 148), byrow = TRUE, nrow = 2)
rownames(t1) <- c("Early", "Delayed")
colnames(t1) <- c("COM", "No_COM")
addmargins(t1)
```

```
        COM No_COM Sum
Early    34    127 161
Delayed  17    148 165
Sum      51    275 326
```

Hypothesis Testing?

The hypotheses being compared can be thought of in several ways…

- H_0: $\pi_1 = \pi_2$, vs. H_A: $\pi_1 \neq \pi_2$.
- H_0: Pr(COM | Early) = Pr(COM | Delayed) vs. H_A: Pr(COM | Early) $\neq$ Pr(COM | Delayed).
- H_0: rows and columns of the table are *independent*, in that the probability of COM in each row is the same vs. H_A: the rows and columns of the table are *associated*.

```
   Exact P-value: 0.0092
Asymptotic P-value: 0.0083
```

- The `Exact P-value` comes from Fisher's exact test, and is technically exact only if we treat the row and column totals as being fixed.
- The `Asymptotic P-value` comes from a Pearson χ^2 test.
- Neither approach is helpful if we don't have sufficient data to justify inference in the first place.

Bayesian Augmentation in a 2x2 Table?

Original command:

```
twobytwo(34, 127, 17, 148,
         "Early", "Delayed", "COM", "NoCOM",
         conf.level = 0.99)
```

Bayesian augmentation approach: Add two successes and add two failures in each row…

```
twobytwo(34+2, 127+2, 17+2, 148+2,
      "Early", "Delayed", "COM", "NoCOM",
      conf.level = 0.99)
```

Output shown on next slide.

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : COM
Comparing : Early vs. Delayed

        COM NoCOM    P(COM) 99% conf. interval
Early     36   129    0.2182    0.1466   0.3120
Delayed  19   150    0.1124    0.0634   0.1917

                                   99% conf. interval
             Relative Risk: 1.9407    0.9893   3.8071
         Sample Odds Ratio: 2.2032    0.9967   4.8701
Conditional MLE Odds Ratio: 2.1980    0.9691   5.2348
    Probability difference: 0.1058    0.0004   0.2105

             Exact P-value: 0.0118
        Asymptotic P-value: 0.0103
------------------------------------------------------
```

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     29      134      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

```r
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`

```r
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?

Coming Soon

Classes 17-18:

- Chi-Square Testing for Larger Contingency Tables
- Analysis of Variance for Comparing $>$ 2 Population Means
- Methods of Dealing with Multiple Comparisons
- More on Statistical Significance and the trouble with p values
- Making Sure You Can Complete Project A's Analyses

and then we return to Multiple Regression

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     29      134      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

```r
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.

```r
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

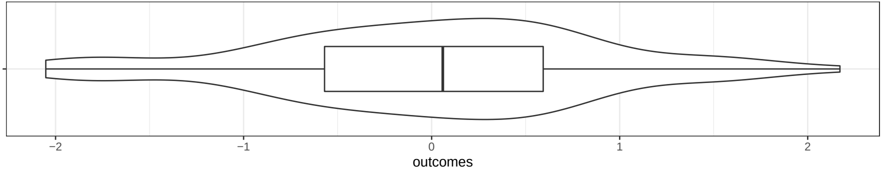
```
ohio20 <- read_xlsx("data/ohio_2020.xlsx") %>%
  mutate(behavior = Hmisc::cut2(rk_behavior, g = 4),
         clin_care = Hmisc::cut2(rk_clin_care, g = 3)) %>%
  mutate(behavior = fct_recode(behavior,
            "Best" = "[ 1,23)", "High" = "[23,45)",
            "Low" = "[45,67)", "Worst" = "[67,88]")) %>%
  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

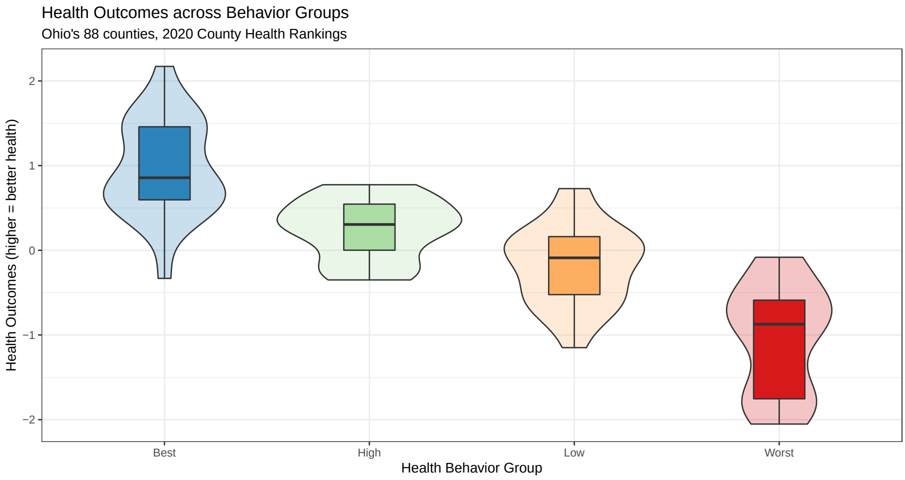
- **outcomes** = quantity that describes the county's premature death and quality of life results, weighted equally and standardized (z scores).
  - Higher (more positive) values indicate better outcomes in this county.
- **behavior** = (Best/High/Low/Worst) reflecting adult smoking, obesity, food environment, 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?



Source: https://www.countyhealthrankings.org/app/ohio/2020/downloads

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?

