Problem Set 09

WRITE YOUR NAME HERE

2018-12-12

Contents

Collaboration

Please indicate who you collaborated with on this problem set:

Background

First load the necessary packages:

```
library(tidyverse)
library(infer)
```

For this Problem Set you will work with some grade-point-average (GPA) data for college freshman. The following will read in the data:

```
sat_gpa <- read_csv("https://rudeboybert.github.io/SDS220/static/PS/sat_gpa.csv")
```

Each row or case in this data frame is a student. The data includes:

- the (binary) gender of each student
- the math, verbal and total SAT scores for each student
- the GPA range of each student in high school (categorized as “low” or “high”)
- the GPA of each student their first year of college on a numeric scale.

We will use hypothesis testing to answer the following questions:

1. Is there a difference in male and female freshman GPAs?
2. Is there a difference in total SAT score for students with a “low” and “high” high-school GPA?

Note, if you get stuck as you are working through this, it will be helpful to review Chapter 9 in ModernDive.

Question 1: Gender differences in first-year GPA?

Exploratory data analysis

Q1.a)

Calculate the mean GPA score for each gender, using the `group_by` and `summarize` commands from the `dplyr` package.

Questions:

1. What is the difference in sample mean GPA's?
2. Make a guess: is this difference statistically significant?

Answers:

1.

2.

Q1.b)

Generate a data visualization that displays the GPAs of the two groups. Be sure to include a title and label your axes.

Stating a null hypothesis

We will now test the null hypothesis that there's no difference in population mean GPA between the genders at the population level. We can write this out in mathematical notation

$$H_0 : \mu_{male} = \mu_{female}$$
$$\text{vs } H_A : \mu_{male} \neq \mu_{female}$$

or expressed differently, that the difference is 0 or not:

$$H_0 : \mu_{male} - \mu_{female} = 0$$
$$\text{vs } H_A : \mu_{male} - \mu_{female} \neq 0$$

Testing the hypothesis

Here's how we use the `infer` package to conduct this hypothesis test:

Step 1: Calculate the observed difference

Note that the order we choose does not matter here (female then male)...but since we used `order = c("Female", "Male")` here, we should do the same in subsequent calculations!

```
obs_diff_gpa_sex <- sat_gpa %>%
  specify(gpa_fy ~ sex) %>%
  calculate(stat = "diff in means", order = c("Female", "Male"))

obs_diff_gpa_sex
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.149
```

Note that this is the difference in the group means we calculated earlier!

```r
2.544587 - 2.396066
```

```
## [1] 0.148521
```

Step 2. Generate the null distribution of δ

This step involves generating simulated values *as if* we lived in a world where there's no difference between the two groups. Going back to the idea of permutation, and tactile sampling, this is akin to shuffling the GPA scores between male and female labels (i.e. removing the structure to the data) just as we could have done with index cards.

```r
gpas_in_null_world <- sat_gpa %>%
  specify(gpa_fy ~ sex) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = 'permute')
```

Q1.c)

Question:

1. What was the size of the "shuffled" (permuted) sample in each run?
2. How many times did we "shuffle" (permute) the sample? How many rows are in the `gpas_in_null_world` data frame?

Answer:

1.

2.

Step 3. Calculate the differences between male and females under the null

The following calculates the differences in mean GPA for males and females for "shuffled" (permuted) data.

```r
gpa_diff_under_null <- gpas_in_null_world %>%
  calculate(stat = "diff in means", order = c("Female", "Male"))

gpa_diff_under_null %>%
  slice(1:5)
```

```
## # A tibble: 5 x 2
##   replicate      stat
##       <int>     <dbl>
## 1         1 -0.0225
## 2         2  0.00445
## 3         3  0.0205
## 4         4 -0.000552
## 5         5 -0.00452
```

Q1.d)

Question:

