GPA Hypothesis Testing

Sean Johnson

2020-09-25

Background

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

Dataset of 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.

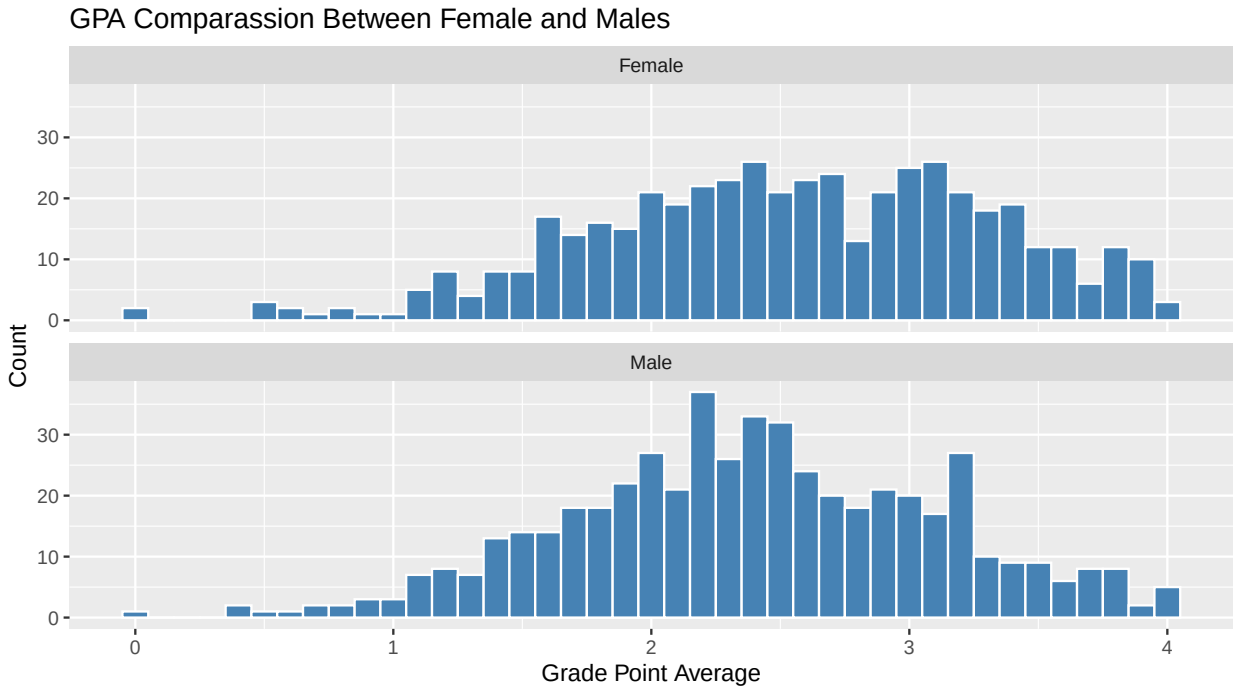
Gender differences in first-year GPA

Exploratory data analysis

```
sat_gpa %>% group_by(sex) %>%
  summarize(mean_gpa = mean(gpa_fy))
```

```
## # A tibble: 2 x 2
##   sex    mean_gpa
##   <chr>     <dbl>
## 1 Female     2.54
## 2 Male       2.40
```

```
ggplot(data = sat_gpa, aes(x = gpa_fy)) +
  geom_histogram(binwidth = .1, color = "white", fill = "steelblue") +
  facet_wrap(~ sex, ncol = 1) +
  labs(x = "Grade Point Average", y = "Count",
  title = "GPA Comparassion Between Female and Males")
```



Null hypothesis

Null hypothesis that there's no difference in population mean GPA between the genders at the population level. The mathematical notation is the following:

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

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

The observed difference

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

obs_diff_gpa_sex
```

```
## Response: gpa_fy (numeric)
## Explanatory: sex (factor)
## # A tibble: 1 x 1
##    stat
```

```
##   <dbl>
## 1 0.149
```

Note that this is the difference in the group means

Generate the null distribution of δ

