Godbey-0.Rmd

Robert Godbey

February 2, 2016

Contents

Data Exploration

First we need to create the datasets we want to examine. We do that here with the programs that were provided. This allows my R Markdown file to work from anywhere.

```
arbuthnot <-
structure(list(year = 1629:1710, boys = c(5218L, 4858L, 4422L,
4994L, 5158L, 5035L, 5106L, 4917L, 4703L, 5359L, 5366L, 5518L,
5470L, 5460L, 4793L, 4107L, 4047L, 3768L, 3796L, 3363L, 3079L,
2890L, 3231L, 3220L, 3196L, 3441L, 3655L, 3668L, 3396L, 3157L,
3209L, 3724L, 4748L, 5216L, 5411L, 6041L, 5114L, 4678L, 5616L,
6073L, 6506L, 6278L, 6449L, 6443L, 6073L, 6113L, 6058L, 6552L,
6423L, 6568L, 6247L, 6548L, 6822L, 6909L, 7577L, 7575L, 7484L,
7575L, 7737L, 7487L, 7604L, 7909L, 7662L, 7602L, 7676L, 6985L,
7263L, 7632L, 8062L, 8426L, 7911L, 7578L, 8102L, 8031L, 7765L,
6113L, 8366L, 7952L, 8379L, 8239L, 7840L, 7640L), girls = c(4683L,
4457L, 4102L, 4590L, 4839L, 4820L, 4928L, 4605L, 4457L, 4952L,
4784L, 5332L, 5200L, 4910L, 4617L, 3997L, 3919L, 3395L, 3536L,
3181L, 2746L, 2722L, 2840L, 2908L, 2959L, 3179L, 3349L, 3382L,
3289L, 3013L, 2781L, 3247L, 4107L, 4803L, 4881L, 5681L, 4858L,
4319L, 5322L, 5560L, 5829L, 5719L, 6061L, 6120L, 5822L, 5738L,
5717L, 5847L, 6203L, 6033L, 6041L, 6299L, 6533L, 6744L, 7158L,
7127L, 7246L, 7119L, 7214L, 7101L, 7167L, 7302L, 7392L, 7316L,
7483L, 6647L, 6713L, 7229L, 7767L, 7626L, 7452L, 7061L, 7514L,
```

```
7656L, 7683L, 5738L, 7779L, 7417L, 7687L, 7623L, 7380L, 7288L
)), .Names = c("year", "boys", "girls"), class = "data.frame", row.names = c(NA,
-82L))

`present` <-
structure(list(year = c(1940, 1941, 1942, 1943, 1944, 1945, 1946,
1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957,
1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968,
1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979,
1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990,
1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001,
2002), boys = c(1211684, 1289734, 1444365, 1508959, 1435301,
1404587, 1691220, 1899876, 1813852, 1826352, 1823555, 1923020,
1971262, 2001798, 2059068, 2073719, 2133588, 2179960, 2152546,
2173638, 2179708, 2186274, 2132466, 2101632, 2060162, 1927054,
1845862, 1803388, 1796326, 1846572, 1915378, 1822910, 1669927,
1608326, 1622114, 1613135, 1624436, 1705916, 1709394, 1791267,
1852616, 1860272, 1885676, 1865553, 1879490, 1927983, 1924868,
1951153, 2002424, 2069490, 2129495, 2101518, 2082097, 2048861,
2022589, 1996355, 1990480, 1985596, 2016205, 2026854, 2076969,
2057922, 2057979), girls = c(1148715, 1223693, 1364631, 1427901,
1359499, 1330869, 1597452, 1800064, 1721216, 1733177, 1730594,
1827830, 1875724, 1900322, 1958294, 1973576, 2029502, 2074824,
2051266, 2071158, 2078142, 2082052, 2034896, 1996388, 1967328,
1833304, 1760412, 1717571, 1705238, 1753634, 1816008, 1733060,
1588484, 1528639, 1537844, 1531063, 1543352, 1620716, 1623885,
1703131, 1759642, 1768966, 1794861, 1773380, 1789651, 1832578,
1831679, 1858241, 1907086, 1971468, 2028717, 2009389, 1982917,
1951379, 1930178, 1903234, 1901014, 1895298, 1925348, 1932563,
1981845, 1968011, 1963747)), .Names = c("year", "boys", "girls"
), row.names = c(NA, 63L), class = "data.frame")
```

What years are included in this data set?

