Lab 7: Birth Ratios

Minsu Kang and Bomin Lyoo

2022-07-16

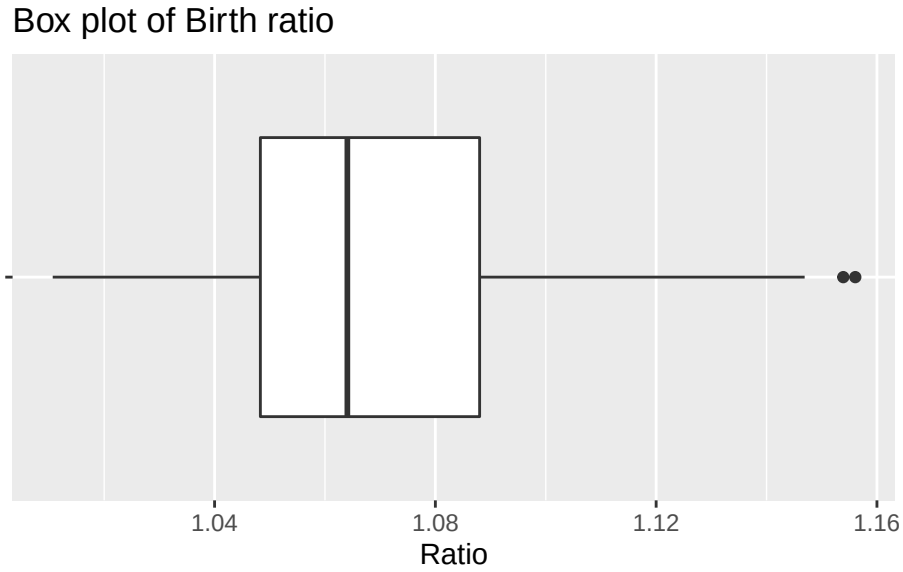
Visualizing and quantifying the distribution

Exercise 1

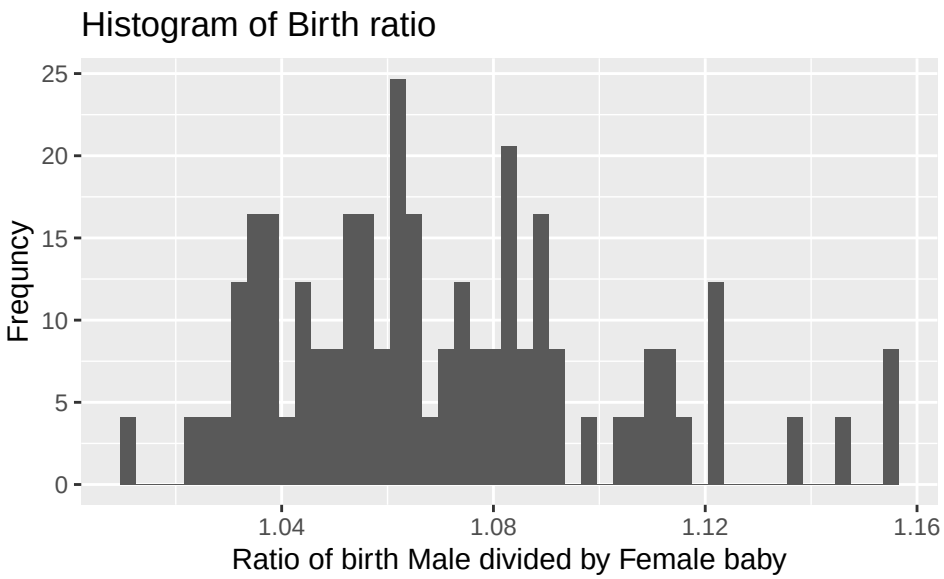
```
arbuthnot <- Arbuthnot %>% filter(Year != 1704)
```

Exercise 2

```
ggplot(data = arbuthnot) +
  geom_boxplot(aes(x = "", y = Ratio)) +
  coord_flip() +
  labs(title = "Box plot of Birth ratio", x="", y="Ratio")
```



