da420\_assignment0\_grahn

jason grahn

9/20/2018

Handout 0 looks like it’s mostly in base-R, so we can jump right in with the first provided command.

# Part1

source("http://www.openintro.org/stat/data/arbuthnot.R")

## Exercise 1

What command would you use to extract just the counts of girls baptized? ###Solution 1

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

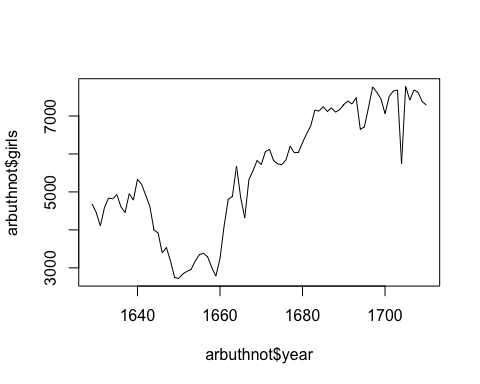
#which provides a pretty poor display so we can use tidy-style with   
select(arbuthnot, girls)

## girls  
## 1 4683  
## 2 4457  
## 3 4102  
## 4 4590  
## 5 4839  
## 6 4820  
## 7 4928  
## 8 4605  
## 9 4457  
## 10 4952  
## 11 4784  
## 12 5332  
## 13 5200  
## 14 4910  
## 15 4617  
## 16 3997  
## 17 3919  
## 18 3395  
## 19 3536  
## 20 3181  
## 21 2746  
## 22 2722  
## 23 2840  
## 24 2908  
## 25 2959  
## 26 3179  
## 27 3349  
## 28 3382  
## 29 3289  
## 30 3013  
## 31 2781  
## 32 3247  
## 33 4107  
## 34 4803  
## 35 4881  
## 36 5681  
## 37 4858  
## 38 4319  
## 39 5322  
## 40 5560  
## 41 5829  
## 42 5719  
## 43 6061  
## 44 6120  
## 45 5822  
## 46 5738  
## 47 5717  
## 48 5847  
## 49 6203  
## 50 6033  
## 51 6041  
## 52 6299  
## 53 6533  
## 54 6744  
## 55 7158  
## 56 7127  
## 57 7246  
## 58 7119  
## 59 7214  
## 60 7101  
## 61 7167  
## 62 7302  
## 63 7392  
## 64 7316  
## 65 7483  
## 66 6647  
## 67 6713  
## 68 7229  
## 69 7767  
## 70 7626  
## 71 7452  
## 72 7061  
## 73 7514  
## 74 7656  
## 75 7683  
## 76 5738  
## 77 7779  
## 78 7417  
## 79 7687  
## 80 7623  
## 81 7380  
## 82 7288

#or with...  
arbuthnot %>%   
 select(girls)

## girls  
## 1 4683  
## 2 4457  
## 3 4102  
## 4 4590  
## 5 4839  
## 6 4820  
## 7 4928  
## 8 4605  
## 9 4457  
## 10 4952  
## 11 4784  
## 12 5332  
## 13 5200  
## 14 4910  
## 15 4617  
## 16 3997  
## 17 3919  
## 18 3395  
## 19 3536  
## 20 3181  
## 21 2746  
## 22 2722  
## 23 2840  
## 24 2908  
## 25 2959  
## 26 3179  
## 27 3349  
## 28 3382  
## 29 3289  
## 30 3013  
## 31 2781  
## 32 3247  
## 33 4107  
## 34 4803  
## 35 4881  
## 36 5681  
## 37 4858  
## 38 4319  
## 39 5322  
## 40 5560  
## 41 5829  
## 42 5719  
## 43 6061  
## 44 6120  
## 45 5822  
## 46 5738  
## 47 5717  
## 48 5847  
## 49 6203  
## 50 6033  
## 51 6041  
## 52 6299  
## 53 6533  
## 54 6744  
## 55 7158  
## 56 7127  
## 57 7246  
## 58 7119  
## 59 7214  
## 60 7101  
## 61 7167  
## 62 7302  
## 63 7392  
## 64 7316  
## 65 7483  
## 66 6647  
## 67 6713  
## 68 7229  
## 69 7767  
## 70 7626  
## 71 7452  
## 72 7061  
## 73 7514  
## 74 7656  
## 75 7683  
## 76 5738  
## 77 7779  
## 78 7417  
## 79 7687  
## 80 7623  
## 81 7380  
## 82 7288

plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l")



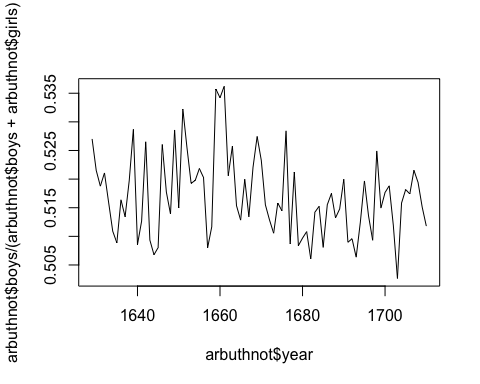
## Exercise 2

Is there an apparent trend in the number of girls baptized over the years? How would you describe it? ###solution 2 Yes, there appears to be an upward trend in the number of girls baptized over the years for the window of time for which we have data. It drops between 1640 and 1660, but increases sharply thereafter and continues to grow.

