HW 04 - Simulation-based Inference

Due: Wednesday, Apr 08 at 11:59pm EST

Nathan Kim

4/9

Packages

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

Data

```
roulette <- read_csv("data/roulette.csv")
```

Set seed

```
set.seed(5768952)
```

Exercises

Exercise 1

Describe the simulation process for testing for a single population standard deviation. Suppose the research question is asking whether the standard deviation of IQ scores is less than 10, and the observed sample standard deviation is 7.5.

First, define the hypotheses based on the given research question. Because the null hypothesis states nothing unusual is occurring, the null hypothesis would be "the standard deviation of IQ scores is 10" whereas the alternative hypothesis would be "the standard deviation of IQ scores is less than 10." Second, bootstrap the data. In this case, a random sample would be taken with replacement from the original sample standard deviation (7.5) the same as the original sample size of 100. Then calculate the standard deviation of the bootstrap sample. Execute step two a total of 15,000 times to create a bootstrap distribution. Third, shift the bootstrap distribution to be centered at the null value by adding 2.5 to each bootstrap standard deviation. Fourth, calculate the p-value as the proportion of simulations that yield a sample standard deviation at least as extreme as the observed sample standard deviation. Finally, form a conclusion based on the findings. If the p-value is greater than the set significance level, we would fail to reject the null hypothesis due to insufficient evidence while if the p-value is less than the set significance level, we would reject the null hypothesis due to the results being statistically significant.

Exercise 2

Describe the simulation process for testing for a single population proportion. Suppose the research question is asking whether the proportion of successes is majority, where we have that the observed sample proportion of success is 0.52.

First, define the hypotheses based on the given research question. Because the null hypothesis is the one that states nothing unusual is occurring, based on the research question, the null hypothesis would be "the population proportion of success is equal tot that of failure" while the alternative would be "the population proportion of success is majority." Second, bootstrap the data. In this case, a random sample would be taken with replacement from the original sample standard deviation (0.52) the same as the original sample size of 100. Then calculate the proportion of success of the bootstrap sample. Execute step two a total of 15,000 times to create a bootstrap distribution. Third, shift the bootstrap distribution to be centered at the null value by subtracting 0.02 from each bootstrap proportion of success. Fourth, calculate the p-value as the proportion of simulations that yield a sample proportion of success at least as extreme as the observed sample proportion of success. Finally, form a conclusion based on the findings. If the p-value is greater than the significance level, we would fail to reject the null hypothesis due to insufficient evidence while if the p-value is less than significance level, we would reject the null hypothesis due to the results being statistically significant.

Exercise 3

Describe the simulation process for creating a 95% confidence interval for the population intercept in a simple linear regression model. Assume the population model is of the form y = B0 + B1x.

First, implement bootstrapping by taking a random sample with replacement from the original sample the same size as the original sample, 100. Second, create a regression model based on the data from the new sample (size 100) to find the intercept. Execute steps one and two a total of 15,000 times to create a bootstrap distribution for the population intercept. Finally, create a 95% confidence interval for the population intercept by taking the intervals containing the middle 95% of the bootstrap distribution.

Exercise 4

The sample size Gallup took in the 2016-2017 study was 337,690. We know this because the website says "Results are based on . . . 337,690 interviews conducted Jan. 2, 2016, through Dec. 30, 2017, as a part of the Gallup-Sharecare Well-Being Index, with a random sample of adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia."

Exercise 5

The quantities 11.5% and 10.8% represent the percentage of U.S. adults surveyed who report having been diagnosed with the disease over that time. 11.5% of U.S. adults surveyed during 2016-2017 report having been diagnosed during that survey period. 10.8% U.S. adults surveyed during the 2014-2015 report having been diagnosed during that survey period.

Exercise 6

Proportion of adult individuals with diabetes in 2016-2017(point estimate): 11.5 Sampling margin of error: 0.2 95 CI given by point estimate plus/minus sampling margin of error: 11.3 and 11.7 (11.3 (11.5 - 0.2) and 11.7 from (11.5 + 0.2)) 95% of all intervals (11.3 –> 11.7) constructed in this manner include the true population parameter, in other words the true proportion of adult individuals diagnosed with diabetes for the entire U.S. adult population in 2016-2017.

