Grando-0

John Grando

January 31, 2017

Lab 0 Assignment

First, let's set the working directory and source the data.

```
setwd("~/Documents/Masters/DATA606/Week1/Lab/Lab0")
source("more/arbuthnot.r")
# Source present day data
source("more/present.R")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

Exercize 1 - What command would you use to extract just the counts of girls baptized? Try it!

Answer:

For a list similar to the example, I would just reference the girls column

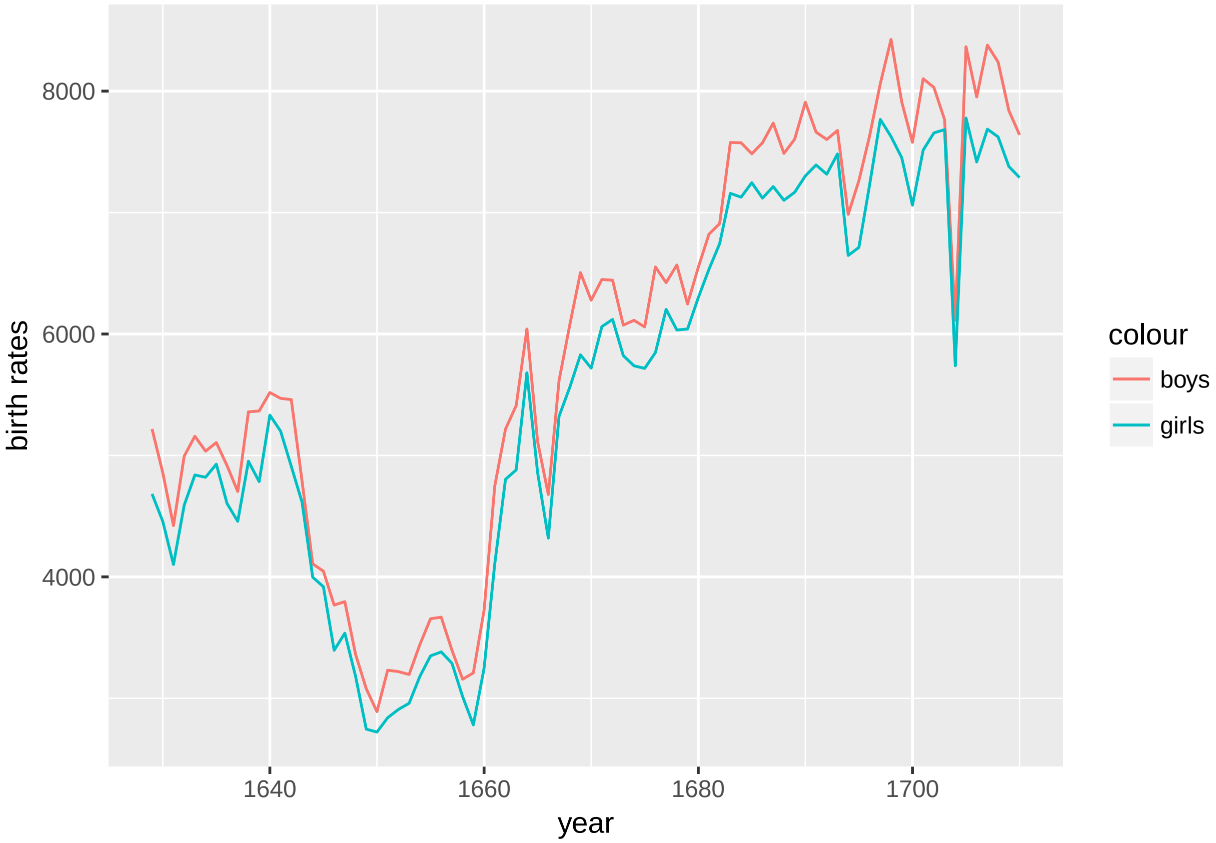
```
arbuthnot$girls
```

