Lab 10 - Math 58B: Inference on Two Means / Medians

your name here

due Tuesday April 11, 2023

Lab Goals

- hypothesis testing of difference in two means and two medians using randomization test
- confidence interval for difference in two means and two medians using bootstrapping

Getting started

- the **infer** package will be used for the R analysis
- the following two applets may help with forming intuition
  - Randomization test applet: http://www.rossmanchance.com/applets/2021/anovashuffle/AnovaShuffle.htm?hideExtras=2
  - Bootstrapping applet: http://www.rossmanchance.com/applets/2021/twoboot/TwoBoot.html?hideExtras=2

Load packages & data

For the lab, we'll use functions from the **tidyverse** and syntax using the **infer** package. The data is on US birth records from the **openintro** package.

> Every year, the US releases to the public a large data set containing information on births recorded in the country. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This is a random sample of 1,000 cases from the data set released in 2014.

```
library(openintro)

data(births14)
```

Let's look at the data!

Try different plots! `geom_histogram`, `geom_point()`, `geom_boxplot()`, `geom_jitter()`, `geom_bar()`,...

(What is the difference between `geom_histogram()` and `geom_bar()` anyway???)

Remember that some plots take both an x and y variable, but some take only one or the other (e.g., `geom_histogram()` only needs one variable, which one?).

```
births14 %>%
  ggplot(aes(x = ___, y = ___, color = ___)) +
  geom_XXX()
```

Variablity of $\overline{X}_1 - \overline{X}_2$ when H_0 is true

The randomization process (which leads to the randomization test) creates different possible values of $\overline{X}_1 - \overline{X}_2$ assuming the null hypothesis is true.

See randomization test applet: http://www.rossmanchance.com/applets/2021/anovashuffle/AnovaShuffle.htm?hideExtras=2

Note that the variability of the statistics (the sampling distribution) can be displayed by applying the same ideas that were used on the difference in sample proportions.

Randomization test comparing the difference in two means

The null hypothesis is that in the population of births, the younger moms and mature moms have, on average, the same gestation time in weeks. Note that the hypotheses are specified using **parameters** which are numbers that describe the population of interest. Because there was no preconceived idea about which group would gestate for longer, the alternative is two-sided.

$H_0 : \mu_Y = \mu_M$

$H_A : \mu_Y \neq \mu_M$

```
set.seed(470)

(wks_obs <- ___ %>%
  specify(___ ~ ___) %>%
  calculate(stat = "diff in means", order = c("___", "___"))  )

null_wks <- ___ %>%
  ___(___ ~ ___) %>%
  ___(null = "___") %>%
  ___(___ = 1000, type = "___") %>%
  ___(stat = "___", order = c("___", "___"))

visualize(___) +
  shade_p_value(obs_stat = ___, direction = "___") +
  xlab("null differences in mean weeks")

null_wks %>%
  get_p_value(obs_stat = ___, direction = "___")
```

Variablity of $\overline{X}_1 - \overline{X}_2$ with no hypothesis stated

The bootstrap process (which leads to a bootstrap confidence interval) creates different possible values of $\overline{X}_1 - \overline{X}_2$ **without** assuming the null hypothesis is true.

See bootstrapping applet: http://www.rossmanchance.com/applets/2021/twoboot/TwoBoot.html?hideExtras=2

Note that the variability of the statistics (the sampling distribution) can be displayed by applying the same ideas that were used on the difference in sample proportions. It is important to observe that the distribution (of $\overline{X}_1 - \overline{X}_2$) is not centered at 0.

```
set.seed(5)
births14 %>%
  specify(___ ~ ___) %>%
  ___(reps = 1000, type = "___") %>%  # no hypothesize step!!!
  ___(stat = "___", order = c("___", "___")) %>%
  visualize() +  # the input gets piped in from the previous steps
  xlab("bootstrap differences in mean weeks")
```

Bootstrap Confidence Interval for $\mu_1 - \mu_2$

Using the distributions above, the CI for $\mu_1 - \mu_2$ can be taken directly from the bootstrap sampling distributions.