Exercise 7

Proportion of adults in Alaska with diabetes in 2016-2017 (point estimate): 8.4 Sampling margin of error: 3.5 95 CI given by point estimate plus/minus sampling margin of error: 4.9 and 11.9. (4.9 (8.4 - 3.5) and 11.9 from (8.4 + 3.5)) 95% of all intervals (4.9 –> 11.9) constructed in this manner include the true population parameter, in other words the true proportion of adult individuals diagnosed with diabetes for the entire Alaskan adult population in 2016-2017.

Exercise 8

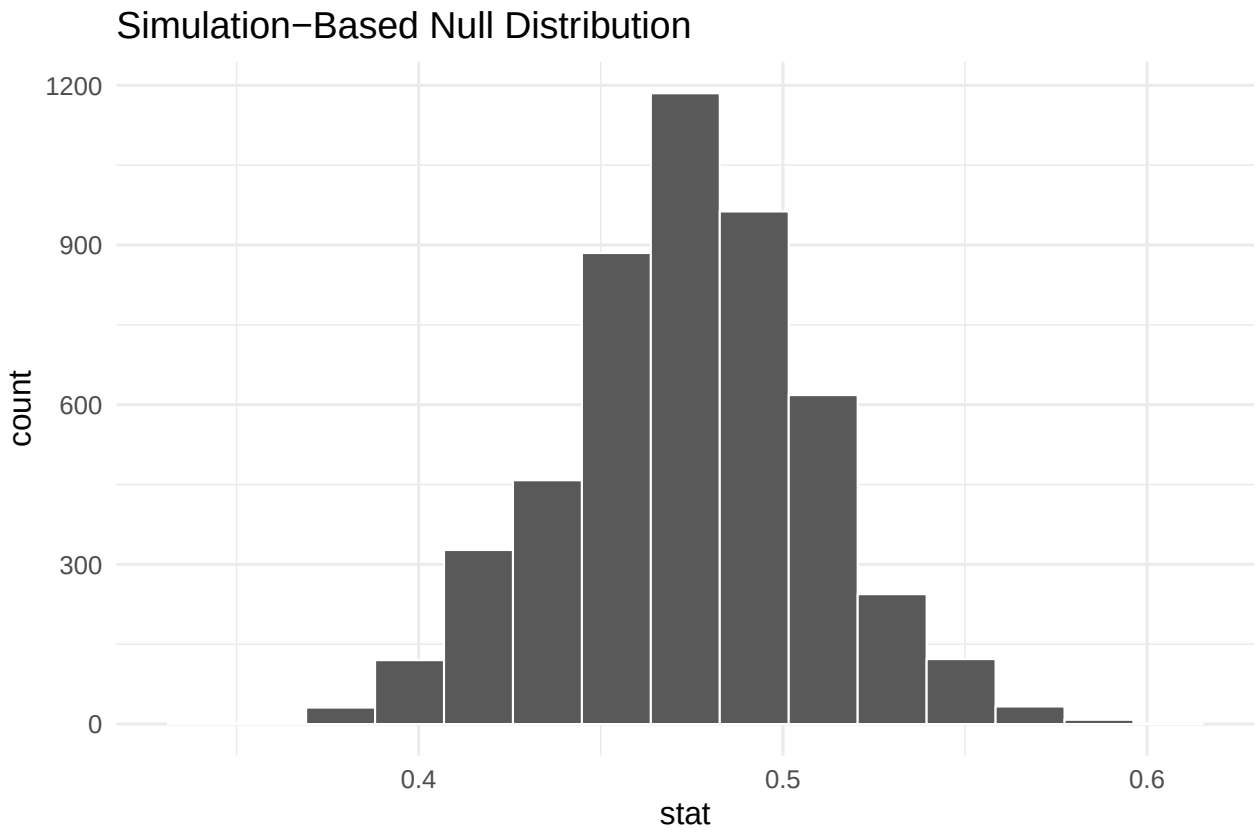
1. The null hypothesis is that the probability of the ball landing on red is exactly how the wheel is laid out (18 red slots out of 38 total, so 18/38). The alternate hypothesis is that the probability of the ball landing on red is not 18/38 and thus biased with regards to the ball landing on red. H0 : p = 0.47368 HA : p = 0.47368
2. Plotting simulated null distribution based on 5,000 replications.

```
other <- c("black", "green")

roulette <- roulette %>%
  mutate(spin = if_else(spin %in% other, "other", "red"))
```

```
sample_stat <- roulette %>%
  count(spin) %>%
  mutate(prop = n/sum(n)) %>%
  filter(spin == "red") %>%
  select(prop)
sample_stat
```