```
##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910
## [15] 4617 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382
## [29] 3289 3013 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719
## [43] 6061 6120 5822 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127
## [57] 7246 7119 7214 7101 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626
## [71] 7452 7061 7514 7656 7683 5738 7779 7417 7687 7623 7380 7288
```

Exercize 2 - Is there an apparent trend in the number of girls baptized over the years? How would you describe it?

Answer:

```
ggplot(arbuthnot, aes(year)) + geom_line(aes(y = arbuthnot$boys,
    color = "boys")) + geom_line(aes(y = arbuthnot$girls, color = "girls")) +
    labs(y = "birth rates")
```

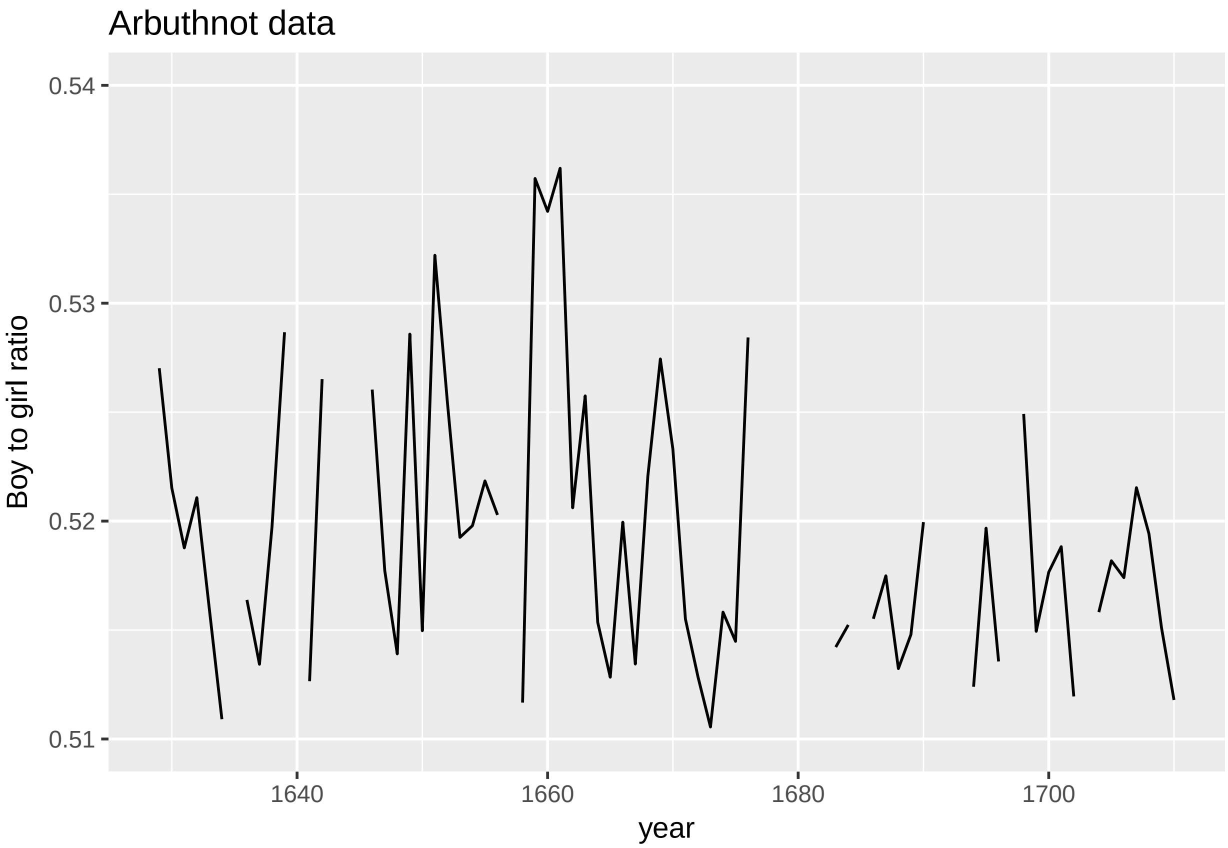


It appears that there is a dip in birth rates from 1640 to 1660; however, the fluctuations appear to be similar to the boy birth rates.

Exercize 3 - Now, make a plot of the proportion of boys over time. What do you see? Tip: If you use the up and down arrow keys, you can scroll through your previous commands, your so-called command history. You can also access it by clicking on the history tab in the upper right panel. This will save you a lot of typing in the future.

Answer:

```
arbuthnot$BGRatio <- arbuthnot$boys/(arbuthnot$girls + arbuthnot$boys)
ggplot(arbuthnot, aes(year)) + geom_line(aes(y = BGRatio)) +
    ggtitle("Arbuthnot data") + scale_y_continuous(limits = c(0.51,
    0.54), breaks = seq(0.51, 0.54, by = 0.01)) + labs(y = "Boy to girl ratio")
```



It appears that more boys than girls are born every year because the ratio is always above 50%.

Question 1 - What years are included in this data set? What are the dimensions of the data frame and what are the variable or column names?

I have chosen to select the levels of the data set, just in case a year is repeated in rows, so i'm sure I have unique values returned.

```
levels(factor(present$year))
```

```
##  [1] "1940" "1941" "1942" "1943" "1944" "1945" "1946" "1947" "1948" "1949"
## [11] "1950" "1951" "1952" "1953" "1954" "1955" "1956" "1957" "1958" "1959"
## [21] "1960" "1961" "1962" "1963" "1964" "1965" "1966" "1967" "1968" "1969"
## [31] "1970" "1971" "1972" "1973" "1974" "1975" "1976" "1977" "1978" "1979"
## [41] "1980" "1981" "1982" "1983" "1984" "1985" "1986" "1987" "1988" "1989"
## [51] "1990" "1991" "1992" "1993" "1994" "1995" "1996" "1997" "1998" "1999"
## [61] "2000" "2001" "2002"
```

I will return the data frame dimension using dim which shows this set has 63 rows and 3 columns. I then show the column names by just using the names() function.

```
dim(x = present)
```

```
## [1] 63  3
```

```
names(present)
```

```
## [1] "year"  "boys"  "girls"
```

Question - 2 How do these counts compare to Arbuthnot's? Are they on a similar scale?

