STA130 revision document

This document includes the code and questions for our class revision of hypothesis tests and some other aspects of the course. The analysis in this example is for a proportion in one sample but you have also seen examples for difference of means, proportions and medians between two samples (Module 4).

Other information for preparing

Previous midterms and practice versions of past quizzes

Read through problem set sample solutions (see the Modules tab).

Feeling distressed? Here are some great resources if you need help.

Learning checklist

Getting started with R (Module 1)

- Understand how this course works, including how you will be assessed and what to do if you miss an assessment
- Recognize the importance of statistical reasoning and data science and provide examples
- Type and run R code in RStudio Cloud and produce a pdf document by knitting an RMarkdown document

Describing distributions; Calculating and interpreting numerical summaries (Module 2)

- Identify the type of a variable (continuous numerical / discrete numerical / nominal categorical / ordinal categorical)
- Describe the distribution of categorical and numerical variables based on data visualizations
- Create visualizations using `ggplot2` to explore the distributions of categorical and numerical variables
  - Centre, spread, shape
    - Think of the mean as the balance point of the data

Tidy data and data wrangling (Module 3)

- Explain why a dataset is not tidy
- Use `dplyr` functions presented in class to wrangle data to extract a subset of a larger dataset
  - Does the order make sense? Do you have what you need to run each line of code from the previous line?
- What is a tibble?
- Use `dplyr` functions presented in class to answer questions via summary tables or visualizations
- Describe the data frame resulting from a sequence of `dplyr` functions applied to a given data frame
- Make sure you're familiar with:
  - `case_when`
  - `mutate`
  - `filter`
  - `arrange`
  - `head`
  - `glimpse`
  - `select`
  - `summarise`
  - `group_by`
  - Pipes `%>%`
  - Assignment `<-`
  - `mean`
  - `median`
  - `sd`
  - `max`
  - `min`

Hypothesis tests (Module 4)

- Describe examples of how random outcomes can be simulated in R using sample() and without a computer (e.g., using coins).
- Explain what a sampling distribution of a statistic is and describe an example.
- Distinguish between the "real world" and the "theoretical world" (under the null hypothesis).
- Modify/write R code that uses for loops and the sample() function to simulate values of the test statistic under a null hypothesis.
- Use R to conduct a statistical test (i.e., hypothesis test) for one proportion.
- Assess evidence against an assumption (i.e., the null hypothesis) and write a conclusion for a hypothesis test based on a p-value.
- Distinguish between correct and incorrect conclusions of hypothesis tests.

- Distinguish between one-sample and two-sample hypothesis tests.
- Explain how test statistic values can be simulated under the null hypothesis of a two-sample test (both with and without a computer).
- Modify sample R code to simulate the sampling distribution of a test statistic under the null hypothesis of a two-sample test.
- Explain how sample() is used differently when simulating values for one-sample and two-sample tests.
- Modify sample R code to conduct a test to compare population parameters (e.g., medians, means, proportions, standard deviations) between two groups.
- Interpret the results of a two-sample hypothesis test by making conclusions based on the p-value and a significance level.
- Recognize correct and incorrect descriptions of "p-value".
- Distinguish between type 1 and type 2 errors in different contexts.
- The big picture for hypothesis testing:
  - One proportion: Is the proportion of the population *this* value or not?
  - Two groups: In the population, is the parameter for these two groups the same or not?
  - We simulate the null and then check if what we saw in the our sample seems pretty normal or really unusual. If our test statistic is not unusual (pretty common to see values like ours or more extreme) than we have no reason to suggest the null might be false. If our test statistic is really unusual if the null is true, that might give us a reason to 'reject' the null, or preferably, claim that we have evidence against it of a certain strength.
  - Make sure you have memorized/easily accessible the strength of evidence table! What is very strong, what is moderate.

Bootstrapping (Module 5)

- Distinguish between the distribution of a variable in the population or based on a sample and sampling distributions of statistics.
- Predict the effect of sample size (n) and number of repetitions on the centre, shape, and spread of the sampling distribution of sample means.
- Estimate sampling distributions of statistics by selecting many samples of the same size from the population or by drawing many bootstrap samples from the original sample.
- Explain the purpose of the bootstrap method and recognize applications where this method might be useful.
- Describe the steps required to obtain a bootstrap sampling distribution.
- Use R to compute bootstrap confidence intervals for parameters (e.g, μ, p).
- Recognize the connection between confidence levels and the widths of confidence interval estimates.
- Distinguish between correct and incorrect interpretations of confidence intervals.
- The big picture for bootstrapping:
  - Our estimate is our single best guess of the population parameter, but we know that while we are hopefully close, we are probably wrong just due to sampling variability.