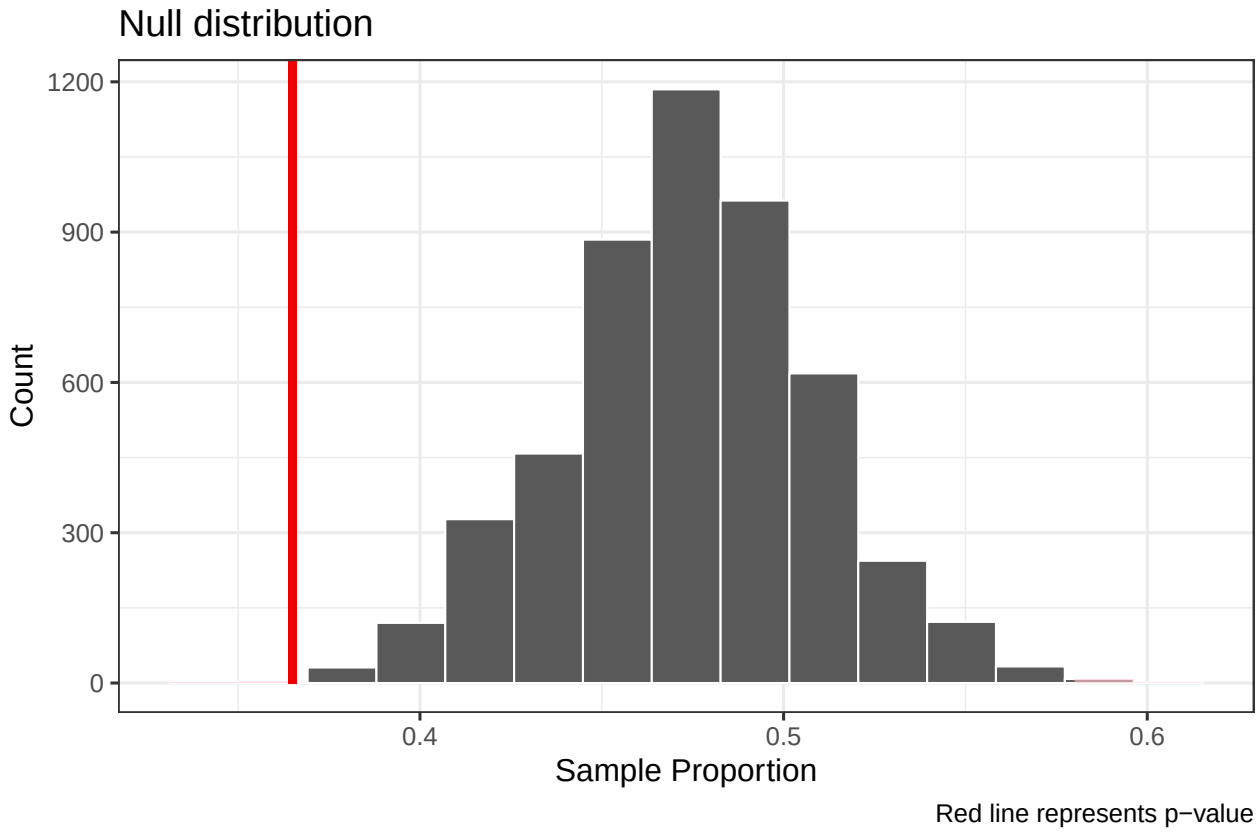
```
# A tibble: 1 x 1
   prop
  <dbl>
1 0.365
```

```
null_dist <- roulette %>%
  specify(response = spin, success = "red") %>%
  hypothesize(null = "point", p = c("red" = 0.47368, "other" = 0.52632)) %>%
  generate(reps = 5000, type = "simulate") %>%
  calculate(stat = "prop")

  visualize(null_dist) + theme_minimal(base_size = 16)
```



3. Compute and display the p-value

```
visualize(null_dist) +
  shade_p_value(obs_stat = sample_stat, direction = "two_sided") +
  theme_minimal(base_size = 16) +
  labs(title = "Null distribution", x = "Sample Proportion", y = "Count",
  caption = "Red line represents p-value") +
  theme_bw(base_size = 16)
```




```
get_p_value(null_dist, obs_stat = sample_stat, direction = "both")
```

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

4. Conclusion: Our p_value of 0.002 is greater than our 0.001 significance level. As a result, we have insufficient evidence to reject the null hypothesis. In other words, we have insufficient evidence to reject the idea that the wheel is NOT biased.