```r
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?

```r
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

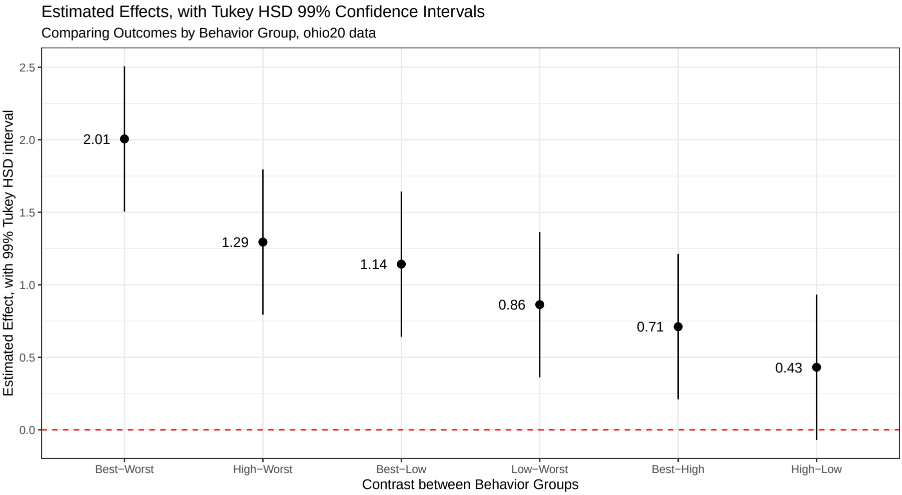
```
  Tukey multiple comparisons of means
    99% family-wise confidence level
    factor levels have been ordered

