Lab 7: Birth Ratios

Sangwon Yum

2023-10-20

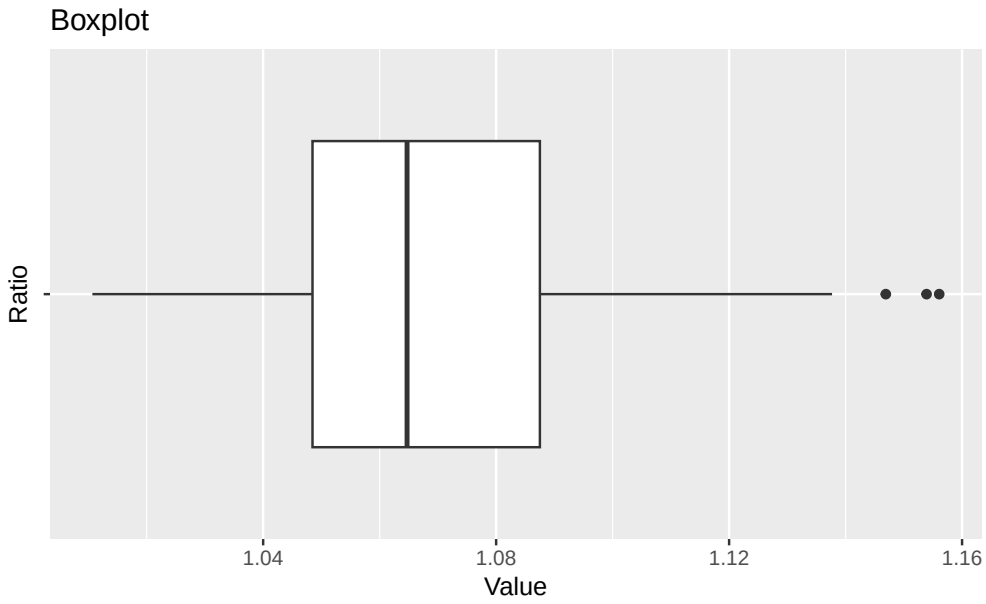
Exercise 1

```
data = Arbuthnot %>%
  tibble::tibble()

help(data)
# View(data)


dupData = data %>%
  group_by(Year) %>%
  count() %>%
  filter(n > 1)

print(dupData)
```

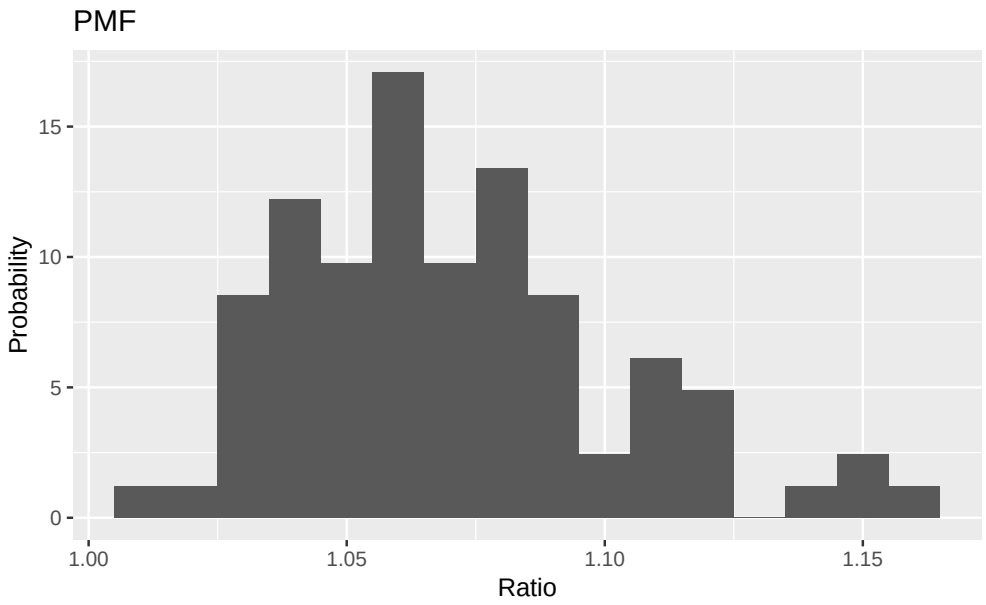
```
## # A tibble: 0 x 2
## # Groups:   Year [0]
## # i 2 variables: Year <int>, n <int>
```

Exercise 2

```
ggplot(data = data) +
  geom_boxplot(aes(x = "", y = Ratio)) +
  coord_flip() +
  labs(title = "Boxplot", x = "Ratio", y = "Value")
```