  - We use bootstrapping to better understand the variability in our sample to help us make an educated guess about the variability in our population, and so create a plausible range of values for the parameter.
  - We use the language 'confident' because we are confident in our method. It will work X% of the time, though we're not sure if our particular time things are working.

Some cross module notes:

- Identifying sample/population
  - Be as specific as possible. Don't assume it is the same as *all* the data. Investigation specific. *Who* are you wanting to make claims about?
- We take a random sample as it should be representative of the population if every unit in the population has an equal chance of being included. We might get unlucky with our sample though, we don't necessarily know!
- Notes for both hypothesis testing and bootstrapping:
  - Make sure you know what you're doing with `sample()` or `sample_n()`
  - We're making sampling distributions in both cases but for different purposes

Practice questions

```
library(tidyverse)

knitr::opts_chunk$set(message = FALSE)
```

The `dpylr` package has some data about Star Wars characters. Let's assume it is a representative sample of all characters seen in Episodes 1 to 9.

Question 1

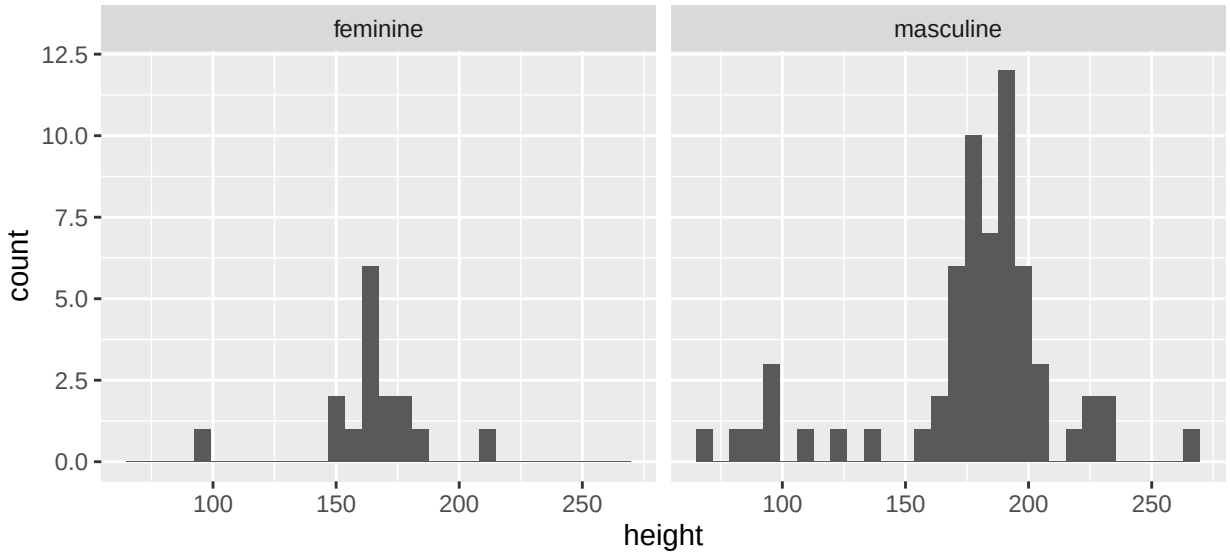
```
my_sw_data <- starwars %>%
  mutate(gender = ifelse(is.na(gender), "none", gender)) %>%
  select(name, gender, height, mass, species)
```

What does this code do? (Just the part starting with `my_sw_data`.)

A. Takes the `starwars` dataset, filters out missing values and selects levels of the name variable that are equal to "gender", "height", "mass" or "species".
B. Takes the `starwars` dataset, makes a new variable called gender with the levels "none" and gender, and then selects only the variables name, gender, height, mass and species.
C. Takes the `starwars` dataset, replaces any missing values for gender with "none" and and then selects only the variables name, gender, height, mass and species.

Question 2

```
my_sw_data %>%
  filter(!is.na(height)) %>%
  filter(gender %in% c("feminine", "masculine")) %>%
  ggplot(aes(height)) +
  geom_histogram(bins=30) +
  facet_wrap(~gender)
```



Which features should you talk about to compare the above distributions?