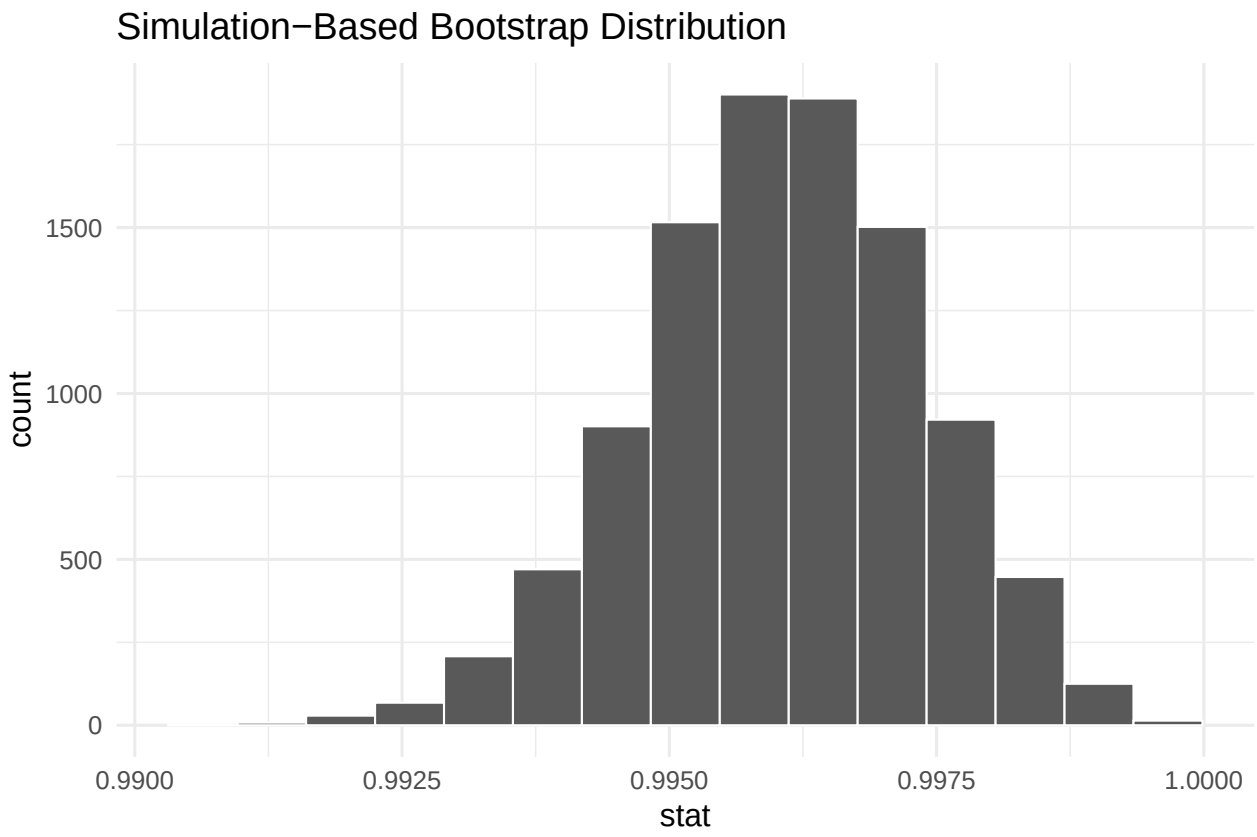
Exercise 9

1. Necessary Assumptions In order to construct a 95% confidence interval the assumptions and conditions of the central limit theorem must be met in order to use the normal model.

We will need to assume independence, that the sample observations are independent, random, and that if sampling without replacement, the sample size is less than 10% of the population size. We also need to assume that the sample is large enough.