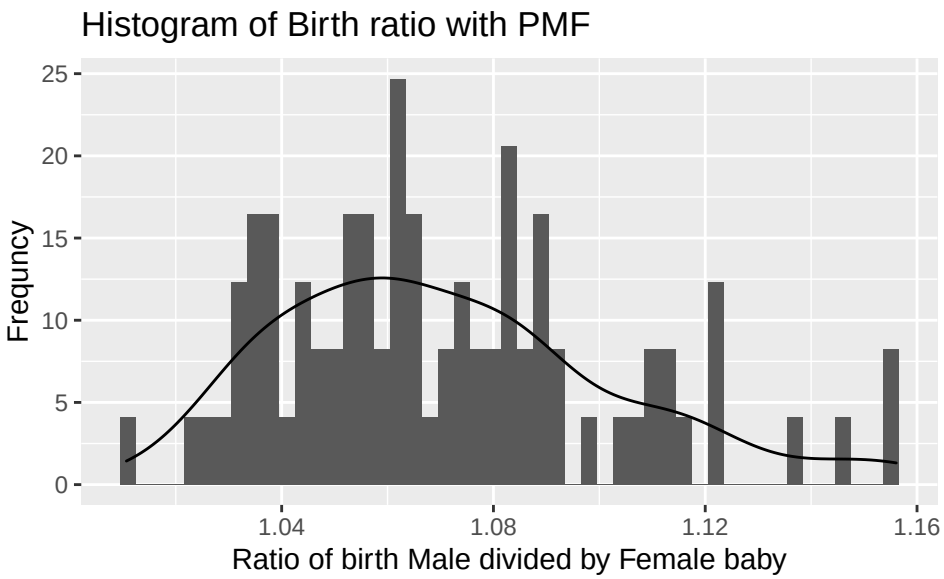
```
ggplot(data = arbuthnot) +
  geom_histogram(mapping = aes(x=Ratio, y = ..density..,), bins=40, binwidth= 0.003) +
  labs(title = "Histogram of Birth ratio",
       x="Ratio of birth Male divided by Female baby",
       y="Frequncy")
```



i) Asymmetrical, right-skewed histogram, and uni-modal. Range of the graph is 1.156 - 1.011, which is 0.145. Outliers lie on approximately 1.153 to 1.156.

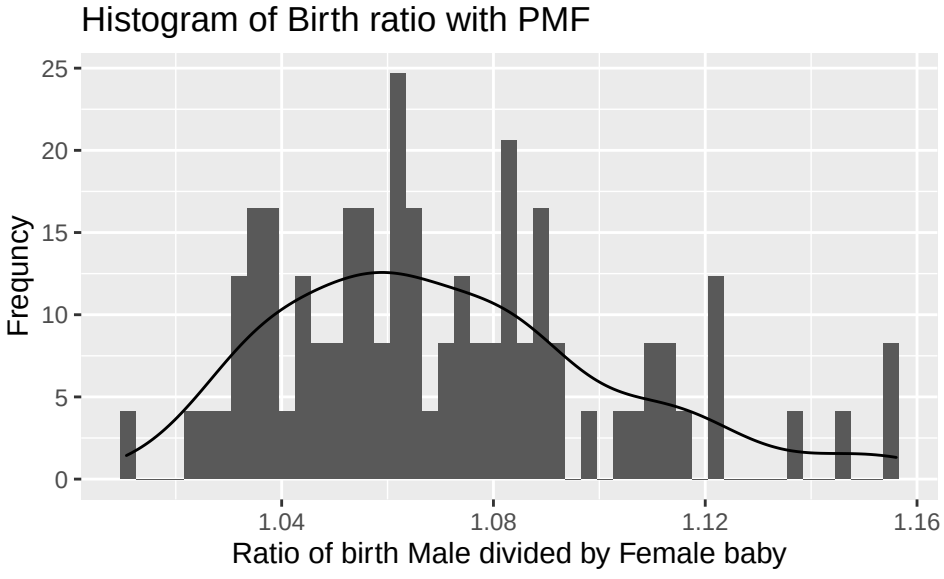
```
ggplot(data = arbuthnot) +
  geom_histogram(mapping = aes(x=Ratio, y = ..density..), bins=40, binwidth= 0.003) +
  geom_density(mapping = aes(x = Ratio),  bins = 11) +
  labs(title = "Histogram of Birth ratio with PMF",
       x="Ratio of birth Male divided by Female baby",
       y="Frequncy")
```

```
## Warning: Ignoring unknown parameters: bins
```



