Assignment 8: Birth Times

Minh Tran

2021-06-05

Exercise 1

```r
live_births <- nsfg6 %>%
  filter(outcome == 1)
```

```r
live_births <- live_births %>%
  mutate(birth_order = if_else(birthord == 1,
                               'first',
                               'other'))
```

```r
pregnancy_length <- live_births %>%
  select(prglngth, birth_order)
```

```r
glimpse(pregnancy_length)
```

```
## Rows: 9,148
## Columns: 2
## $ prglngth    <int> 39, 39, 39, 39, 39, 38, 40, 42, 39, 35, 39, 39, 37, 33,...
## $ birth_order <chr> "first", "other", "first", "other", "other", "first", "...
```

Exercise 2

```r
ggplot(data = pregnancy_length) +
  geom_histogram(aes(x = prglngth, fill = birth_order), binwidth = 1) +
  coord_cartesian(xlim = combine(27, 46)) +
  labs(title = 'Spread of Pregnancy Length between First and Other Birth Orders',
       x = 'Pregnancy Length',
       y = 'Count')
```

```
## Warning: `combine()` is deprecated as of dplyr 1.0.0.
## Please use `vctrs::vec_c()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

Spread of Pregnancy Length between First and Other E
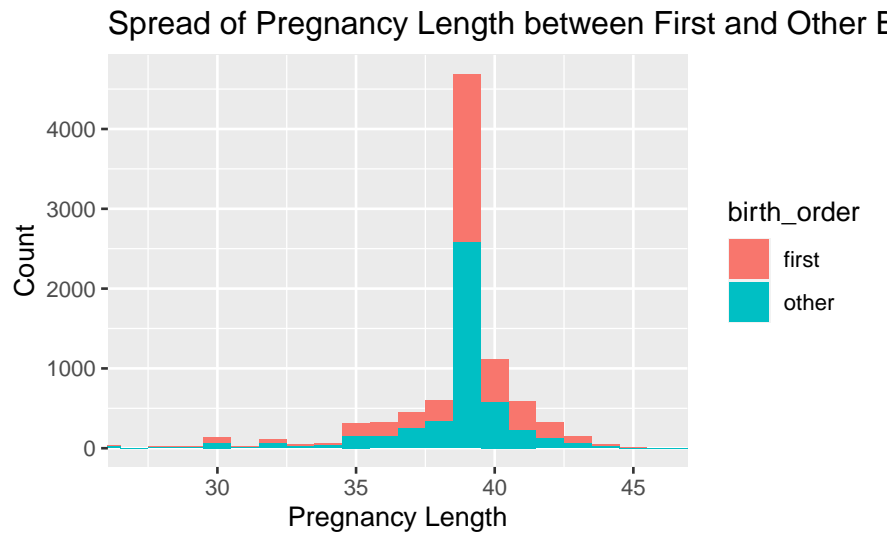
The mode of the pregnancy length lies around 39. Based on the visuals itself, it is kind of hard to determine if there is a difference in pregnancy length between 'first' and 'other' birth order because they are spread out similarly. As a result, a hypothesis test will be needed to make sure if there is a difference or not.

Exercise 3

```
pregnancy_length %>%
  group_by(birth_order) %>%
  summarise(mean = mean(prglngth),
            median = median(prglngth),
            sd = sd(prglngth),
            iqr = IQR(prglngth),
            minimum = min(prglngth),
            maximum = max(prglngth))
```

birth_order	mean	median	sd	iqr	minimum	maximum
first	38.60095	39	2.791901	1	0	48
other	38.52291	39	2.615852	0	4	50

Comparing the summary statistics between the two groups shows that there isn't that much significant difference between the two groups. There is a noticable difference in the standard deviation but nothing too significant.

Exercise 4

The test statistic is the difference in means in pregnancy length between the first and other group

Null Hypothesis - There is no difference in pregnancy length between first-borns and non-first-borns Alternative Hypothesis - There is a difference in pregnancy length between first-borns and non-first-borns

We will be using the two-sided test since we want to find out if first born children either arrive early or late when compared with non-first-borns.

Exercise 5

```
preg_null <- pregnancy_length %>%
  specify(prglngth ~ birth_order) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000,  type = "permute") %>%
  calculate(stat = "diff in means",  order = c("first", "other"))
```

```
preg_obs_stat <- pregnancy_length %>%
  specify(prglngth ~ birth_order) %>%
  calculate(stat = "diff in means",  order = c("first", "other"))
```

```
preg_null %>%
  get_p_value(obs_stat = preg_obs_stat, direction = "two_sided")
```

p_value
0.1694