A. Pattern, strength, direction
B. Centre, spread, shape
C. Strength, mean, range

Question 3

name	gender_male	gender_female	height
Luke Skywalker	TRUE	FALSE	172
Finn	TRUE	FALSE	NA
Rey	FALSE	TRUE	NA

Which of the following statements is appropriate?

A. This version of the data is not tidy because there are missing values.
B. This version of the data is not tidy because one or more variable(s) does not have exactly one column.
C. Both A and B.
D. This version of the data is tidy.

Question 4

```
table1
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

```
table3
```

```
## # A tibble: 6 x 3
##   country      year rate
## * <chr>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

```
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

Which ONE of the following is true?

A. Only `table4a` is not tidy. There is more than one observation per row because of the multiple year columns.
B. Only `table3` is not tidy. Storing cases and population as a character vector means each value doesn't have its own cell.
C. Only `table1` is tidy.
D. All three tables are tidy.

I want to know a plausible range of values for the mean height of feminine characters in Star Wars Episodes 1 to 9.

```
my_sw_data %>%
  filter(!is.na(height)) %>%
  mutate(gender = ifelse(is.na(gender), "none", gender)) %>%
  select(name, gender, height, mass, species) %>%
  group_by(gender) %>%
  summarise(n=n(),
            mean_height=mean(height),
            median_height=median(height))
```

```
## # A tibble: 3 x 4
##   gender        n mean_height median_height
##   <chr>     <int>       <dbl>         <dbl>
## 1 feminine     16        165.          166.
## 2 masculine    62        177.          183
## 3 none          3        181.          183
```

```
fem_heights <- starwars %>%
  filter(gender == "feminine", !is.na(height))

glimpse(fem_heights)
```

```
## Rows: 16
## Columns: 14
## $ name       <chr> "Leia Organa", "Beru Whitesun lars", "Mon Mothma", "Shmi Sk~
## $ height     <int> 150, 165, 150, 163, 178, 184, 157, 170, 166, 165, 168, 213,~
## $ mass       <dbl> 49.0, 75.0, NA, NA, 55.0, 50.0, NA, 56.2, 50.0, NA, 55.0, N~
## $ hair_color <chr> "brown", "brown", "auburn", "black", "none", "none", "brown~
## $ skin_color <chr> "light", "light", "fair", "fair", "blue", "dark", "light", ~
## $ eye_color  <chr> "brown", "blue", "blue", "brown", "hazel", "blue", "brown",~
## $ birth_year <dbl> 19, 47, 48, 72, 48, NA, NA, 58, 40, NA, NA, NA, NA, NA, NA,~
## $ sex        <chr> "female", "female", "female", "female", "female", "female",~
## $ gender     <chr> "feminine", "feminine", "feminine", "feminine", "feminine",~
## $ homeworld  <chr> "Alderaan", "Tatooine", "Chandrila", "Tatooine", "Ryloth", ~
## $ species    <chr> "Human", "Human", "Human", "Human", "Twi'lek", "Tholothian"~
## $ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return~
## $ vehicles   <list> "Imperial Speeder Bike", <>, <>, <>, <>, <>, <>, <>, <>, <~
## $ starships  <list> <>, <>, <>, <>, <>, <>, <>, <>, <>, <>, <>, <>, <>, <>, <>~
```

Question 5

How many rows are in the `fem_heights` dataset?

Question 6

How many columns are in the `fem_heights` dataset?

Question 7

Which of the following are always true, if we do enough bootstrap resamples?

1. The mean of the bootstrap sampling distribution will be approximately the test statistic from our sample.
2. The statistic of interest calculated for the bootstrap sampling distribution will be exactly the same as the test statistic from our sample.
3. The mean of bootstrap sampling distribution will be approximately median from our sample.

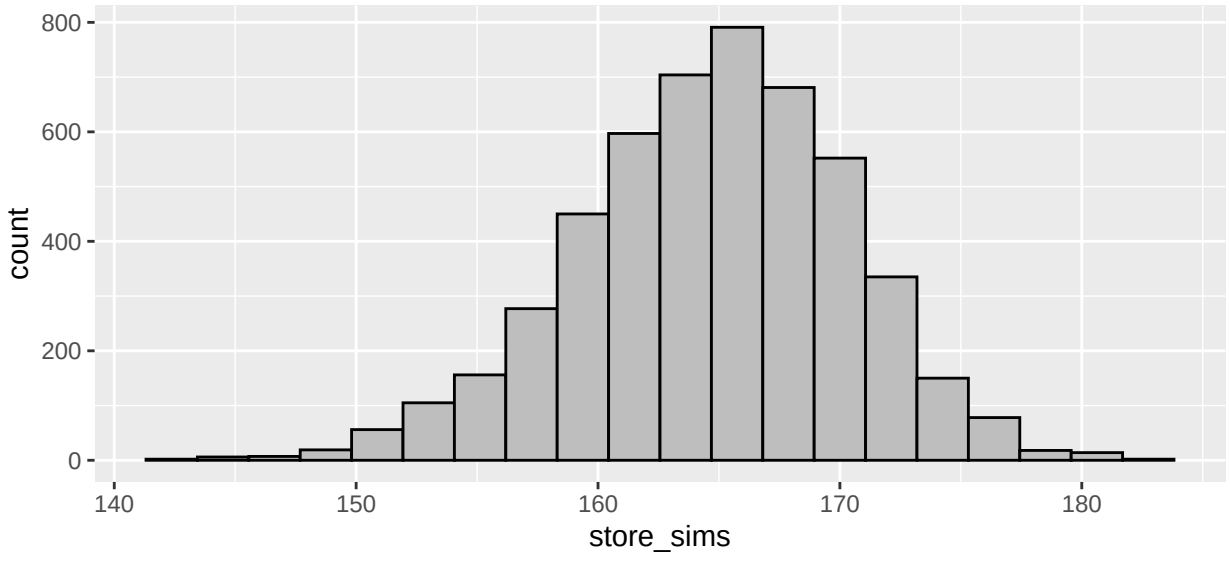
A. Only 1.
B. Only 1 and 2.
C. Only 3.
D. None of these statements.