1. How many rows are in this data frame? Why?

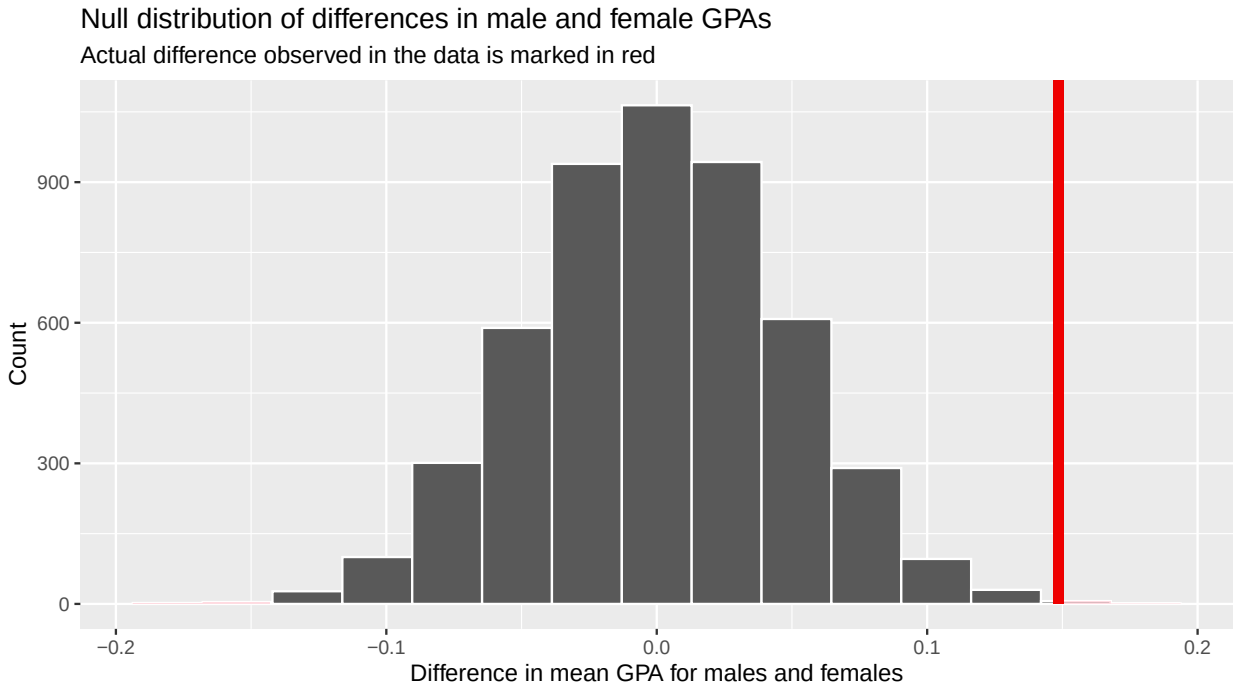
Answer:

1.

Step 4. Visualize how the observed difference compares to the null distribution of δ

The following plots the δ values we calculated for each of the different "shuffled" replicates. This is the null distribution of δ. The red line shows the observed difference between male and female scores in the data (-0.1485209) from step 1.

```
visualize(gpa_diff_under_null) +
  shade_p_value(obs_stat = obs_diff_gpa_sex, direction = "both") +
  labs(x = "Difference in mean GPA for males and females", y = "Count",
       title = "Null distribution of differences in male and female GPAs",
       subtitle = "Actual difference observed in the data is marked in red"
       )
```



Note that zero is the center of this null distribution. The null hypothesis is that there is no difference between males and females in GPA score. In the permutations, zero was the most common difference, because observed GPA values were re-assigned to males and females **at random**. Differences as large as ~ 0.1 and -0.1 occurred, but much less frequently, because they are just not as likely when structure is removed from the data.

Step 5: Calculate a p-value

```
gpa_diff_under_null %>%
  get_pvalue(obs_stat = obs_diff_gpa_sex, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.002
```

This result indicates that there is a 0.1% chance (very low) chance that we would see a difference of 0.15 in GPA scores between males and females (or a bigger difference) if in fact there was truly no difference between the sexes in GPA scores in the population.

Q1.e)

Fill in the blanks below to write up the results & conclusions for this test:

The mean GPA scores for females in our sample ($\bar{x}$ = _____) was greater than that of males ($\bar{x}$ = _____). This difference **(was/was not)**_ statistically significant at $\alpha = 0.05$, (p = _____). Given this I **(would/would not)** reject the Null hypothesis and conclude that _____ have higher GPAs than _____ at the population level.

Step 6: Calculate a confidence interval for the difference

The following will allow us to calculate a 95% confidence interval for the difference between mean GPA scores for males and females.

```
ci_diff_gpa_means <- sat_gpa %>%
  specify(gpa_fy ~ sex) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("Female", "Male")) %>%
  get_confidence_interval(level = 0.95)
ci_diff_gpa_means
```

```
## # A tibble: 1 x 2
##   `2.5%` `97.5%`
##    <dbl>   <dbl>
## 1 0.0565   0.235
```

Complete all the above tasks with a t-test

Note that all the above steps can be done with one line of code **if a slew of assumptions** like normality and equal variance of the groups are met.

```
t.test(gpa_fy ~ sex, var.equal = TRUE, data = sat_gpa)
```

```
##
##  Two Sample t-test
##
## data:  gpa_fy by sex
## t = 3.1828, df = 998, p-value = 0.001504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05695029 0.24009148
## sample estimates:
## mean in group Female   mean in group Male
##             2.544587             2.396066
```

Question 2: Relationship between high-school GPA category and Total SAT score?

For this analysis `sat_total` is the outcome variable, and `gpa_hs` is the predictor variable, with two levels “low” and “high”.

Exploratory data analysis

We can first calculate the mean total SAT score for each group (i.e students with a low and high GPA), using the `group_by` and `summarize` commands from the `dplyr` package.