```
gpas_in_null_world <- sat_gpa %>%
  specify(gpa_fy ~ sex) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = 'permute')
```

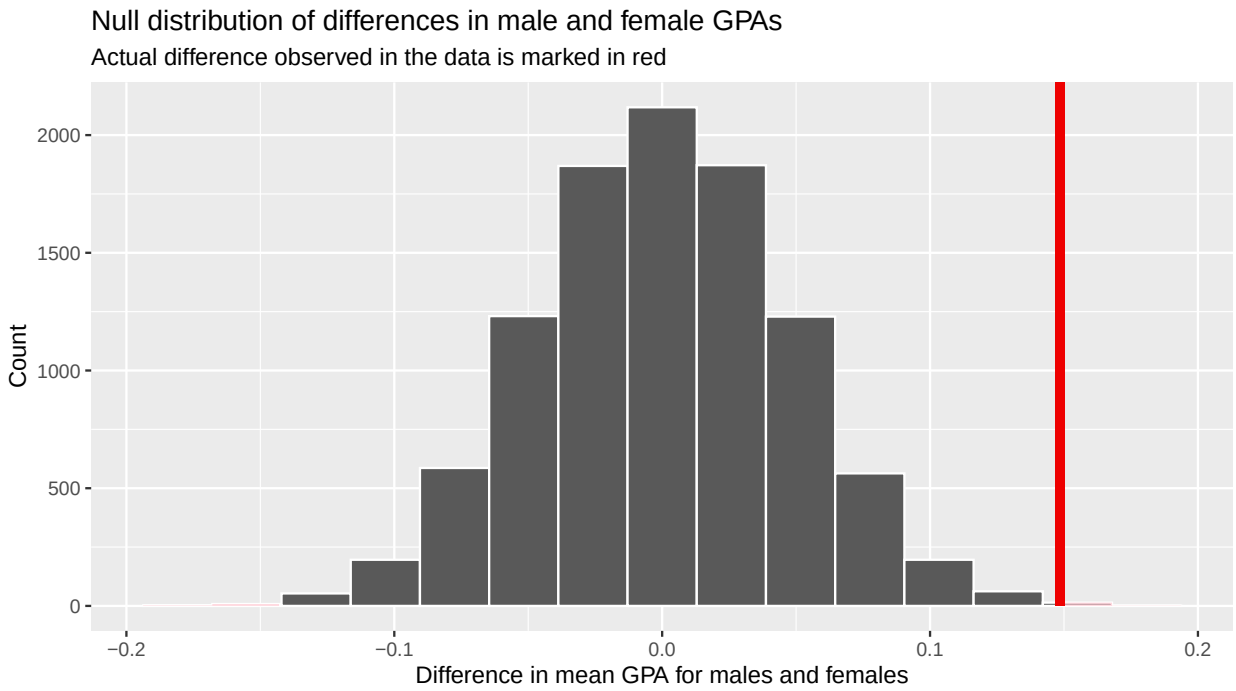
The differences between male and females under the null

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

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

```
## Response: gpa_fy (numeric)
## Explanatory: sex (factor)
## Null Hypothesis: independence
## # 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
```

Visualization of how the observed difference compares to the null distribution of δ

```
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.

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  0.0018
```

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.

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
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0579    0.241
```