```
set.seed(42)
repetitions <- 5000
store_sims <- rep(NA, repetitions)

for(i in 1:repetitions){
  store_sims[i] <- fem_heights %>%
    sample_n(size = nrow(fem_heights), replace=TRUE) %>%
    summarise(x = mean(height)) %>%
    as.numeric()
}

store_sims <- tibble(store_sims)

store_sims %>%
  ggplot(aes(x = store_sims)) +
  geom_histogram(bins = 20, color = "black", fill = "grey")
```



Question 8

How might we expect the histogram above to change if we had a sample of feminine characters in Episodes 1 to 9 that was twice as large?

A. The mean should be roughly the same, but it might be less symmetrical and more spread out.
B. The mean should be roughly the same, but it might be more symmetrical and less spread out.
C. The mean, skew and spread should decrease.
D. The mean, skew and spread should increase.

```
quantile(store_sims$store_sims,
         probs = c(0.005, 0.01, 0.025, 0.05, 0.2, 0.25, 0.5,
                   0.75, 0.8, 0.95, 0.975, 0.99, 0.995))
```

```
##     0.5%       1%     2.5%       5%      20%      25%      50%      75%
## 149.3103 150.5619 152.8750 154.7500 160.1250 161.1875 165.0000 168.6875
##      80%      95%    97.5%      99%    99.5%
## 169.5000 173.3125 174.8141 176.8756 178.3763
```

Question 9

Which of these is the 95% CI for this analysis?

A. (150.6, 176.9)
B. (152.9, 174.8)
C. (154.8, 173.3)
D. (149.3, 173.3)

Question 10

Which of these is a correct interpretation of the CI (149.3, 178.4)?

A. We are 99% certain that each feminine character in Episodes 1 to 9 was between 149.3 and 178.4 cm tall.
B. We expect 99% of feminine characters in Episodes 1 to 9 to be between 149.3 and 178.4 cm tall.
D. We are 99% confident that the true mean height of feminine characters in Episodes 1 to 9 is between 149.3 and 178.4 cm.

Question 11

```
set.seed(1)
opt1 <- fem_heights %>%
    sample_n(size = nrow(fem_heights), replace=FALSE) %>%
    summarise(x = mean(height)) %>%
    as.numeric()
opt1
```

```
## [1] 164.6875
```

```
set.seed(2)
opt2 <- fem_heights %>%
    sample_n(size = nrow(fem_heights), replace=TRUE) %>%
    summarise(x = mean(height)) %>%
    as.numeric()
opt2
```

```
## [1] 169.0625
```

Which of the following statements about the above code is correct?

A. The results are different ONLY because we haven't used the same set.seed() value.
B. The results are different because one sets replace=FALSE and so samples exactly the same observations as in the original sample, while the other set replace=TRUE and so some of the values can be duplicated or not selected into the sample at all.
C. The results are different because the mean of the bootstrap sampling distribution is the same as the mean height from our sample.