Lab 10 - Math 58 / 58b: inference on quantitative data

done during lab April 8 or 10, 2020

not due

Lab Goals

- one and two sample inference on mean(s) using `t.test`
- we won't cover this, but included in the lab is the `infer` syntax

Getting Started

Load packages & data

For the lab, we'll use functions from the `tidyverse` and for the material not covered, there is syntax using the `infer` package. The data is on births from North Carolina in 2004.

> In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. 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 this data set.

```
library(tidyverse)

ncbirths <- read_csv("https://www.openintro.org/data/csv/ncbirths.csv")
```

Structure of the lab

EDA

The first part of the lab will be focused on exploring the data and learning about any nuances that might be relevant for our analysis.

> Recall, there is a data wrangling cheat sheet at: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf
>
> Recall, there is a `ggplot2` cheat sheet at: https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf

`t.test`

The second part of the lab will use the function `t.test` to address the following two hypotheses:

Research Question 1: is "40 weeks" an accurate measure for the average human gestation?

$H_0 : \mu = 40$

$H_A : \mu \neq 40$

From: https://en.wikipedia.org/wiki/Gestation

> Human pregnancy can be divided roughly into three trimesters, each approximately three months long. The first trimester is from the last period through the 13th week, the second trimester is 14th–27th week, and the third trimester is 28th–42nd week. Birth normally occurs at a gestational age of about 40 weeks.

Research Question 2: is average gestation independent of the age of the mom (where here age is categorized as "mature mom" or "younger mom"). A "mature mom" is 35 years old or older. [Note: you might suspect

that older moms gestate less because they have higher risk pregnancies, more likely to have multiples, etc. However, without the knowledge, you'd do a two-sided test which is what we will do here.]

$H_0 : \mu_Y = \mu_M$

$H_A : \mu_Y \neq \mu_M$

`infer`

Although we will not cover the methods used within the `infer` syntax (bootstrapping for one sample, randomly permuting for two samples), the code for the relevant analyses is given below. Even without understanding the sampling mechanism, you can see that conclusions are quite similar to the conclusions given by the mathematical model of the t-distribution.

Analysis

Exploratory data analysis

1. What are the cases (observational units) in this data set? How many cases are there in our sample?
2. From how many mothers are we missing gestation data (measured by the `weeks` variable)?
3. Visualize the data using a boxplot and a histogram. What do the plots highlight about the distribution of the data?

Inference with `t.test`

4. Are the technical conditions necessary for inference satisfied? Comment.
5. Write the hypotheses for testing if the average gestation period is consistent with the commonly held belief that humans gestate for 40 weeks.
6. Give the full conclusion associated with the results of the hypothesis test. The conclusion should include English words like "gestation" and "40 weeks".
7. Using the CI part of the `t.test` function, find a 99% confidence interval for the true gestation period in the population. Interpret the interval in context of the data (use English words that give the interpretation – say things like "gestation").
8. Give three possible explanations for why the confidence interval(s) on mean / trimmed mean gestation don't overlap the number 40. Which reason do you think it is? Explain.
9. Repeat the analysis: t-test, interpretation, CI, interpretation for the second research question.

$H_0 : \mu_Y = \mu_M$ $H_A : \mu_Y \neq \mu_M$

Inference with `infer`

None of this will be covered in class, but I thought you might find it interesting to compare the the mathematical approximation (i.e., t-test) above.

CI for one sample mean (using bootstrapping)