Answer: 1940 to 2002

If we type present at the R prompt the dataframe is listed to the screen. It flys by so fast it is difficult to see the top of the dataframe, but we can see the bottom (row 48 to 63 on my screen). The last year is 2002. We can scroll up to see row 1 and the date 1940. We can also use the head function to display the first 6 rows and the tail function to show the last 6 rows.

```
present
```

```
##    year    boys   girls
## 1  1940 1211684 1148715
## 2  1941 1289734 1223693
## 3  1942 1444365 1364631
## 4  1943 1508959 1427901
## 5  1944 1435301 1359499
## 6  1945 1404587 1330869
## 7  1946 1691220 1597452
## 8  1947 1899876 1800064
## 9  1948 1813852 1721216
```

```
## 10 1949 1826352 1733177
## 11 1950 1823555 1730594
## 12 1951 1923020 1827830
## 13 1952 1971262 1875724
## 14 1953 2001798 1900322
## 15 1954 2059068 1958294
## 16 1955 2073719 1973576
## 17 1956 2133588 2029502
## 18 1957 2179960 2074824
## 19 1958 2152546 2051266
## 20 1959 2173638 2071158
## 21 1960 2179708 2078142
## 22 1961 2186274 2082052
## 23 1962 2132466 2034896
## 24 1963 2101632 1996388
## 25 1964 2060162 1967328
## 26 1965 1927054 1833304
## 27 1966 1845862 1760412
## 28 1967 1803388 1717571
## 29 1968 1796326 1705238
## 30 1969 1846572 1753634
## 31 1970 1915378 1816008
## 32 1971 1822910 1733060
## 33 1972 1669927 1588484
## 34 1973 1608326 1528639
## 35 1974 1622114 1537844
## 36 1975 1613135 1531063
## 37 1976 1624436 1543352
## 38 1977 1705916 1620716
## 39 1978 1709394 1623885
## 40 1979 1791267 1703131
## 41 1980 1852616 1759642
## 42 1981 1860272 1768966
## 43 1982 1885676 1794861
## 44 1983 1865553 1773380
## 45 1984 1879490 1789651
## 46 1985 1927983 1832578
## 47 1986 1924868 1831679
## 48 1987 1951153 1858241
## 49 1988 2002424 1907086
## 50 1989 2069490 1971468
## 51 1990 2129495 2028717
## 52 1991 2101518 2009389
## 53 1992 2082097 1982917
## 54 1993 2048861 1951379
## 55 1994 2022589 1930178
## 56 1995 1996355 1903234
## 57 1996 1990480 1901014
## 58 1997 1985596 1895298
## 59 1998 2016205 1925348
## 60 1999 2026854 1932563
## 61 2000 2076969 1981845
## 62 2001 2057922 1968011
## 63 2002 2057979 1963747
```

```
head(present)
```

```
##   year    boys   girls
## 1 1940 1211684 1148715
## 2 1941 1289734 1223693
## 3 1942 1444365 1364631
## 4 1943 1508959 1427901
## 5 1944 1435301 1359499
## 6 1945 1404587 1330869
```

```
tail(present)
```

```
##    year    boys   girls
## 58 1997 1985596 1895298
## 59 1998 2016205 1925348
## 60 1999 2026854 1932563
## 61 2000 2076969 1981845
## 62 2001 2057922 1968011
## 63 2002 2057979 1963747
```

What are the dimensions of the data frame?

Answer: 63 observations (rows) of 3 variables (columns)

We can look in the global environment window (upper right) and see 63 obs. of 3 variables, or we can run the dim function on present. I also like the structure or str function for looking at data.

```
dim(present)
```

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

```
str(present)
```

```
## 'data.frame':    63 obs. of  3 variables:
##  $ year : num  1940 1941 1942 1943 1944 ...
##  $ boys : num  1211684 1289734 1444365 1508959 1435301 ...
##  $ girls: num  1148715 1223693 1364631 1427901 1359499 ...
```

And, what are the variable or column names?