```r
present$total <- present$boys + present$girls
arbuthnot$total <- arbuthnot$boys + arbuthnot$girls
summary(arbuthnot$total)
```

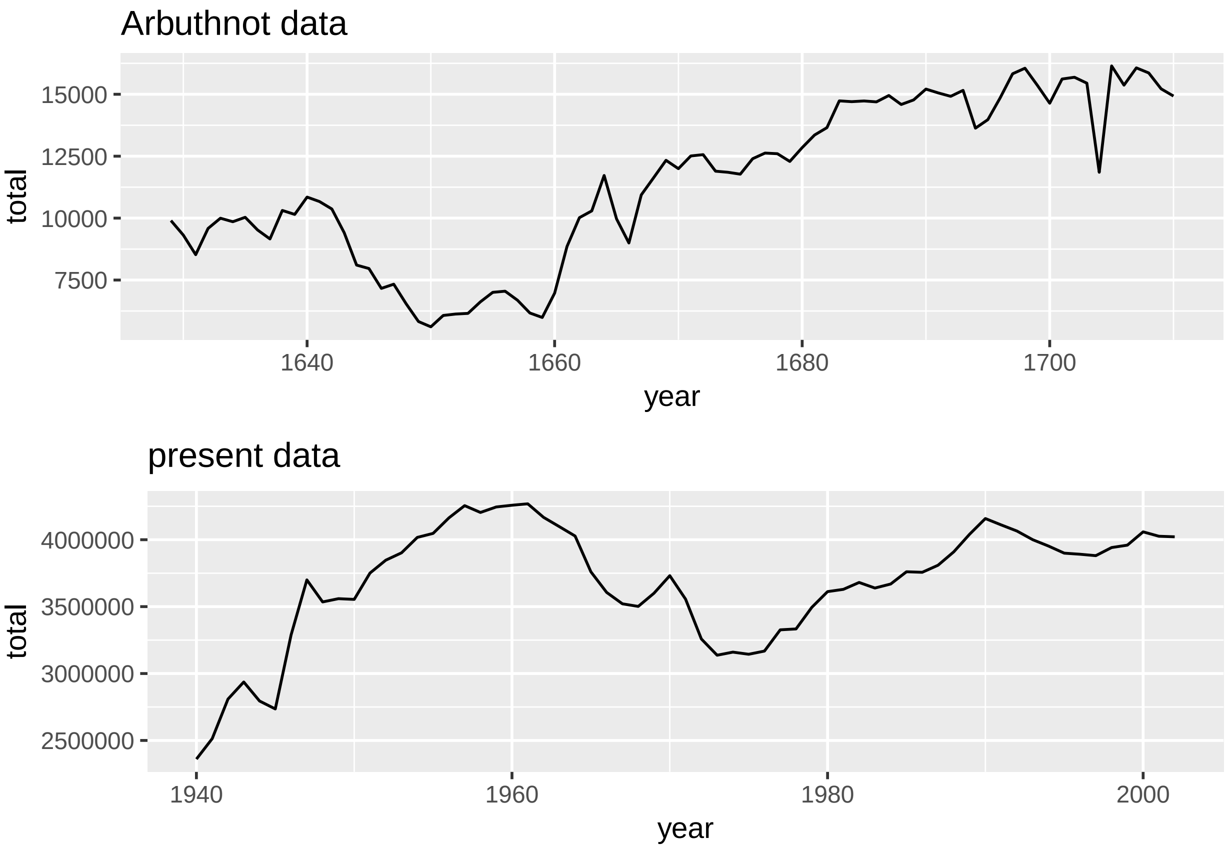
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5612    9199   11810   11440   14720   16140
```

```r
summary(present$total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 2360000 3511000 3757000 3680000 4024000 4268000
```

From the summary information, it can be seen that that the number of births tracked is much larger for the present day data than the arbuthnot data. They are not on a similar scale.

```r
plot1 <- ggplot(arbuthnot, aes(year)) + geom_line(aes(y = total)) +
    ggtitle("Arbuthnot data")
plot2 <- ggplot(present, aes(year)) + geom_line(aes(y = total)) +
    ggtitle("present data")
grid.arrange(plot1, plot2)
```