```
preg_null %>%
  visualize() +
  shade_p_value(obs_stat = preg_obs_stat, direction = "two_sided") +
  labs(title = 'First-Borns and Non-First Borns Null Distribution' ,
       x = 'Difference in Means of Pregnancy Length')
```
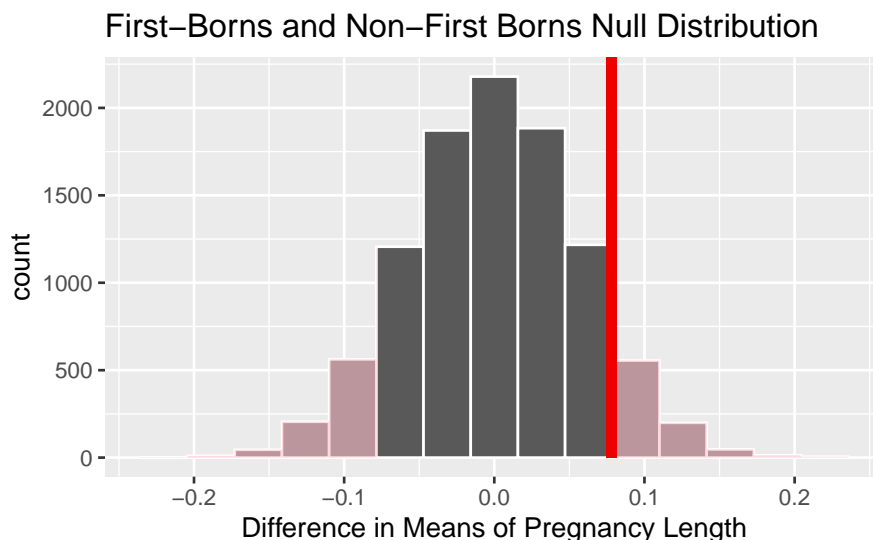

We fail to reject the null hypothesis since the p-value is greater than $\alpha = 0.05$. As a result, we conclude that there is no significant difference in pregnancy length first-borns and non-first-borns.

Exercise 6

```
birth_bootstraps <- pregnancy_length %>%
  specify(prglngth ~ birth_order) %>%
  generate(10000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c('first', 'other'))
```

```
bootstrap_ci <- birth_bootstraps %>%
  get_confidence_interval()
bootstrap_ci
```

2.5%	97.5%
-0.0345438	0.1901824

The observed statistic does fall within the range of the 95% confidence interval.

```
birth_bootstraps %>%
  visualize() +
  shade_confidence_interval(bootstrap_ci) +
  labs(title = '95% Confidence Interval of Bootstrap Distribution' ,
       x = 'Difference in Means of Pregnancy Length')
```
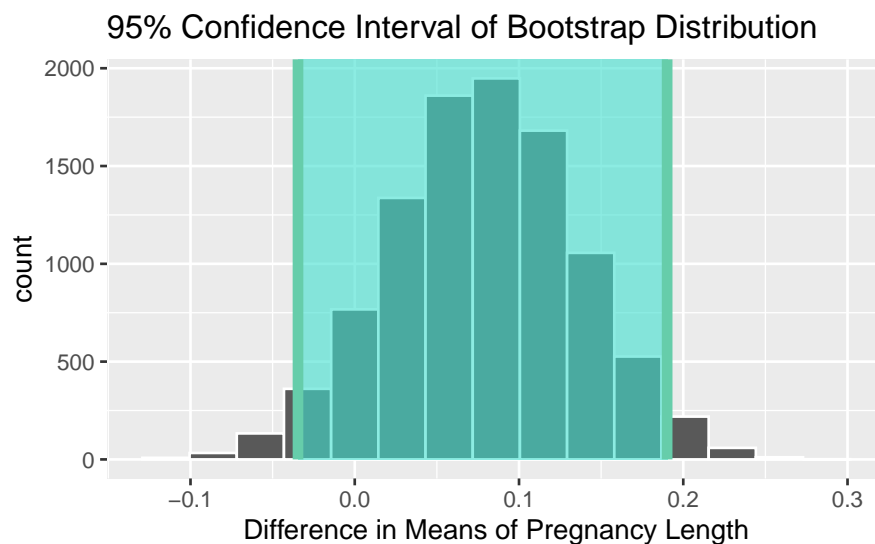

Exercise 7

```
bootstrap_results <- cohens_d_bootstrap(data = pregnancy_length, model = prglngth ~ birth_order
```

```
bootstrap_report(bootstrap_results)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
```

```
## boot::boot.ci(boot.out = cohens_d_bootstrap_sim, type = c("perc"))
##
## Intervals :
## Level      Percentile
## 95%    (-0.0124,  0.0708 )
## Calculations and Intervals on Original Scale
##
## Response variable
## prglngth
##
## Explanatory variable
## birth_order
##
## Explanatory category with larger mean
## first
##
## Explanatory category with smaller mean
## other
##
## Cohen's d observed value
## 0.0288791
```

```r
plot_ci(bootstrap_results)
```
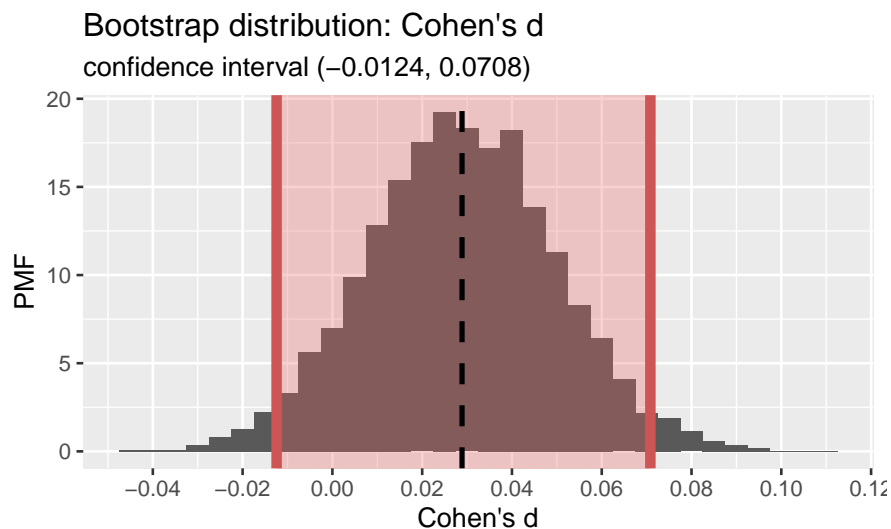

According to the table, the Cohen's d observed value of 0.0288791 indicates that it is a very small effect size.