Lab 08 - Simulation-based inference

Due: Friday, Apr 10 at 11:59pm EST

Team gitData - Avanti Shah, Nathan Kim, Blossom Mojekwu

Packages

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

Data

```
data(ncbirths)
```

Setting seed

```
set.seed(71189752)
```

Exercises

Exercise 1

The null hypothesis is that the true average weight of Caucasian babies in NC is equal to that found by the 1995 study (7.43 pounds). On the other hand, the alternative hypothesis is that the true average weight of Caucasian babies in NC is less than or greater than that found by the 1995 study, i.e. it has changed from 7.43 pounds.

Exercise 2

H0: μ `= 7.43`

H1: μ `!= 7.43` (i.e., μ `> 7.43`, μ `< 7.43`)

where μ is the true average weight of Caucasian babies born in North Carolina

Exercise 3

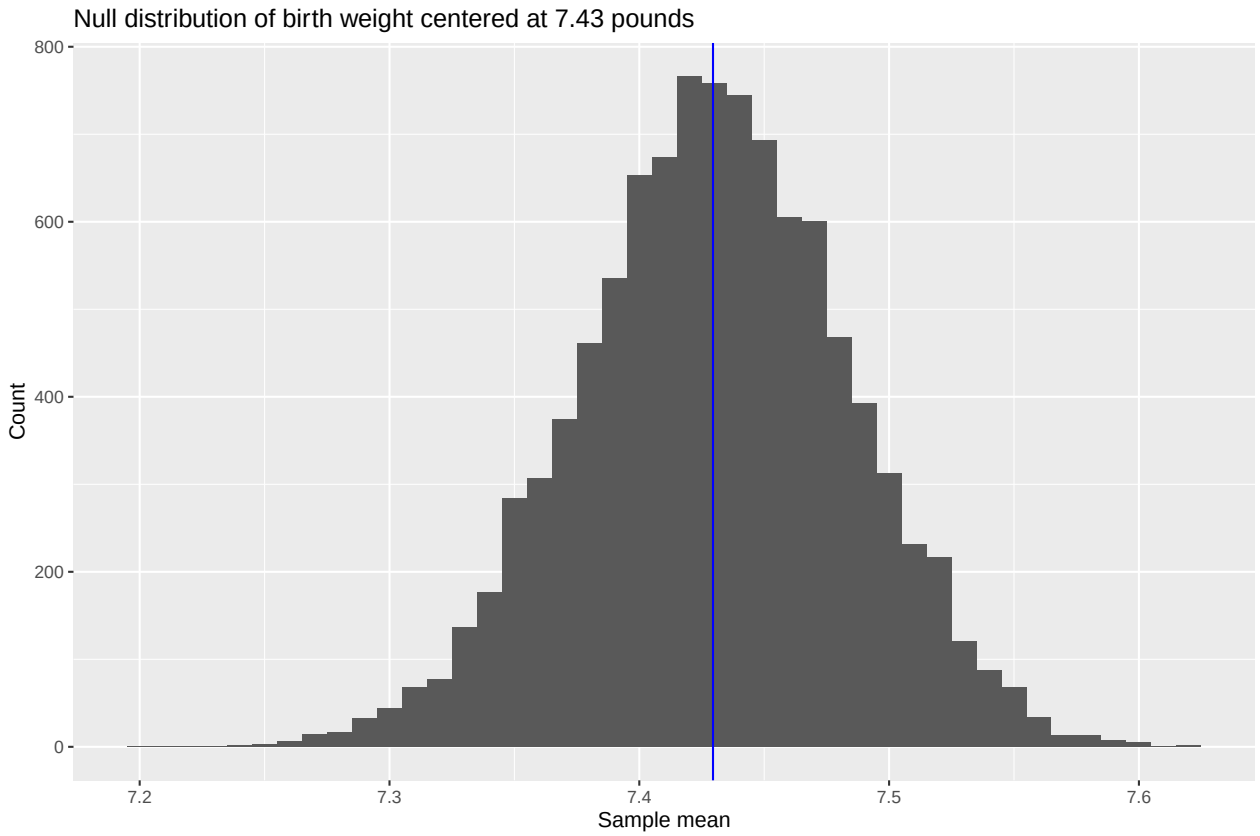
```
ex_3_sample <- ncbirths %>%
  filter(whitemom == "white")

ex_3_sample %>%
  summarise(mean = mean(weight))
```

```
      mean
1 7.250462
```

```
ex_3_weight_boot <- ex_3_sample %>%
  specify(response = weight) %>%
  hypothesize(null = "point", mu = 7.43) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