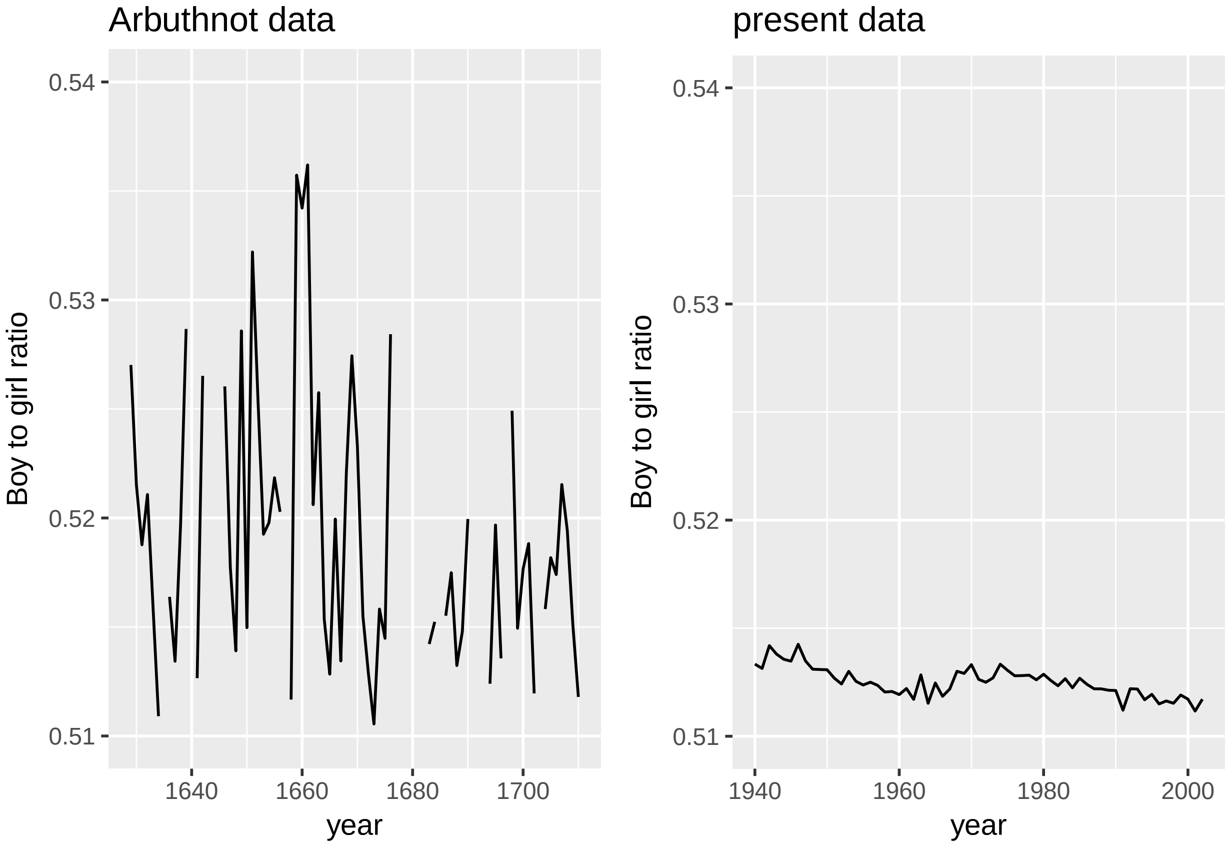


The plots show that there was a large population dip between 1640 and 1660 in the arbuthnot data. This also appears to happen in the 1970s in the present data; however, the present day data always has much larger counts than the arbuthnot data.

Question - 3 Make a plot that displays the boy-to-girl ratio for every year in the data set. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response.

I will create a new column that calculates the ratio of boys-to-girls for each year. I'll also make the same column for the arbuthnot data.

```
present$BGRatio <- present$boys/(present$girls + present$boys)
arbuthnot$BGRatio <- arbuthnot$boys/(arbuthnot$girls + arbuthnot$boys)
plot1 <- ggplot(arbuthnot, aes(year)) + geom_line(aes(y = BGRatio)) +
    ggtitle("Arbuthnot data") + scale_y_continuous(limits = c(0.51,
    0.54), breaks = seq(0.51, 0.54, by = 0.01)) + labs(y = "Boy to girl ratio")
plot2 <- ggplot(present, aes(year)) + geom_line(aes(y = BGRatio)) +
    ggtitle("present data") + scale_y_continuous(limits = c(0.51,
    0.54), breaks = seq(0.51, 0.54, by = 0.01)) + labs(y = "Boy to girl ratio")
grid.arrange(plot1, plot2, ncol = 2)
```



The present day data appears to show that a boy-to-girl ratio of greater than 0.5 is consistently reported every year. This data suggests that Arbuthnot's theory may be valid but further analysis would be required since the average present day ratio is so close to the value of 0.5 (which would mean Arbuthnot's theory is not supported by this data). Also, I have set the y-scale equal between the side by side graphs to highlight how much the variability between years is reduced in the present day data. This is most likely due to the much larger sample size which better represents the population.

Question 4 - In what year did we see the most total number of births in the U.S.? You can refer to the help files or the R reference card http://cran.r-project.org/doc/contrib/Short-refcard.pdf to find helpful commands.

This action can be done in one line using the which.max() function.

```
present[which.max(present[, "total"]), ]
```

```
##    year    boys   girls   total   BGRatio
## 22 1961 2186274 2082052 4268326 0.5122088
```