We are 93% confident that the true difference in average weeks of gestation between mature versus younger moms is between -0.468 weeks and 0.220 weeks. Note that because the confidence interval overlaps zero, we know that the value of zero is a plausible parameter.

```
set.seed(314)
boot_diff <- ___ %>%
  ___(___ ~ ___) %>%
  ___(reps = 1000, type = "___") %>%  # no hypothesize step!!!
  ___(stat = "___", order = c("___", "___"))


ci_diff_93 <- get_ci(___, level = 0.93)
ci_diff_93

visualize(___) +
  shade_confidence_interval(endpoints = ___, color = "red", fill = "pink") +
  xlab("bootstrap differences in mean weeks")
```

To Turn In

The data

Continue to work with the births dataset as above. But now consider the variables of premature baby or not (`premie`) and `weight` of baby.

Q1. Learning Community Q Describe one thing you learned from someone in your learning community this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Variables

In the study on premature delivery and weight of infant, what are the variables; which is the explanatory and which is the response variable? For each variable, indicate if the variable is continuous, categorical, or binary (where binary is a special case of categorical). What is an observational unit?

Q3. Visualize the data

Using ggplot, create a box plot with the data values (points) on top of the boxplot. One of the axes should be whether or not the baby was premature, and the other axis should be the weight of the baby. Make sure that your plot layers are such that the points are on top of the box. (Also, you'll be able to see the points better if you jitter then than if you just plot the points. What is the difference between `geom_point()` and `geom_jitter()`? Try both!)

Pick a third (categorical) variable to use to color the points.

Q4. Randomization Test for the difference in means

Use a randomization test to test whether or not the **average** weight is higher for full term babies as compared with premature babies. Your result should be consistent with your knowledge of how gestation works (i.e., babies grow inside wombs).

For the hypothesis test, include the following:

- Statement of the null and alternative hypotheses (remember that the test is of a **parameter**, so either state the parameter(s) in words or in symbols or both)
- A visualization of the null sampling distribution. Report what **variable** is being plotted in the distribution? That is, what is the appropriate label for the variable on the x-axis?
- A visualization and calculation for the p-value of the test.
- A conclusion using words like "premature" and "birth weight."

Q5. Randomization Test for the difference in medians

Use a randomization test to test whether or not the **median** weight is higher for full term babies as compared with premature babies. Your result should be consistent with your knowledge of how gestation works (i.e., babies grow inside wombs).

For the hypothesis test, include the following:

- Statement of the null and alternative hypotheses (remember that the test is of a **parameter**, so either state the parameter(s) in words or in symbols or both)
- A visualization of the null sampling distribution. Report what **variable** is being plotted in the distribution? That is, what is the appropriate label for the variable on the x-axis?
- A visualization and calculation for the p-value of the test.
- A conclusion using words like "premature" and "birth weight."

Q6. Bootstrap Confidence Interval for the difference in means

Create a 96% confidence interval for the difference in average birth weight for full term versus premature babies. Your analysis should include:

- Code for the complete bootstrap process.
- A visualization for the bootstrap sampling distribution. Report what **variable** is being plotted in the distribution? That is, what is the appropriate label for the variable on the x-axis?
- A calculation / report of the endpoints of the confidence interval.
- A conclusion in terms of the problem. Your conclusion should report the interval endpoints and use words like "premature" and "birth weight."

Q7. Bootstrap Confidence Interval for the difference in medians

Create a 96% confidence interval for the difference in median birth weight for full term versus premature babies. Your analysis should include:

- Code for the complete bootstrap process.
- A visualization for the bootstrap sampling distribution. Report what **variable** is being plotted in the distribution? That is, what is the appropriate label for the variable on the x-axis?
- A calculation / report of the endpoints of the confidence interval.
- A conclusion in terms of the problem. Your conclusion should report the interval endpoints and use words like "premature" and "birth weight."

Q8. Mean vs Median

- Give one advantage to using means instead of medians to do randomization and bootstrap methods (hint: think power and the plots you made above, the difference is slight, but it is real).
- Give one advantage to using medians instead of means to do randomization and bootstrap methods (hint: think about outliers).

```
praise()
```

```
## [1] "You are glorious!"
```