lab02

Part I: Understanding the Context of the Data

1. **What question did John Arbuthnot set out to answer in collecting this data?**

   What do you think the probability is that a newborn child is recorded as a girl?

2. **What is the unit of observation in the original christening records? What are the possible variables that may have been recorded?**

   A child (a record). Name, sex, birth of date, parent's name.

3. **What do you think the probability is that a newborn child is recorded as a girl? What form of evidence or reasoning did you use to come to that determination?**

   0.5 based on biological evidence.

Part II: Computing on the Data

```
library(stat20data)
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.0      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
data(arbuthnot)
```

4. **What does every row in the data frame published by John Arbuthnot correspond to? What are the names of the variables and what is the type of each one?**

   Each row means the numbers of boys and girls were born.

   - year: the year of the data collected / discrete numerical
   - boys: the numbers of boys were born in the specific year / nominal categorical data
   - girls: the numbers of girls were born in the specific year / nominal categorical data

5. **What is the time frame covered by Arbuthnot's data?**

   ```r
   temp <- select(arbuthnot, year)
   year1 <- max(temp)
   year2 <- min(temp)
   ```

   From 1629 to 1710.

6. **Which year saw the greatest number of children christened?**

   ```r
   arbuthnot <- mutate(arbuthnot, total = boys + girls)
   max_children <- max(select(arbuthnot, total))
   filter(arbuthnot, total == max_children)
   ```

   ```
   # A tibble: 1 x 4
      year  boys girls total
     <int> <int> <int> <int>
   1  1705  8366  7779 16145
   ```

   In year 1705, the greatest number of children christened.

7. **What is the proportion of girls christened in 1700?**

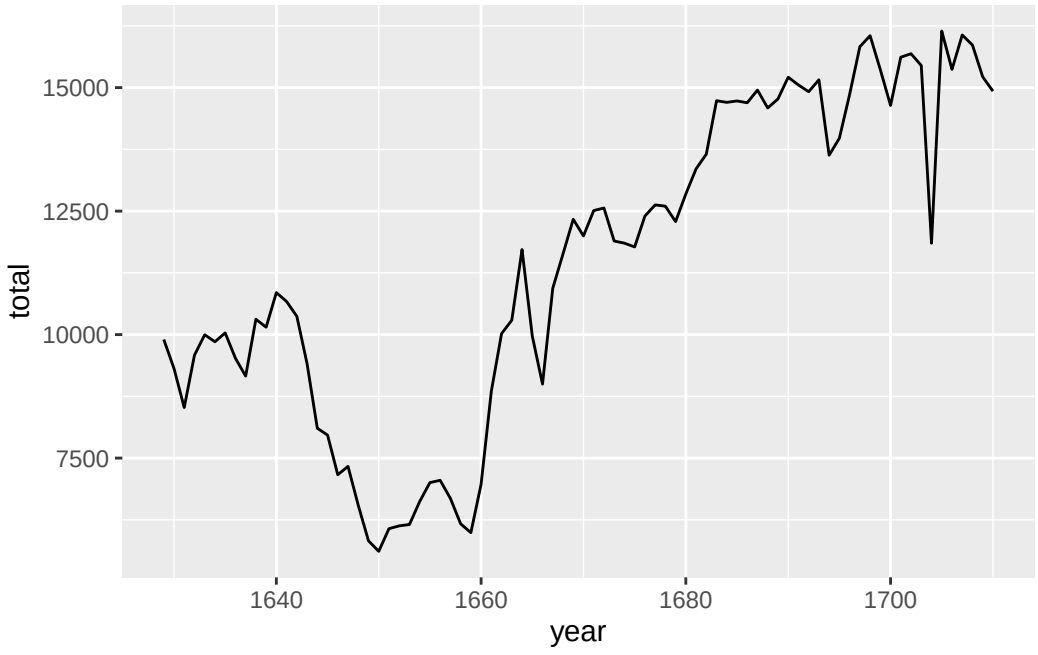
   ```r
   arbuthnot <- mutate(arbuthnot, prop_girl = (girls / (boys + girls)))
   filter(arbuthnot, year == 1700)
   ```

   ```
   # A tibble: 1 x 5
      year  boys girls total prop_girl
     <int> <int> <int> <int>     <dbl>
   1  1700  7578  7061 14639     0.482
   ```

The proportion of girl christened in 1700 is 0.4823.