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
```

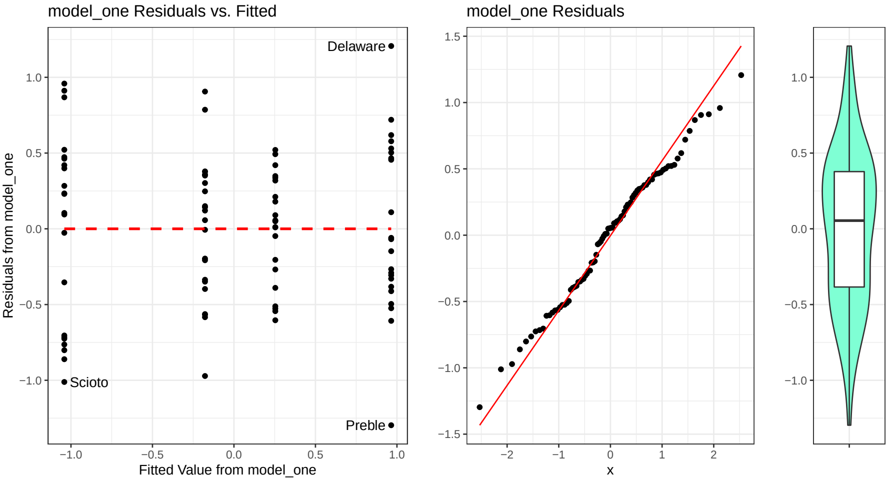
ANOVA Assumptions

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

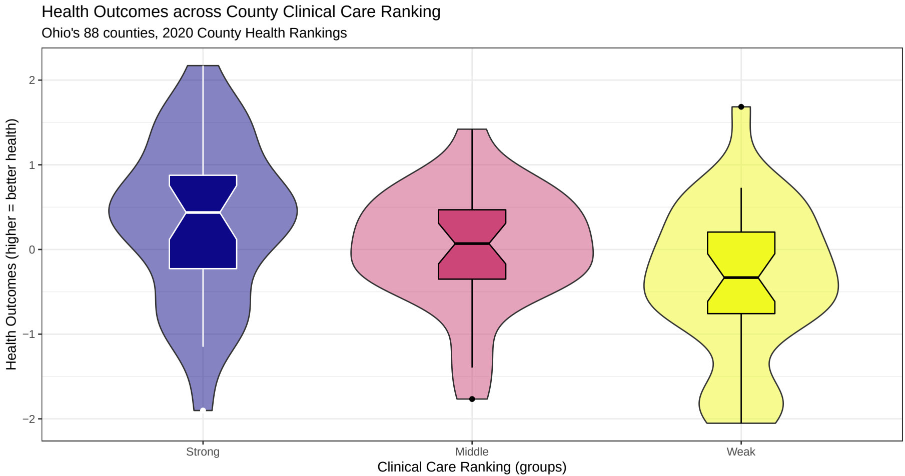
1. What is the outcome under study?
2. What are the (in this case, $K \geq 2$) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the 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?



Source: https://www.countyhealthrankings.org/app/ohio/2020/downloads

Question 2 Numerical Summaries

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

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

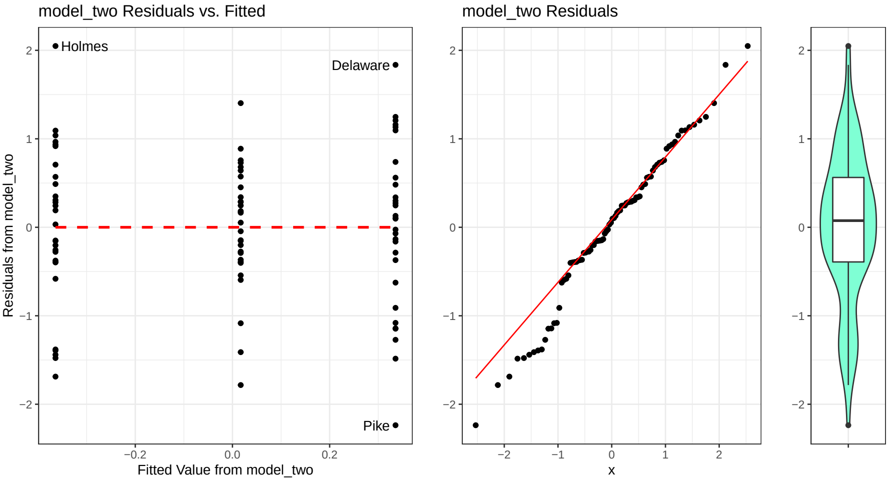
clin_care	min	Q1	median	Q3	max	mean	sd	n	na
Strong	-1.90	-0.23	0.44	0.88	2.17	0.34	0.94	30	0
Middle	-1.77	-0.35	0.07	0.47	1.42	0.02	0.69	29	0
Weak	-2.05	-0.76	-0.33	0.21	1.68	-0.36	0.90	29	0

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

```r
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

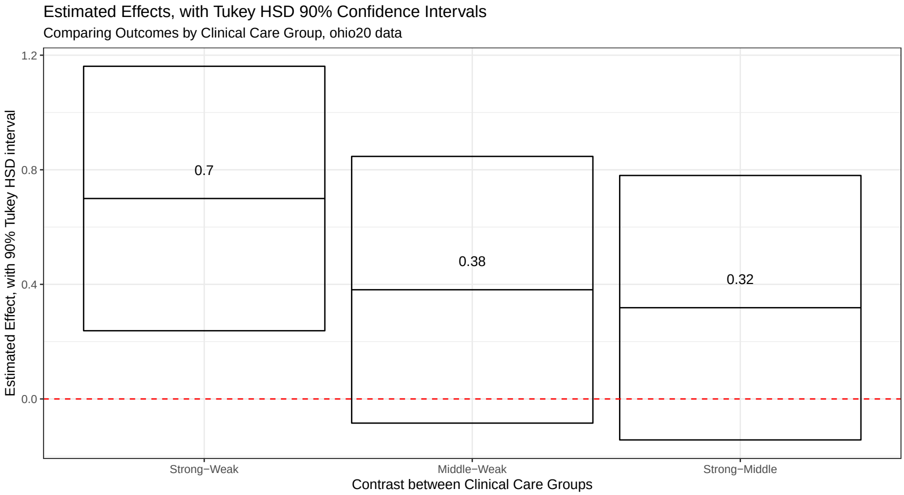
1. What is the outcome under study?
2. What are the (in this case, $K \geq 2$) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the 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% Confiden
       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:

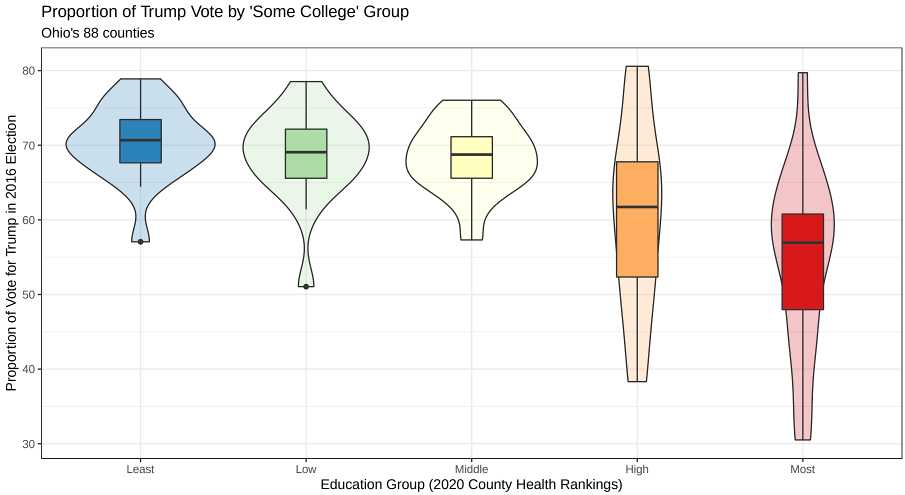
- trump16 = proportion of the vote cast in 2016 in the county that went to President Trump
- somecollege = percentage of adults ages 25-44 with some post-secondary education in the county

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

```r
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

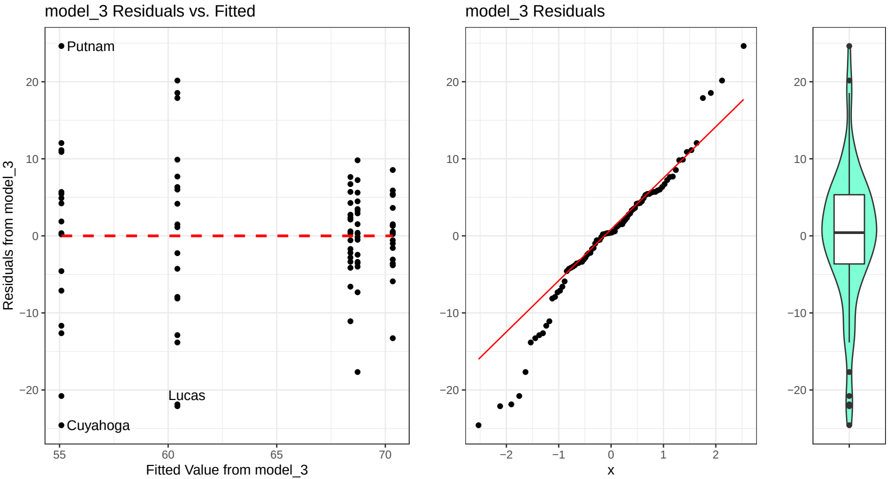
```
anova(model_3)
```

```
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)
```

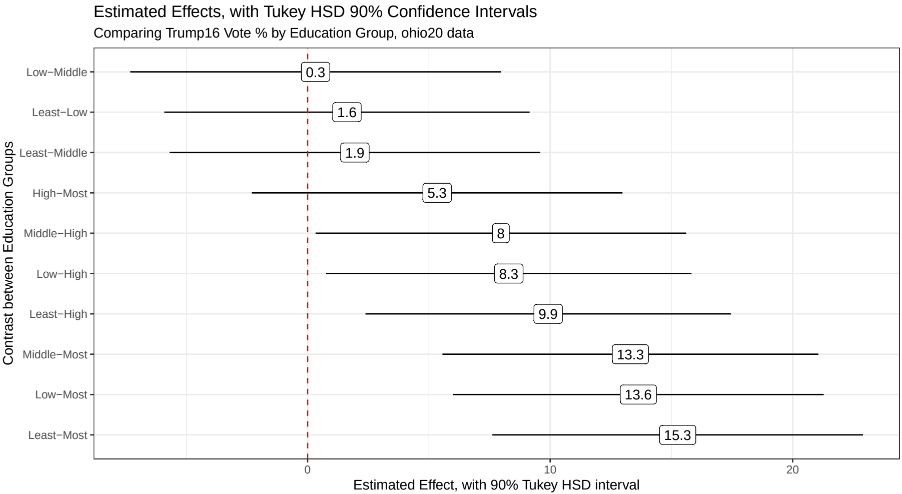
```
    Kruskal-Wallis rank sum test