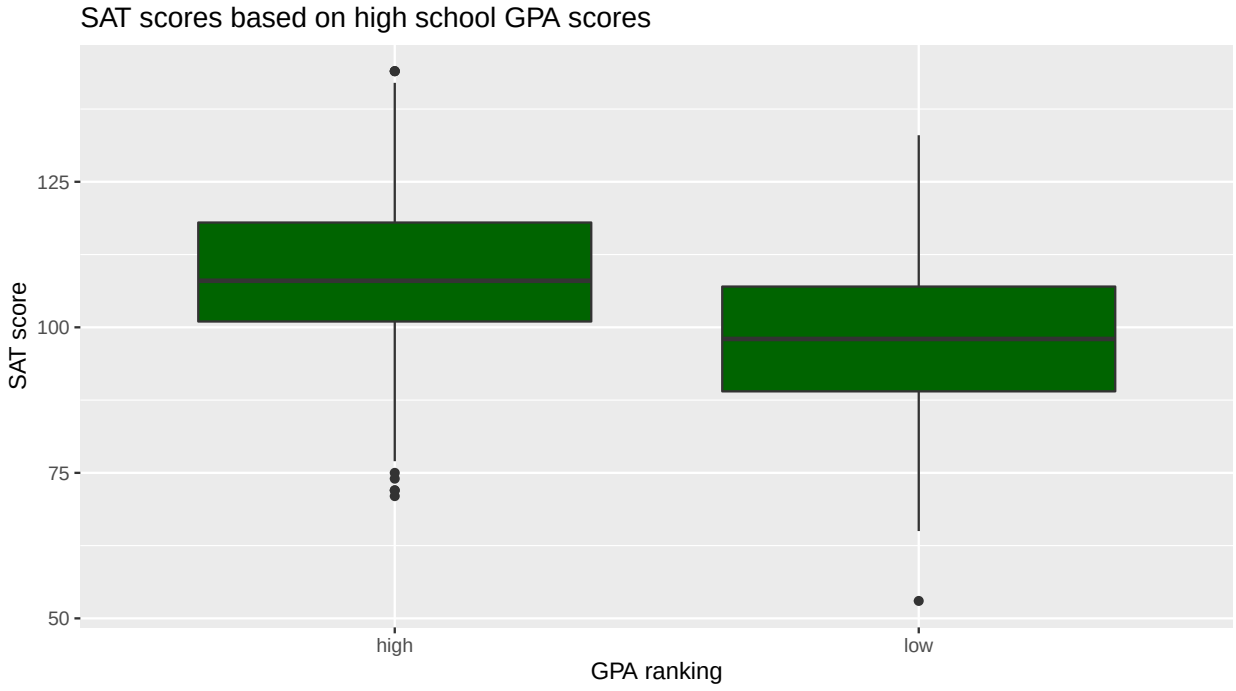
```
avg_sat_gpa <- sat_gpa %>%
  group_by(gpa_hs) %>%
  summarize(sat_total = mean(sat_total))

avg_sat_gpa
```

```
## # A tibble: 2 x 2
```

```
##   gpa_hs sat_total
##   <chr>      <dbl>
## 1 high       109.
## 2 low         98.2
```

We will next generate a data visualization that displays the total SAT scores of the two groups. Be sure to include a title and label your axes.

```
ggplot(sat_gpa, aes(x = gpa_hs, y = sat_total)) +
  geom_boxplot(fill = "darkgreen") +
  labs(title = "SAT scores based on high school GPA scores",
       x = "GPA ranking", y = "SAT score")
```



Stating a null hypothesis

Q2.a)

State the null hypothesis that you are testing (using either words or symbols)

Answer:

1.

Testing the hypothesis

Q2.b)

Calculate the observed difference between the mean total SAT scores of the low and high GPA high-school students.

```
# you finish this code....

# obs_diff_sat_hs_gpa <- sat_gpa %>%
```

Q2.c)

Generate the null distribution of δ. Here you need to generate simulated values *as if* we lived in a world where there's no difference in SAT scores between high school students with low and high GPAs.

```
# you finish this code....

# sat_in_null_world <- sat_gpa
```

Q2.d)

Calculate the differences in mean SAT scores between students with low and high GPA scores under the Null. Note…you should use whatever order you chose above…i.e. `order = c("low", "high")` or `order = c("high", "low")`.

```
# you finish this code....

# sat_diff_under_null <-
```

Q2.e)

Visualize how the observed difference compares to the null distribution of δ. Generate a histogram of the null distribution, with a vertical red line showing the observed difference in SAT scores between high school students with a high and low GPA.

```
# you finish this code....

# sat_diff_under_null %>%
```

Q2.f)

Calculate a p-value

Answer:

1.

Q2.g)

Write up the results & conclusions for this hypothesis test. Note, p-values less than 0.001 are often reported as p < 0.001.

Answer:

1.

Q2.h)

Calculate a confidence interval for the difference in total SAT scores for students with high and low high-school GPA scores. Note…you should use whatever order you chose above…i.e. `order = c("low", "high")` or `order = c("high", "low")`.

```
# you finish this code....

# ci_diff_sat_means <- sat_gpa %>%
```

Q2.i)

Use a t-test to test the null hypothesis that total SAT scores do not differ between students with high and low high school GPA scores at the population level.