8. **What is the trend over time in the total number of children christened? Please answer with a plot and written interpretation.**

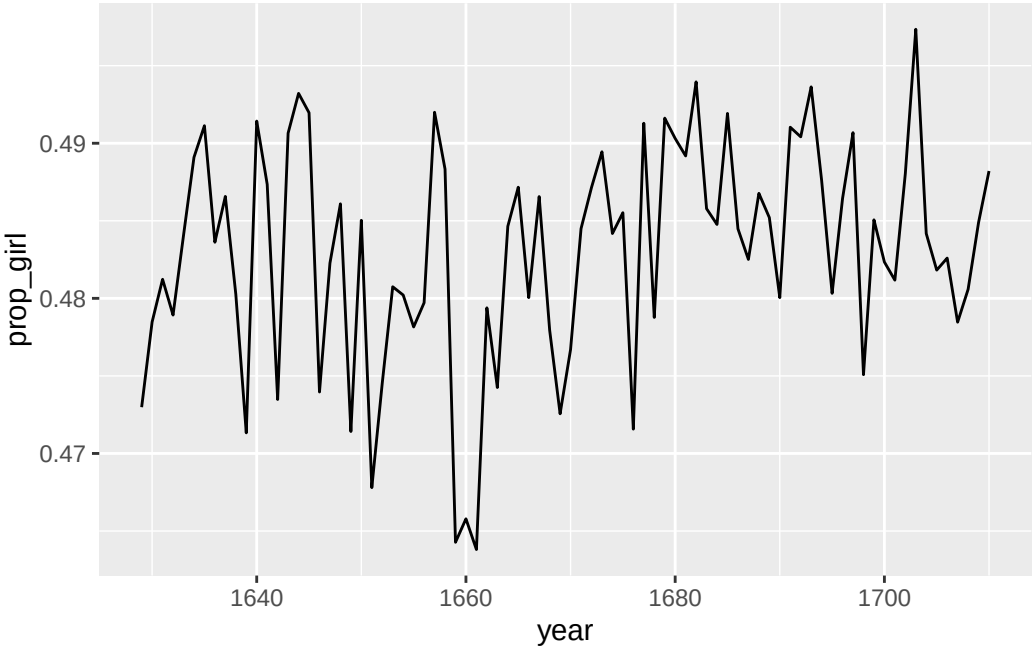
```
ggplot(arbuthnot, aes(x = year, y = total)) +
  geom_line()
```



This data plot shows that there is a positive correlation between year and total number of children christened. When the year increased, the number of total children christened also increased.

9. **What is the trend over time in the proportion of girls christened? Please answer with a plot and written interpretation.**

```
ggplot(arbuthnot, aes(x = year, y = prop_girl)) +
  geom_line()
```



This data plot shows the relationship between year and the proportion of girls christened. When the year increased, the proportion of girls christened does not increased as well. For example, during 1940 to 1942, the proportion of girls christened decreased a lot. However, during 1942 to 1945, the proportion of girls christened increased dramatically. Therefore, we can say that the proportion of girls christened does not have correlation with year.

Part III: Extensions

```
data(present)
```

10. **What is the time frame covered by the present-day data?**

```
temp1 <- select(present, year)
max_year <- max(temp1)
min_year2 <- min(temp1)
```

From 1940 to 2013.