Repeating the above 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 between group Female and group Male is not equal to
## 95 percent confidence interval:
##  0.05695029 0.24009148
## sample estimates:
## mean in group Female   mean in group Male
##             2.544587             2.396066
```

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

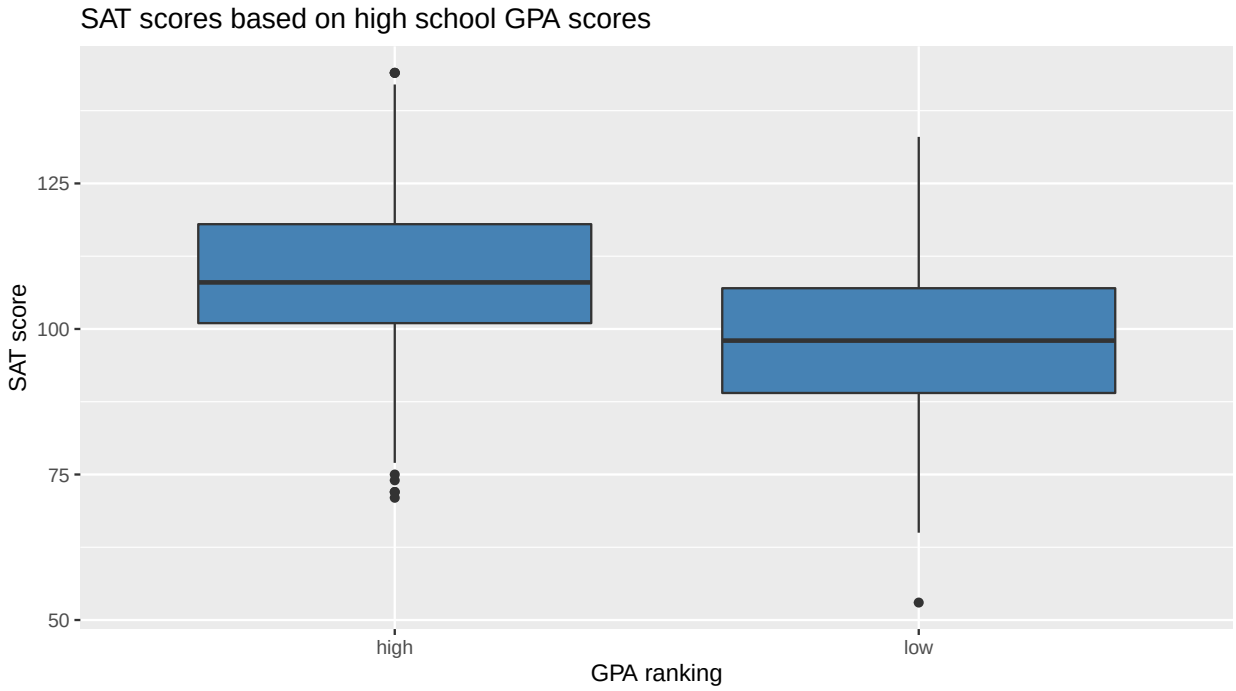
For this analysis `sat_total` is the outcome variable, and `gpa_hs` is the predictor variable, with two levels "low" and "high".

Exploratory data analysis

```
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
```

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



Null hypothesis

$$H_0 : \mu_{high} = \mu_{low}$$
$$\text{vs}$$
$$H_A : \mu_{high} \neq \mu_{low}$$

Testing the hypothesis

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

```
obs_diff_sat_hs_gpa <- sat_gpa %>%
  specify(sat_total ~ gpa_hs) %>%
  calculate(stat = "diff in means", order = c("high", "low"))

obs_diff_sat_hs_gpa
```

```
## Response: sat_total (numeric)
## Explanatory: gpa_hs (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  10.4
```