```
weight_boot_mean <- ex_3_weight_boot %>%
  summarise(mean = mean(stat)) %>% pull()
```

```
ggplot(data = ex_3_weight_boot, mapping = aes(x = stat)) +
  geom_histogram(binwidth = 0.01) +
  geom_vline(xintercept = weight_boot_mean, colour = "blue") +
  labs(title = "Null distribution of birth weight centered at 7.43 pounds",
       x = "Sample mean",
       y = "Count")
```



Exercise 4

```
ex_3_weight_boot %>%
  mutate(stat = round(stat, 2)) %>%
  get_p_value(obs_stat = 7.25, direction = "two_sided")
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1  0.0014
```

At the $\alpha = 0.05$ level, we have sufficient evidence to reject the null hypothesis, since the calculated p-value is 0.0014, which shows that it is quite unlikely that we would have gotten these data if the mean weight was 7.43 pounds. Hence, there is insufficient evidence to suggest that the average weight of babies born to Caucasian mothers in North Carolina is equal to 7.43 pounds.

Exercise 5

```
ex_3_sample %>%
  specify(response = weight) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean") %>%
  get_ci(level = .95)
```

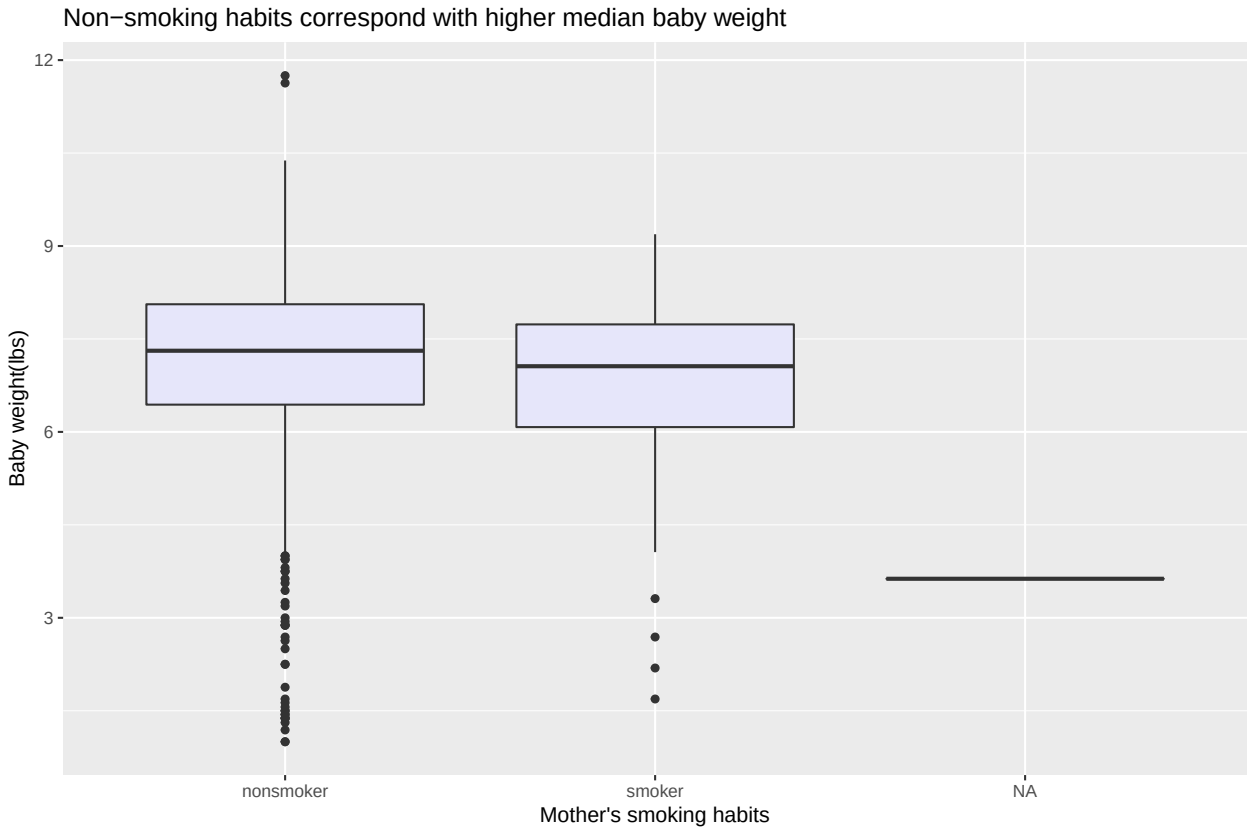
```
# A tibble: 1 x 2
  `2.5%` `97.5%`
   <dbl>   <dbl>
1   7.14    7.36
```

From this confidence interval, we are 95% confident that the true average weight of Caucasian babies born in NC lies between 7.14 pounds and 7.36 pounds. This interval does not cover the value of μ, i.e. 7.43, under the null hypothesis.

A confidence interval simply represents a plausible range of values for the population parameter, in this case, the true mean birth weight of babies born to Caucasian mothers in NC. Since Exercise 4 showed us that there is sufficient evidence to reject the null hypothesis which states that the mean birth-weight has stayed at 7.43 pounds, we can assume that this interval, with 95% confidence, captures the true mean weight which not remained at 7.43 pounds. As a result, this confidence interval is supported by the rejection of the null hypothesis.

Exercise 6