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
```

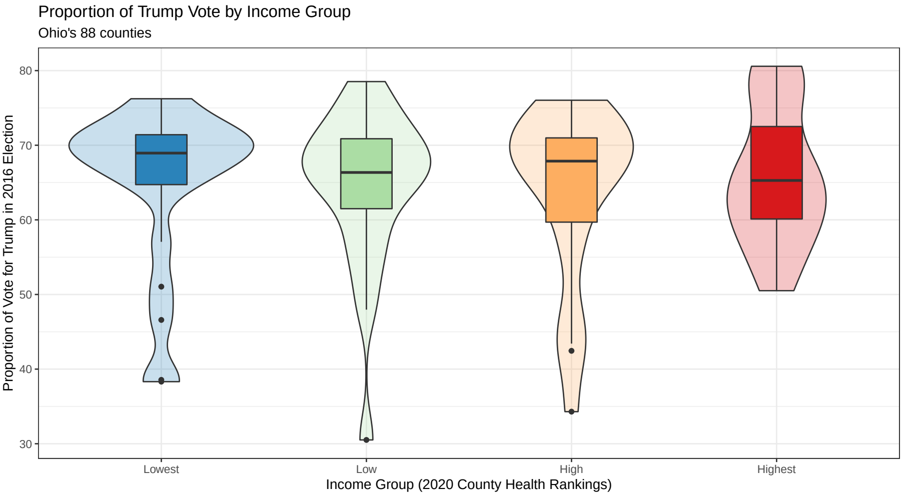
Question 4

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)
```

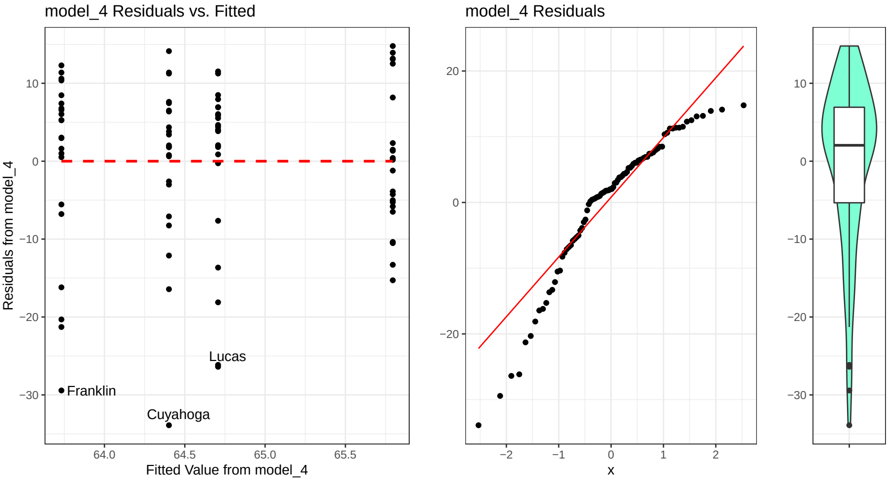
```
Analysis of Variance Table

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?

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

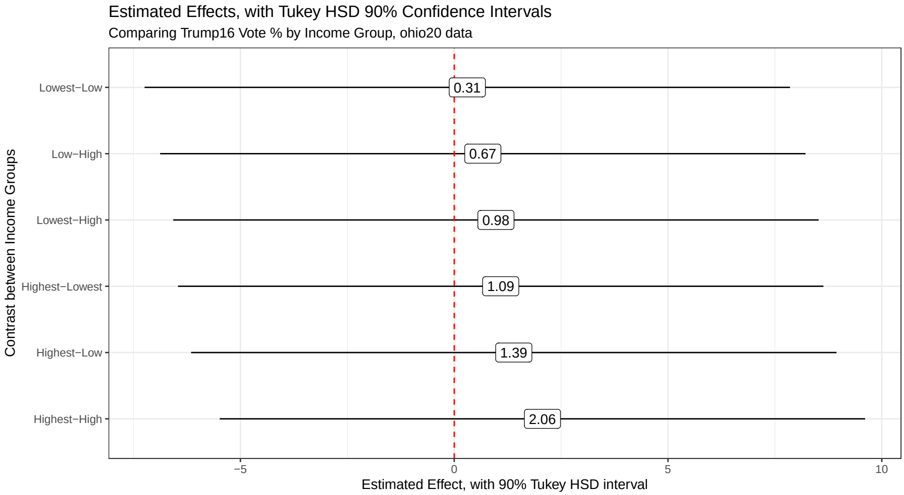
```

    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?