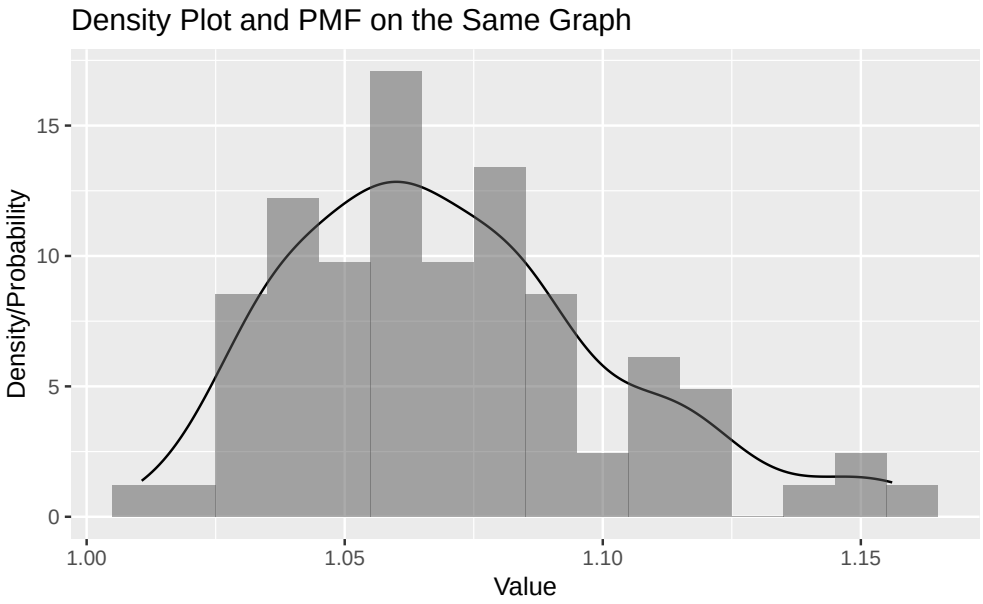



```
ggplot(data = data, aes(x = Ratio)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.01) +
  labs(title = "PMF", x = "Ratio", y = "Probability")
```



Exercise 3

```
ggplot(data = data) +
  geom_density(aes(x = Ratio, y = ..density..), alpha=0.5) +
  geom_histogram(aes(x = Ratio, y = ..density..), binwidth = 0.01, alpha=0.5) +
  labs(title="Density Plot and PMF on the Same Graph", x="Value", y="Density/Probability")
```



Exercise 4

- Q) Which summary statistics are sensitive to outliers?
- A) mean, sd, min, max
- Q) Which summary statistics are not sensitive to outliers?
- A) median, iqr

```
data %>%
  summarise(
    mean = mean(Ratio, na.rm = TRUE)
    , median = median(Ratio, na.rm = TRUE)
    , sd = sd(Ratio, na.rm = TRUE)
    , iqr = IQR(Ratio, na.rm = TRUE)
    , min = min(Ratio, na.rm = TRUE)
    , max = max(Ratio, na.rm = TRUE)
  )
```

mean	median	sd	iqr	min	max
1.070748	1.064704	0.0312537	0.0390408	1.010673	1.156075

Exercise 5

- Q) After performing the two-sided hypothesis test, explain the result of your hypothesis testing

- A) (p-value) 0.05 1

```
# Exercise 5 part 1
data_null = data %>%
  specify(formula = Ratio ~ NULL) %>%
  hypothesise(null = "point", mu = 1) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
print(data_null)
```

```
## Response: Ratio (numeric)
## Null Hypothesis: point
## # A tibble: 10,000 x 2
##    replicate  stat
##        <int> <dbl>
##  1         1 1.00
##  2         2 1.00
##  3         3 0.999
##  4         4 0.996
##  5         5 1.00
##  6         6 0.996
##  7         7 1.00
##  8         8 0.998
##  9         9 0.999
## 10        10 1.00
## # i 9,990 more rows
```