```
ncbirths %>%
  #filter(!is.na(habit)) %>%
  ggplot(mapping = aes(x = habit, y = weight, na.rm = TRUE)) +
  geom_boxplot(fill = "lavender") +
  labs(title = "Non-smoking habits correspond with higher median baby weight",
       x = "Mother's smoking habits", y = "Baby weight(lbs)")
```



This boxplot shows that there is a higher median baby weight for mothers who don't smoke compared to mothers who do smoke.The first and third quartile values for non-smoking mothers are also higher than those for smoking mothers. This indicates that there is likely a relationship between smoking habit and baby weight, where non-smoking corresponds with higher average baby weights.

However, the plot also shows a number of outliers in the data for mothers who don't smoke in the form of several very low baby weights. This could just be a function of the quantity and randomness of the data in the sample, or it could indicate that there are other factors affecting the relationship between smoking habit and baby weight.

Exercise 7

```
ncbirths_habitgiven <- ncbirths %>%
  filter(!is.na(habit))
```

Exercise 8

H_0: The true mean weight of babies born to smoking mothers is equal to the true mean weight of babies born to non-smoking mothers.
$(H_0 : \mu_1 = \mu_2)$

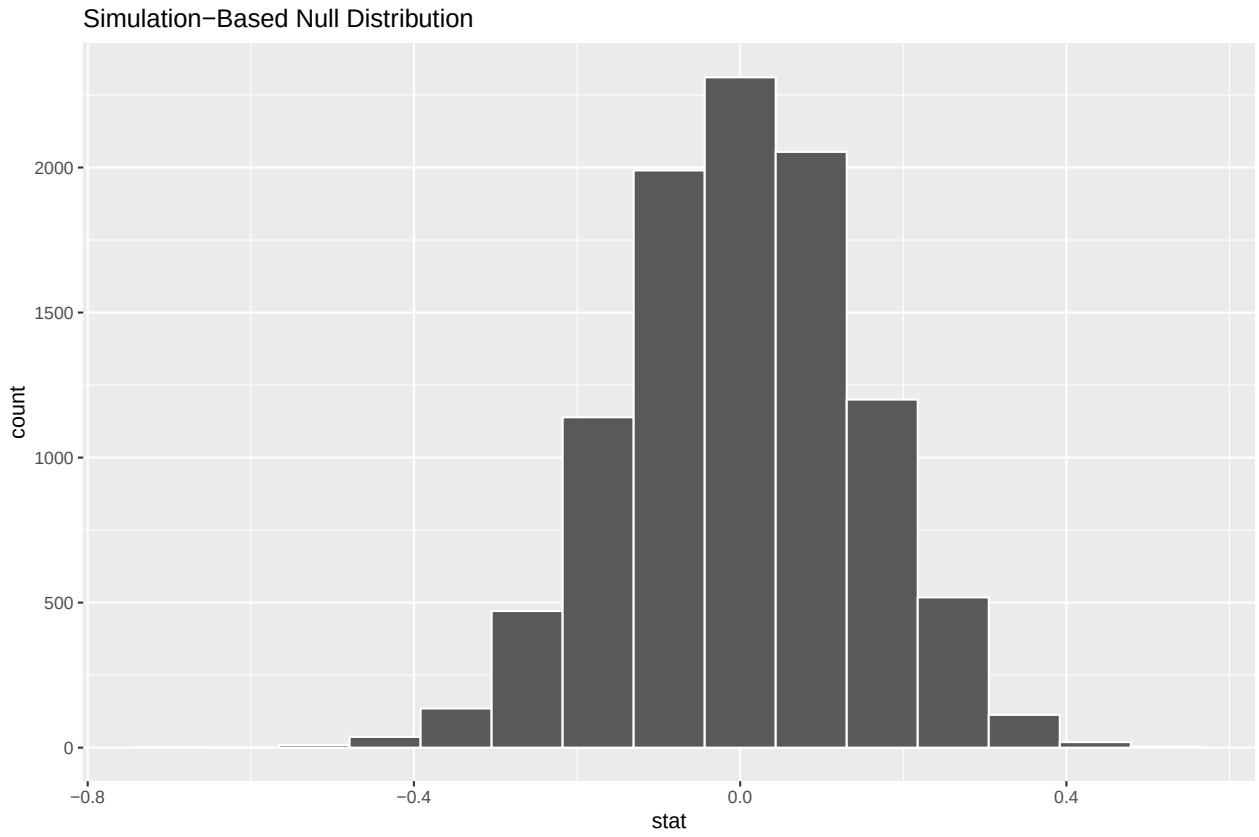
H_1: The true mean weight of babies born to smoking mothers is not equal to the true mean weight of babies born to non-smoking mothers. $(H_1 : \mu_1 ! = \mu_2)$

Exercise 9

```
obs_diff <- ncbirths_habitgiven %>%