2. Plot the Simulated Distribution Based on 10,000 replications

```
simul_dist <- women %>%
  specify(explanatory = height, response = weight) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "correlation")

visualize(simul_dist) + theme_minimal(base_size = 16)
```

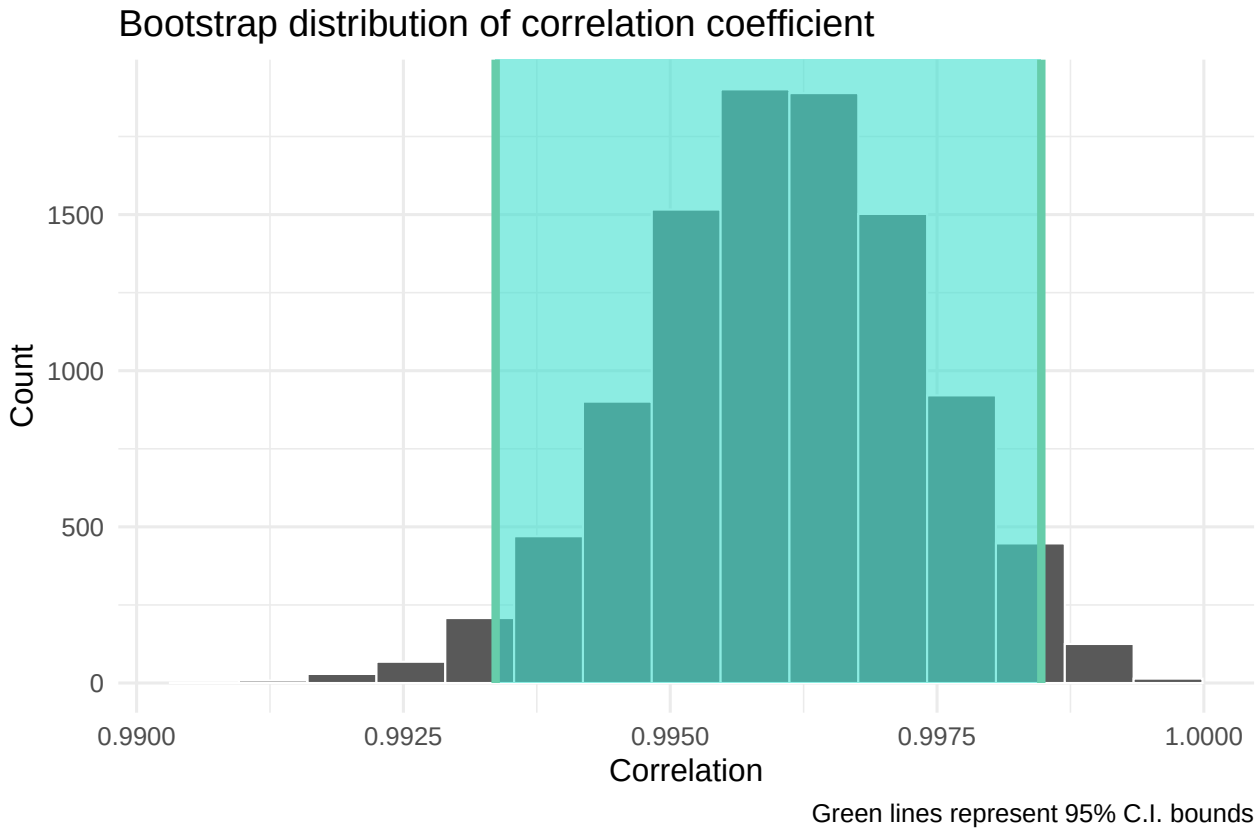


3. Compute and Display 95% CI

```
 percentile_ci <- get_ci(simul_dist, level = .95)
percentile_ci
```

```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1  0.993   0.998
```

```
visualize(simul_dist) +
  shade_confidence_interval(endpoints = percentile_ci) +
  labs(title = "Bootstrap distribution of correlation coefficient",
       x = "Correlation", y = "Count",
       caption = "Green lines represent 95% C.I. bounds") +
  theme_minimal(base_size = 16)
```



4. Interpretation: We are 95% confident that the true population correlation coefficient between the height and weight of all American women (aged 30-39) is between 0.99336 and 0.99848.