DATA 606 - Lab0

Joshua Sturm

August 28, 2017

Note: I spent some time trying to find a way to compact the data to speed up the knitting process. I came across the `DataTables` library, which I think is amazing, and I've incorporated throughout my lab.

Load the arbuthnot dataset

```
source("more/arbuthnot.R")
```

```
library(DT)
datatable(arbuthnot, extensions = 'Scroller', options = list(
  deferRender = TRUE,
  scrollY = 300,
  scroller = TRUE
))
```

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

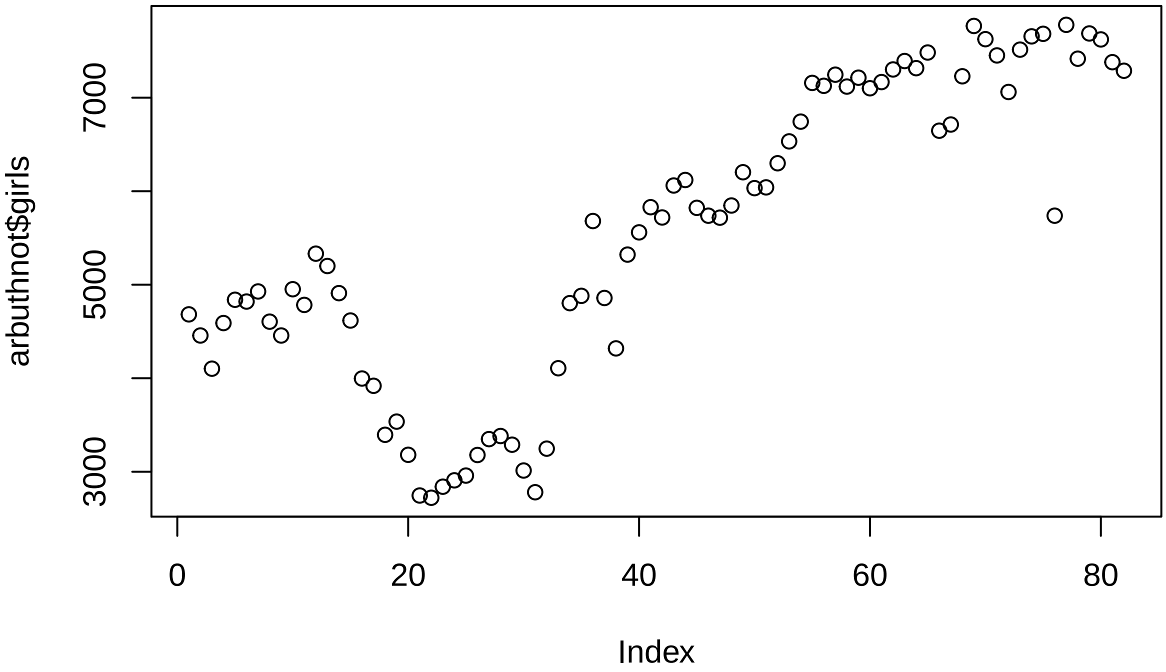
```
arbuthnot$girls
```

```
##  [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
```

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

```
To better illustrate the trend, I will plot it using
```