Answer: year, boys, girls

The listings above (dataframe, head, tail) and str() all listed the column names.

Data Comparison

How do these counts compare to Arbuthnot's?

One quick way to get some comparsion info is to use the summary function. We can see the results on both data sets below.

```r
summary(arbuthnot)
```

```
##       year           boys          girls
##  Min.   :1629   Min.   :2890   Min.   :2722
##  1st Qu.:1649   1st Qu.:4759   1st Qu.:4457
##  Median :1670   Median :6073   Median :5718
##  Mean   :1670   Mean   :5907   Mean   :5535
##  3rd Qu.:1690   3rd Qu.:7576   3rd Qu.:7150
##  Max.   :1710   Max.   :8426   Max.   :7779
```

```r
summary(present)
```

```
##       year           boys             girls
##  Min.   :1940   Min.   :1211684   Min.   :1148715
##  1st Qu.:1956   1st Qu.:1799857   1st Qu.:1711405
##  Median :1971   Median :1924868   Median :1831679
##  Mean   :1971   Mean   :1885600   Mean   :1793915
##  3rd Qu.:1986   3rd Qu.:2058524   3rd Qu.:1965538
##  Max.   :2002   Max.   :2186274   Max.   :2082052
```

Are they on a similar scale?

Answer: No, the Present counts are 3 orders of magnitude higher

Arbuthnot's count data is in thousands as in the maximum number of boys being 8.4 thousand and the maximum number of girls being 7.8 thousand. The more recent US data is in millions, as in 2.2 million boys born one year. This is 3 orders of magnitude higher, so they are not on the same scale.

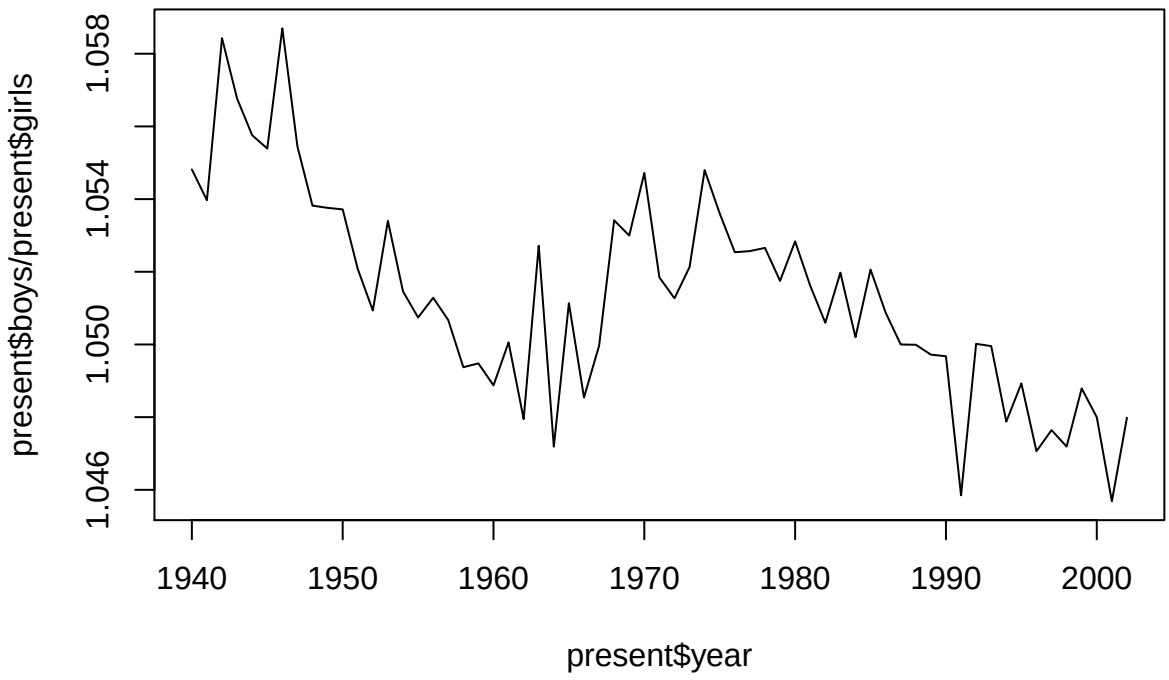
The ratio of boys to girls is also different. Arbuthnot's data had a minimum ratio of 1.011 to a maximum of 1.156. The Present data set ran 1.046 to 1.059 (max(present-boys / present-girls)). Although the Present Boy-to-Girl ratio never gets as low as Arbuthnot's the difference never gets as high. The number of boys born in the US each year is close to the the number of girls (and seems to be getting closer, see below).