## Exercise 3

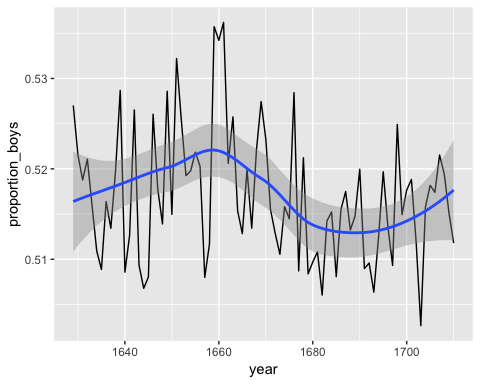
Make a plot of the proportion of boys over time. What do you see? ###Solution 3 First, the plot:

plot(arbuthnot$year, arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls), type = "l")



#but this is really hard to understand / read. So if we use tidyverse...  
arbuthnot %>%  
 mutate(proportion\_boys = boys / (boys + girls)) %>%  
 ggplot(aes(year, proportion\_boys)) +  
 geom\_line() +   
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



# Part 2

Repeat all commands with present day birth records.

source("http://www.openintro.org/stat/data/present.R")

## Questions:

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?

head(present, 5)

## year boys girls  
## 1 1940 1211684 1148715  
## 2 1941 1289734 1223693  
## 3 1942 1444365 1364631  
## 4 1943 1508959 1427901  
## 5 1944 1435301 1359499

tail(present, 5)

## year boys girls  
## 59 1998 2016205 1925348  
## 60 1999 2026854 1932563  
## 61 2000 2076969 1981845  
## 62 2001 2057922 1968011  
## 63 2002 2057979 1963747

dim(present)

## [1] 63 3

The data covers 1940 through 2002. The dimensions are 63 observations by 3 variables. The 3 variables are *year*, *boys*, *girls*.

1. How do these counts compare to Arbuthnot’s? Are they on a similar scale?

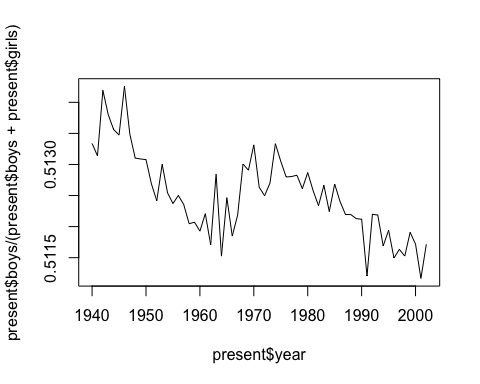
summary(head(arbuthnot$girls + arbuthnot$boys, 40) /   
 head(present$girls + present$boys, 40)   
 \* 100)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1315 0.1756 0.2624 0.2508 0.3159 0.4195

The *Arbuthnot* data is approximately 0.25% of the *Present* data, on average, for a given year. The *Present* data is on a scale much more massive than the *Arbuthnot* data. A likely way to compare these would be log-scale.

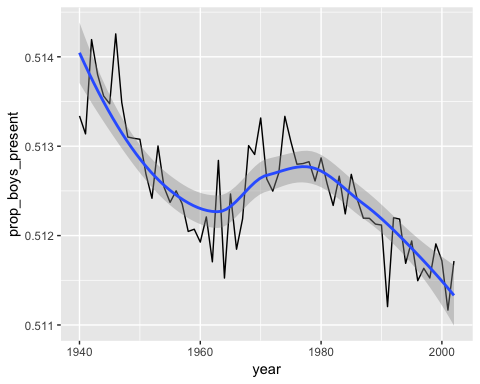
1. Make a plot that displays the boy-to-girl ratio for every year in the data set. What do you see? Does Arbuthnot’s observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response.

plot(present$year, present$boys / (present$boys + present$girls), type = "l")



#The joy of the tidyverse is not JUST more readable code, but less code duplication.  
present %>%  
 mutate(prop\_boys\_present = boys / (boys + girls)) %>%  
 ggplot(aes(year, prop\_boys\_present)) +  
 geom\_line() +  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



We see the proportion of boys is generally around the same values in the Present data that we do in the Arbuthnot data. In contrast to the Arbuthnot data, we see less variation in Present data *and* we see a significant downward trend. The Arbuthnot data showed larger variation and a generally flat trend.

1. 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.

present %>%   
 mutate(total = girls + boys) %>%   
 filter(total == max(total))

## year boys girls total  
## 1 1961 2186274 2082052 4268326

The highest year was 1961.