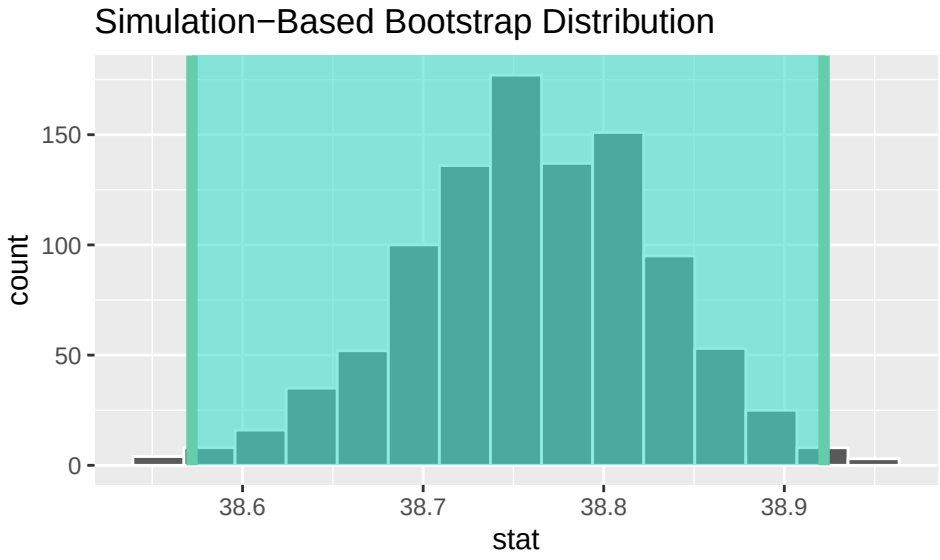
Maybe really what the research shows is that the 50% trimmed mean (the average of the middle 50% of births) is 40 weeks. Using a 99% bootstrap CI, evaluate whether or not the data are consistent with a trimmed mean gestation of 40 weeks.

```
set.seed(47)
# code to use the infer syntax
library(infer)
nc_nona <- ncbirths %>%
```

```
  filter(!is.na(weeks))        # get rid of missing observations

boot_wks <- nc_nona %>%
   specify(response = weeks) %>%
   generate(reps = 1000, type = "bootstrap") %>%
   calculate(stat = "mean", trim=0.25)

(percentile_ci <- get_ci(boot_wks, level = 0.99) )
```

```
## # A tibble: 1 x 2
##   `0.5%` `99.5%`
##    <dbl>   <dbl>
## 1   38.6    38.9
```

```
visualize(boot_wks) +
  shade_confidence_interval(endpoints = percentile_ci)
```




```
# code as is written in the ISCAM textbook
gweeks <- na.omit(ncbirths$weeks)  # keep only the non missing data
resamples <- lapply(1:1000, function(i) sample(gweeks, 998, replace=T))
bootstraptmeans <- sapply(resamples, mean, trim=.25)  # trim 25% off each end
bootstraptmeans %>% quantile(c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 38.56199 38.91802
```

The 99% confidence interval is (38.6 weeks, 38.9 weeks). That is, we are 99% confident that the true population 50% trimmed average gestation time is between 38.6 weeks and 38.9 weeks. Again, 40 weeks is not in the interval, so it does not appear that 40 weeks is the true trimmed average.

Randomization test comparing two means

$H_0 : \mu_Y = \mu_M$

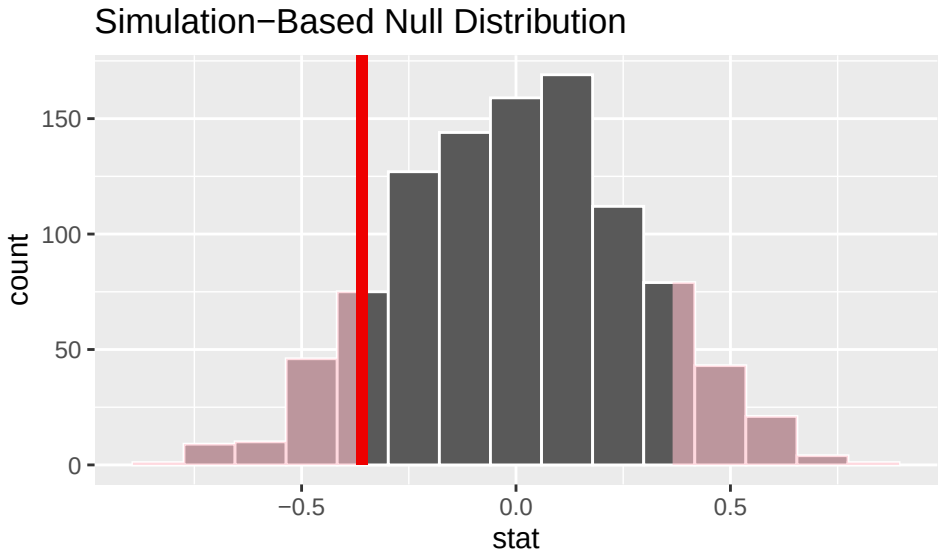
$H_A : \mu_Y \neq \mu_M$

```
set.seed(470)

(diff_obs <- ncbirths %>%
```

```
  specify(weeks ~ mature) %>%
  calculate(stat = "diff in means", order = c("mature mom", "younger mom"))  )
```

```
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 -0.359
```

```
null_wks <- ncbirths %>%
  specify(weeks ~ mature) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("mature mom", "younger mom"))

visualize(null_wks) +
  shade_p_value(obs_stat = diff_obs, direction = "two_sided")
```




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

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.194
```