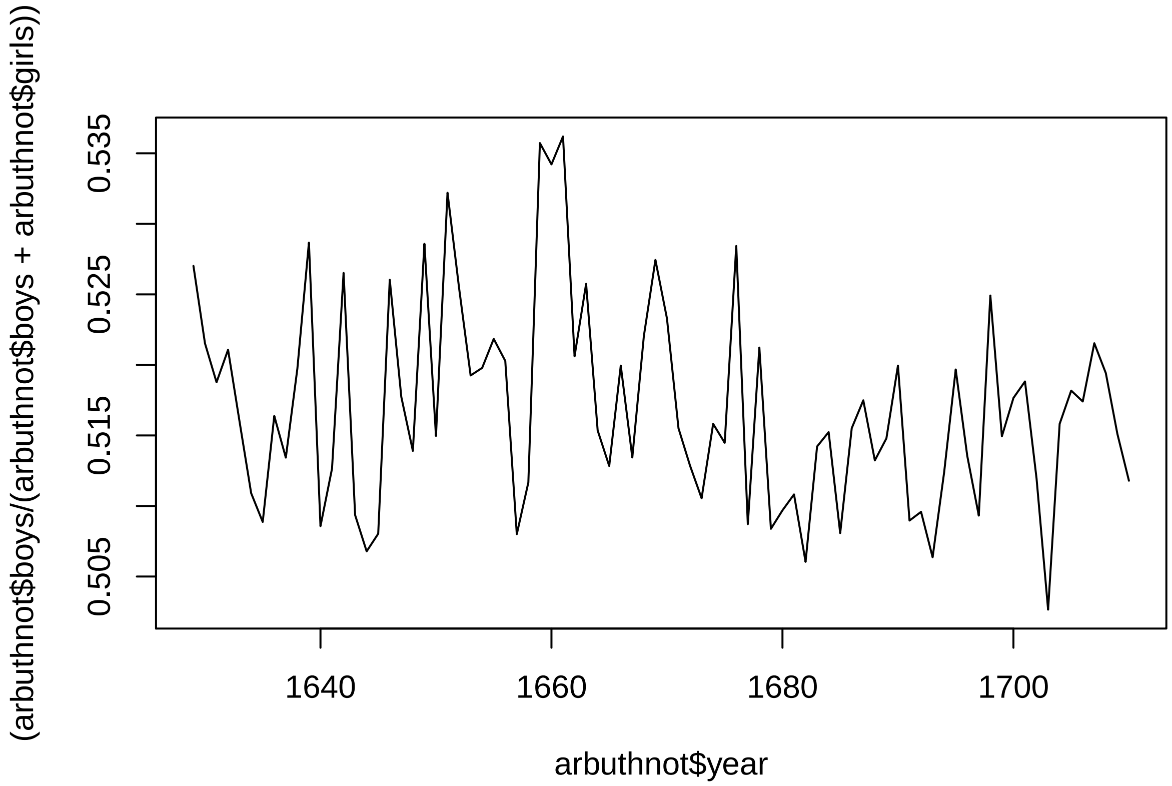
```
plot(arbuthnot$girls)
```



```
The number of girls being baptized increases extremely quickly from 1660 through 1700.
```

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.

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



```
The proportion of boys baptized decreases between 1629 and 1710. However, the graph is always above 0.5
```

On Your Own

Load the present dataset:

```
source("more/present.R")
```

a. What years are included in this data set? What are the dimensions of the data frame and what are the variable or column names?

Load the data in the file "present.R"

```
    library(DT)
datatable(present, extensions = 'Scroller', options = list(
  deferRender = TRUE,
  scrollY = 300,
  scroller = TRUE
))
```

```
present$year
```

```
##  [1] 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953
## [15] 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967
## [29] 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981
## [43] 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
## [57] 1996 1997 1998 1999 2000 2001 2002
```

We can see the file has values for the year 1940-2002.

Alternatively, we can find the exact values with the min and max functions.

Get the dimensions

```
dim(present)
```

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

There are 63 year's worth of data, and three variables.

Get the column names

```
names(present)
```

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

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

If we compare the sums of the two variables in both data sets, we can see how much we're dealing with in

```
arbuthnot$girls + arbuthnot$boys
```

```
##  [1]  9901  9315  8524  9584  9997  9855 10034  9522  9160 10311 10150
## [12] 10850 10670 10370  9410  8104  7966  7163  7332  6544  5825  5612
## [23]  6071  6128  6155  6620  7004  7050  6685  6170  5990  6971  8855
## [34] 10019 10292 11722  9972  8997 10938 11633 12335 11997 12510 12563
## [45] 11895 11851 11775 12399 12626 12601 12288 12847 13355 13653 14735
## [56] 14702 14730 14694 14951 14588 14771 15211 15054 14918 15159 13632
## [67] 13976 14861 15829 16052 15363 14639 15616 15687 15448 11851 16145
## [78] 15369 16066 15862 15220 14928
```