11. **In terms of general magnitude (size), how do the counts in Arbuthnot's data compare to the counts in the present-day data?**

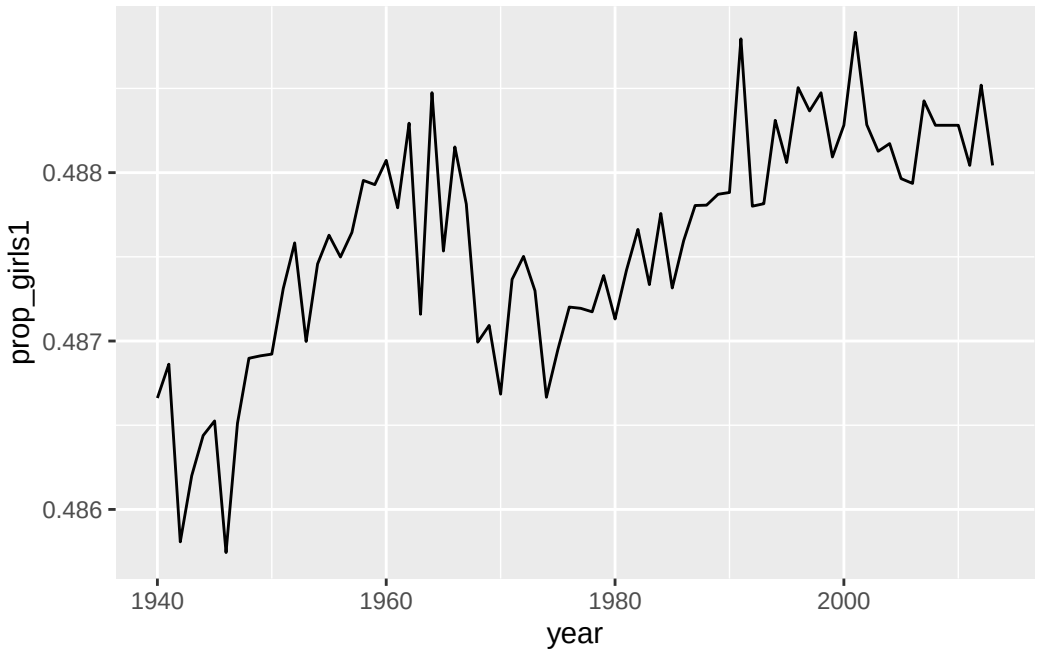
The general magnitude of Arbuthnot's data compares to the counts in the present-day data is quite small because Arbuthnot is just a village which means that the data of Arbuthnot is sample data. The size of present-day data is much larger than the size of Arbuthnot's data.

12. **What is the trend over time in the proportion of births that are girls? Please answer with a plot and written interpretation.**

```
present <- mutate(present, prop_girls1 = (girls / (girls + boys)))
select(present, prop_girls1)
```

```
# A tibble: 74 x 1
   prop_girls1
         <dbl>
 1       0.487
 2       0.487
 3       0.486
 4       0.486
 5       0.486
 6       0.487
 7       0.486
 8       0.487
 9       0.487
10       0.487
# ... with 64 more rows
```

```
ggplot(present, aes(x = year, y = prop_girls1)) +
  geom_line()
```



This plot shows a positive correlation between year and the proportion of births that are girls. The trend over time in the proportion of births that are girls in the present-day is increasing. In the plot, we can see that the data of the trend over time in the proportion of births that are girls are gradually increasing with the year pass.

13. **Based on these two data sets, what claim are you prepared to make regarding John Arbuthnot's original question? What reservations, if any, do you have about using this data to make the claim?**

    Based on these two data sets, I would like to refute my previous conjecture. In both data sets, the proportions of girls christened have never reached 0.5. It is hard to come with a general conclusion for year of 1629 to 1710 from Arbuthnot's data because this data only represents the situation of this village. However, we can conclude that during 1629 to 1710 the proportions of girls christened in Arbuthnot are unstable but do not change a lot over time. In present-day data, the proportion of girls christened has grown gradually over the years collected, and it almost reach 0.49, which is close to my original thought. Several situations may cause this data unreliable such as abortion, non-reporting, lack of tracing. Nevertheless, when the loss of reporting happens every year, the error can somehow be neglected. Therefore, this data is still able to be used for checking the new birth population.