Exercise 3

```
ggplot(data = arbuthnot) +
  geom_histogram(mapping = aes(x=Ratio, y = ..density..),
                 bins=40, binwidth= 0.003) +
    geom_density(mapping = aes(x=Ratio)) +
  labs(title = "Histogram of Birth ratio with PMF",
       x="Ratio of birth Male divided by Female baby",
       y="Frequncy")
```



Exercise 4

```
arbuthnot %>%
  summarize(
    mean = mean(Ratio),
    median = median(Ratio),
    sd = sd(Ratio),
    iqr = IQR(Ratio),
    min = min(Ratio),
    max = max(Ratio)
  )
```

mean	median	sd	iqr	min	max
1.070815	1.064054	0.0314426	0.0397189	1.010673	1.156075

```
i) max and minimum is the most sensitive to outliers.
The outliers can be over or less than 1.5 IQR,
which means if all variables are in 1.5 IQR,
outliers equals to maximum or minimum number.

ii) The IQR and medians are less effected by outliers.
    This is because outliers are over 1.5IQR,
    but IQR decided by the number in 25% quartile(second) and 75% quartile(third).

    Also median is center number, emergence of new outliers can change the order of count by
    adding more numbers, but next closest number became median.
    If the number is not far from former median, there will be minuscule effect.

    Moreover, standard deviation robust to outliers.
    The outliers influence to distribution (as it distributed over normal values),
    standard deviatioin rise when more outliers comes out,
    but other values deviations are within the stats,
    so it become neutralized.

Same reson, the mean is robust to outliers.
If outliers increase the number of numberator in mean formula,
it mixed with other number which other normal values made.
Therefore, we can say these stats are 'robust' to outliers.
```

infering a trend

Exercise 5

```
i) The null hypothesis is 'the mean of male and female births is 1'.
The alternative hypothesis is flipped, 'the mean of ratio male and female birth is not 1'.
```

```
arbuthnot_null <- arbuthnot %>%
  specify(formula = Ratio ~ NULL) %>%
  hypothesize(null = "point", mu = 1) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

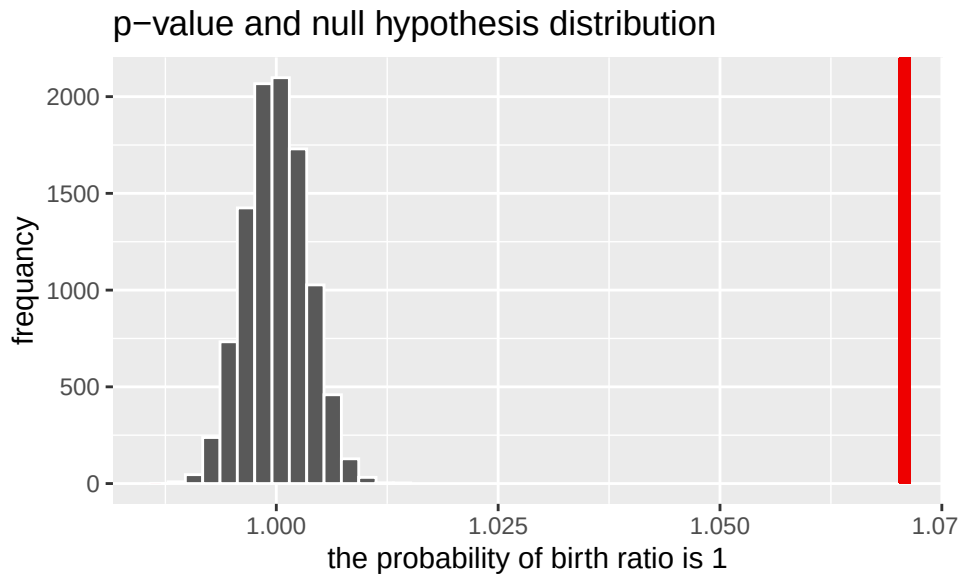
```
arbuthnot_obs_stat <- arbuthnot %>%
  specify(formula = Ratio ~ NULL) %>%
  calculate(stat = "mean")
```