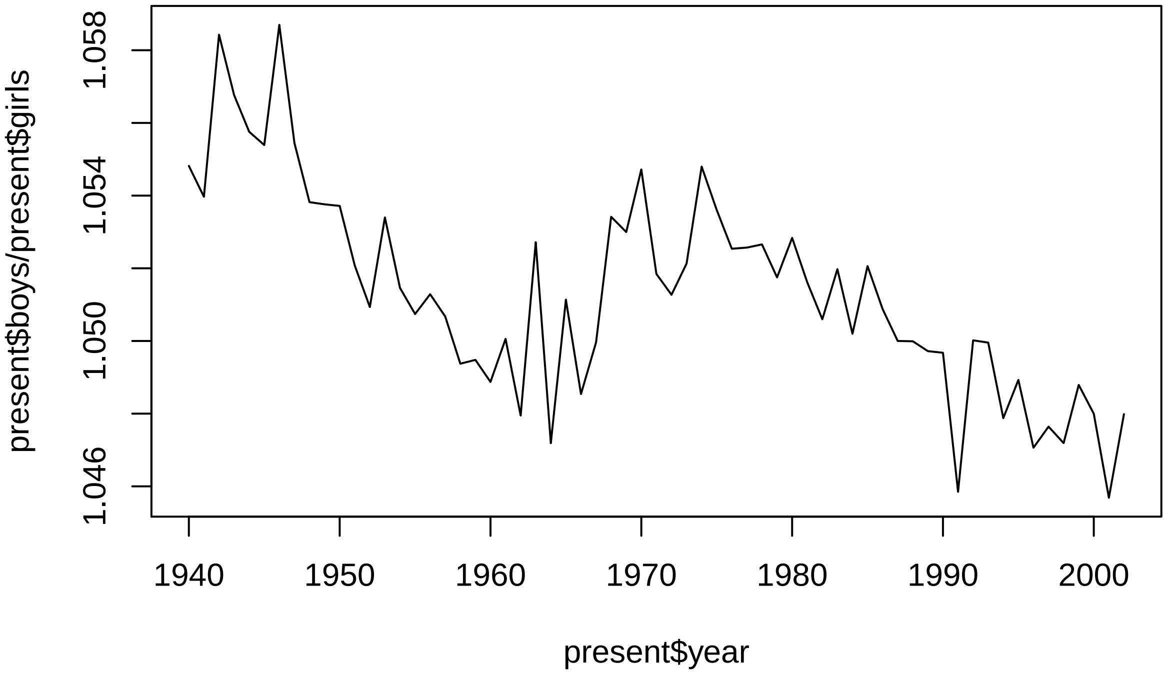
```
present$girls + present$boys
```

```
##  [1] 2360399 2513427 2808996 2936860 2794800 2735456 3288672 3699940
##  [9] 3535068 3559529 3554149 3750850 3846986 3902120 4017362 4047295
## [17] 4163090 4254784 4203812 4244796 4257850 4268326 4167362 4098020
## [25] 4027490 3760358 3606274 3520959 3501564 3600206 3731386 3555970
## [33] 3258411 3136965 3159958 3144198 3167788 3326632 3333279 3494398
## [41] 3612258 3629238 3680537 3638933 3669141 3760561 3756547 3809394
## [49] 3909510 4040958 4158212 4110907 4065014 4000240 3952767 3899589
## [57] 3891494 3880894 3941553 3959417 4058814 4025933 4021726
```

We can see that the present data set does not have less than a seven figure sum, whereas the arbuthnot

c. 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$girls, type="l")
```



Like Arbuthnot's data, there were more boys (being born, in this case) in the U.s. than girls. The ratio declined over the observed years, but is much higher in the present data set.

d. In what year did we see the most total number of births in the U.S.?

```
    which.max(present$boys+present$girls)
```

```
## [1] 22
```

```
This returns the value '22'. To find the corresponding year, we can do
```

```
present[22,c(1)]
```

```
## [1] 1961
```

```
which gives us the year 1961.
```

To find the actual value of the max, we can use the command

```
    sum(present[22,c(2,3)])
```

```
## [1] 4268326
```