  group_by(habit) %>%
  summarise(mean_weight = round(mean(weight),3)) %>%
  summarise(diff_means = diff(mean_weight))

obs_diff
```

```
# A tibble: 1 x 1
  diff_means
       <dbl>
1     -0.315
```

```
null_dist <- ncbirths_habitgiven %>%
  specify(response = weight, explanatory = habit) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("smoker", "nonsmoker"))

visualize(null_dist)
```




```
get_p_value(null_dist, obs_stat = 0.315, direction = "two_sided")
```

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

```
1   0.022
```

At the $\alpha = 0.10$ level, the results are statistically significant and we reject the null hypothesis because the p-value (0.023) is less than α. It is very unlikely to observe our data or more extreme if the null hypothesis were actually true. Thus, there is insufficient evidence to suggest that the true mean weight of babies born to smoking mothers is equal to the true mean weight of babies born to non-smoking mothers.

Exercise 10

Because I concluded that we reject the null hypothesis, it is possible that I could have made a Type I error. If a Type I error was made, the null hypothesis would be true. In the context of this testing problem, a Type I error would mean that assuming the null hypothesis, the true mean weight of babies born to smoking mothers is equal to the mean weight of babies born to non-smoking mothers.

Exercise 11

The probability of a Type I error is α which is specified by the researcher prior to conducting the test as 0.10.