```
arbuthnot_null %>%
  get_p_value(obs_stat = arbuthnot_obs_stat, direction = "two_sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

p_value
0

```
visualize(arbuthnot_null) +
  shade_p_value(obs_stat = arbuthnot_obs_stat, direction = "two_sided") +
  labs(title= "p-value and null hypothesis distribution",
       x= "the probability of birth ratio is 1",
       y="frequancy")
```



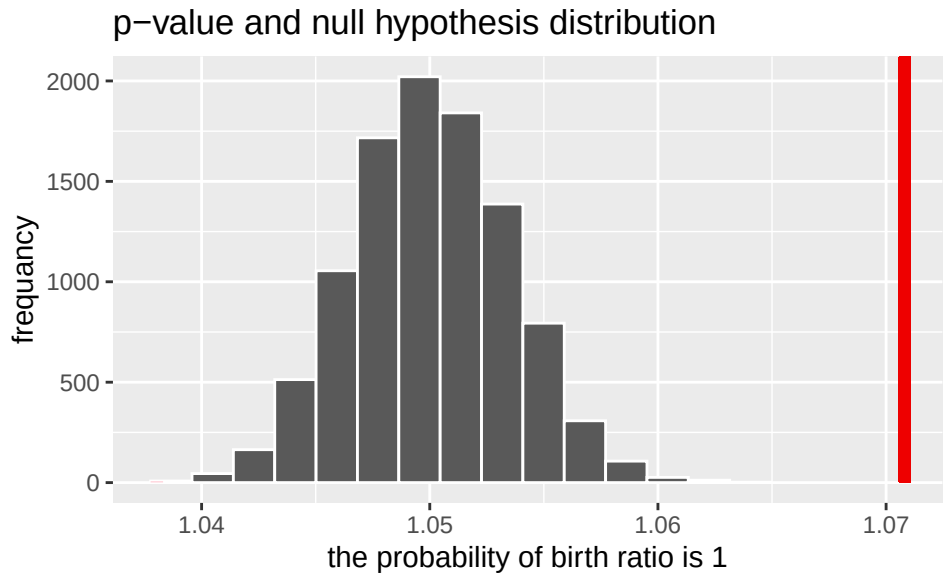
Exercise 6

```
arbuthnot_null <- arbuthnot %>%
  specify(formula = Ratio ~ NULL) %>%
  hypothesize(null = "point", mu = 1.05) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

arbuthnot_obs_stat <- arbuthnot %>%
  specify(Ratio ~ NULL) %>%
  calculate(stat = "mean")

arbuthnot_null %>%
  get_p_value(obs_stat = arbuthnot_obs_stat, direction = "two_sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

p_value
0

```
visualize(arbuthnot_null) +
  shade_p_value(obs_stat = arbuthnot_obs_stat, direction = "two_sided") +
  labs(title= "p-value and null hypothesis distribution",
       x= "the probability of birth ratio is 1",
       y="frequancy")
```



i) The p-value and distribution histogram is moved to left, and the difference between p-value and null-distribution is reduced. It contains the meaning that 1.05 is more closer to the real world observation, which was ratio 1 we did in exercise5. However, still it is not over the p-value, so the mean of the birth ratio have to over 1.07.