```r
sat_in_null_world <- sat_gpa %>%
  specify(sat_total ~ gpa_hs) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = 'permute')
```

Generate the null distribution of δ.

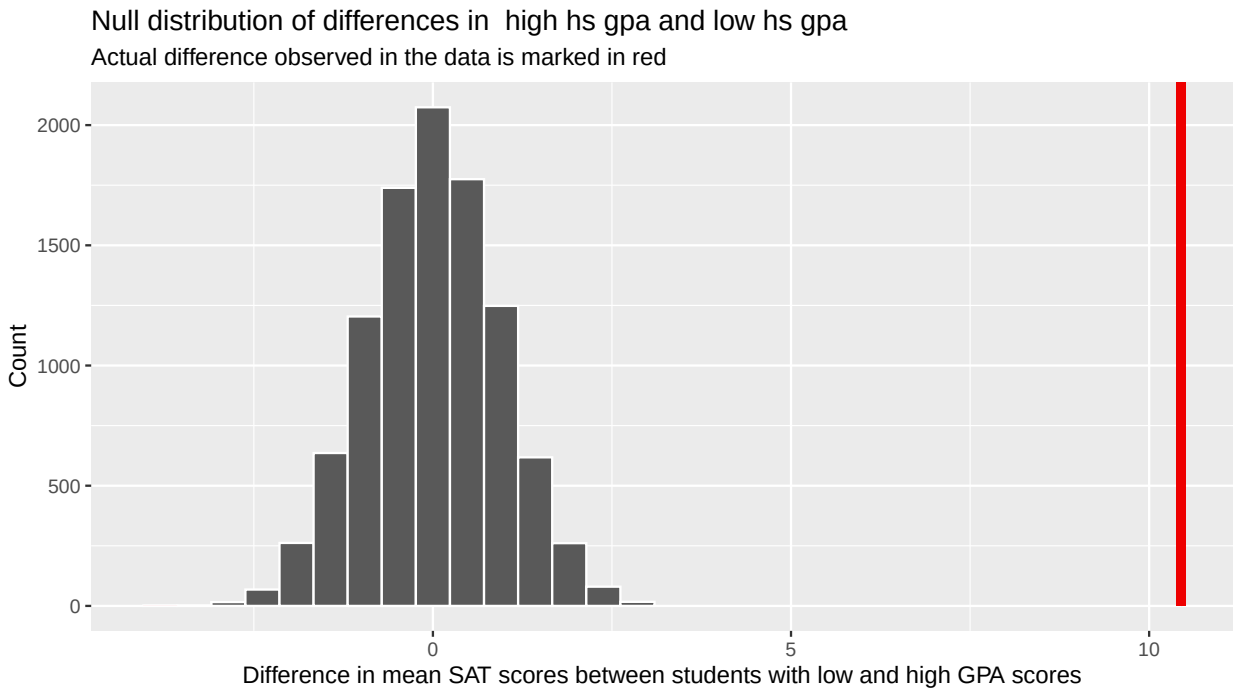
```r
sat_diff_under_null <- sat_in_null_world %>%
  calculate(stat = "diff in means", order = c("high", "low"))

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

The differences in mean SAT scores between students with low and high GPA scores under the Null.

```
## Response: sat_total (numeric)
## Explanatory: gpa_hs (factor)
## Null Hypothesis: independence
## # A tibble: 5 x 2
##   replicate   stat
##       <int>  <dbl>
## 1         1 -0.895
## 2         2  0.398
## 3         3  0.602
## 4         4 -0.118
## 5         5 -0.891
```

Visualization of how the observed difference compares to the null distribution of δ. Generating 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.

```r
visualize(sat_diff_under_null)+
  shade_p_value(obs_stat = obs_diff_sat_hs_gpa, direction = "both") +
  labs(x = "Difference in mean SAT scores between students with low and high GPA scores", y = "Count",
       title = "Null distribution of differences in  high hs gpa and low hs gpa",
       subtitle = "Actual difference observed in the data is marked in red"
       )
```




```
sat_diff_under_null %>%
  get_pvalue(obs_stat = obs_diff_sat_hs_gpa, direction = "both")
```

Calculate a p-value

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Results & conclusions for this hypothesis test. Note, p-values less than 0.001 will be reported as p < 0.001.

The mean SAT scores for students with high GPA scores in our sample ($\bar{x} = 108.67828$) was greater than that of students with low GPA scores ($\bar{x} = 98.23047$). This difference was statistically significant at $\alpha = 0.05$, ($p < 0.001$). Given this I would reject the Null hypothesis and conclude that high-gpa students have higher SATs than low-gpa students at the population level.

```
ci_diff_sat_means <- sat_gpa %>%
  specify(sat_total ~ gpa_hs) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("high", "low")) %>%
  get_confidence_interval(level = 0.95)
ci_diff_sat_means
```

Confidence interval for the difference in total SAT scores for students with high and low high-school GPA scores.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     8.76     12.1
```

```
t.test(sat_total ~ gpa_hs, var.equal = TRUE, data = sat_gpa)
```

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.

```
##
##  Two Sample t-test
##
## data:  sat_total by gpa_hs
## t = 12.413, df = 998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group high and group low is not equal to 0
## 95 percent confidence interval:
##   8.79614 12.09948
## sample estimates:
## mean in group high  mean in group low
##          108.67828           98.23047
```