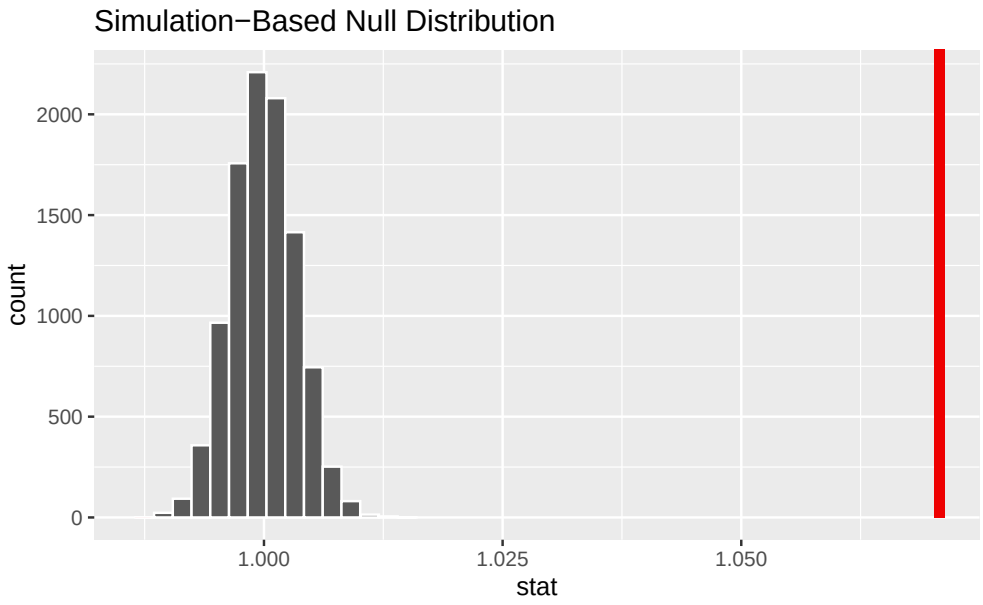
```
# Exercise 5 part 2
data_obs_stat = data %>%
  specify(formula = Ratio ~ NULL) %>%
  calculate(stat = "mean")
print(data_obs_stat)
```

```
## Response: Ratio (numeric)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  1.07
```

```
# Exercise 5 part 3
data_null %>%
  get_p_value(obs_stat = data_obs_stat, direction = "two-sided")
```

p_value
0

```
# Exercise 5 part 4
data_null %>%
  visualize() +
  shade_p_value(obs_stat = data_obs_stat, direction = "two-sided")
```



Exercise 6

- Q) Is there a difference between this hypothesis test result and the result of Ex 5?
- A) Exercise 5
- Q) Explain why
- A) mu mu

```
# Exercise 6 part 1
data_null2 = data %>%
  specify(formula = Ratio ~ NULL) %>%
  hypothesise(null = "point", mu = 1.05) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
print(data_null2)
```

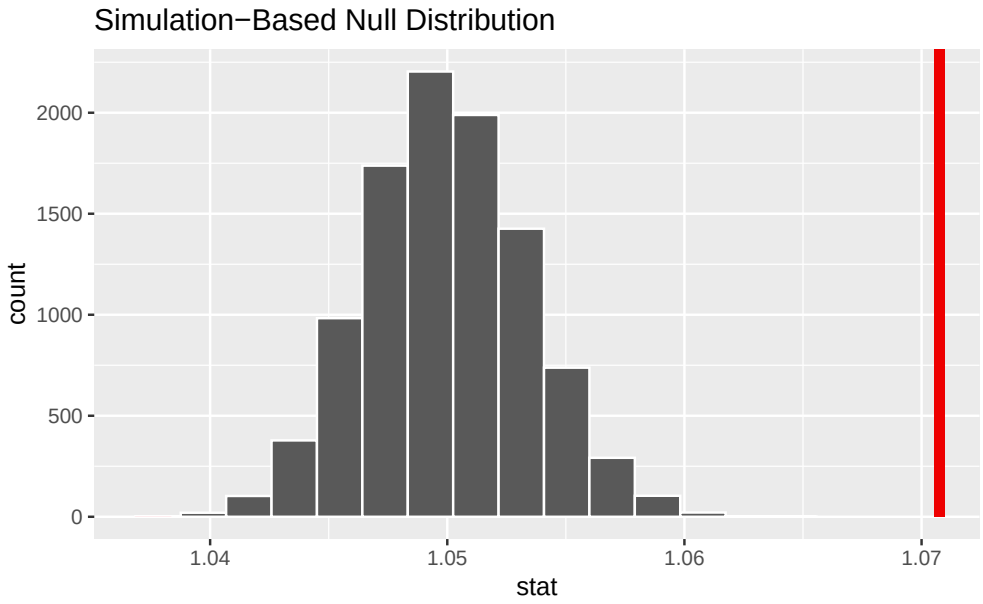
```
## Response: Ratio (numeric)
## Null Hypothesis: point
```

```
## # A tibble: 10,000 x 2
##    replicate  stat
##        <int> <dbl>
##  1         1  1.05
##  2         2  1.05
##  3         3  1.06
##  4         4  1.05
##  5         5  1.06
##  6         6  1.05
##  7         7  1.05
##  8         8  1.04
##  9         9  1.04
## 10        10  1.05
## # i 9,990 more rows
```

```
# Exercise 6 part 2
# Exercise 5 part 2

# Exercise 6 part 3
data_null2 %>%
  get_p_value(obs_stat = data_obs_stat, direction = "two-sided")
```

p_value
0

```
# Exercise 6 part 4
data_null2 %>%
  visualize() +
  shade_p_value(obs_stat = data_obs_stat, direction = "two-sided")
```


Simulation−Based Null Distribution
count
2000
1500
1000
500
0
1.04
1.05
1.06
1.07
stat