Data Analysis (and plotting)

Make a plot that displays the boy-to-girl ratio for every year in the data set.

Using the approach from the lab example I plotted boys divided by girls by year.

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



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.

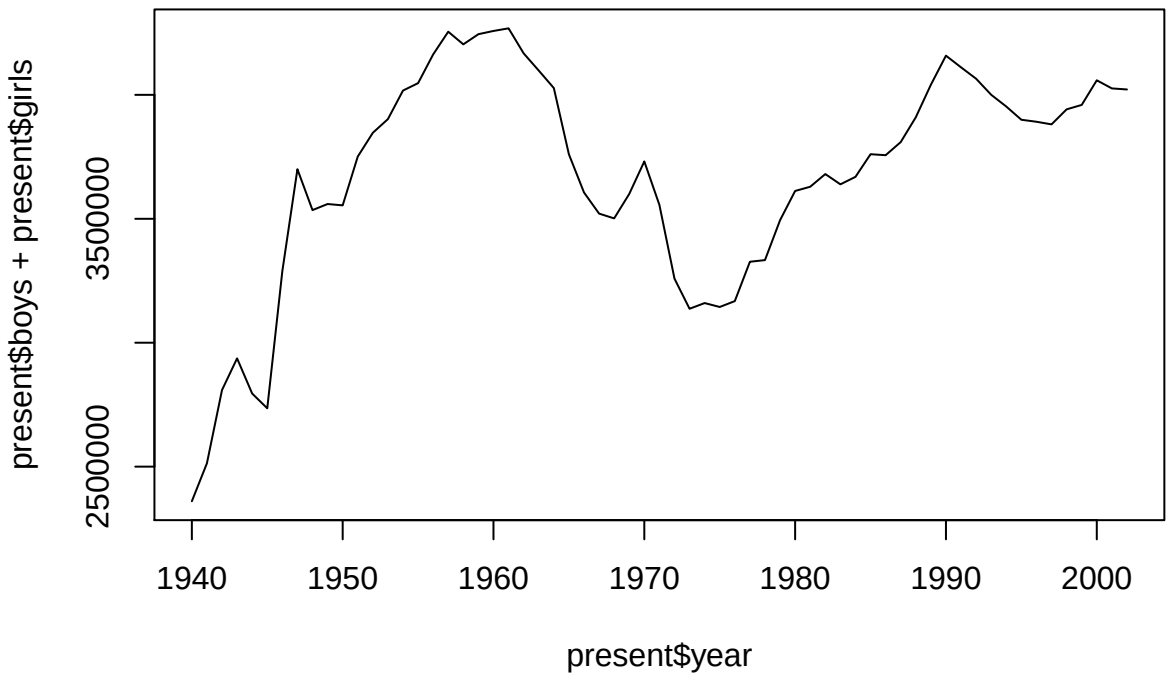
Our proportion would be greater than 1 when more boys are born, equal to 1 when they are the same, and less than 1 when more girls are born. Our plot shows values greater than one for each year (~1.058 to 1.046), so more boys are born each year. However, the trend of the graph is down and means the difference is getting smaller.

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

Answer: 1961

We can do this a few ways. Plotting the sum of boys and girls by year gives us a quick visual way of seeing it is near 1960. Looking at the present dataframe we can see that the totals around 1960 are as follows in the table below. The year 1961 has the highest total births.

year	boys	girls	total
1959	2,173,638	2,071,158	4,244,796
1960	2,179,708	2,078,142	4,257,850
1961	2,186,274	2,082,052	4,268,326
1962	2,132,466	2,034,896	4,167,362


present$boys + present$girls
2500000
3500000
1940
1950
1960
1970
1980
1990
2000
present$year