Intro_to_R

SMWadgymar, KGSmith

2022-08-15

This is an R markdown document. These documents are ideal for combining chunks of code embedded within notes and other text.

We will be working with a dataset called present, which includes the number of individuals assigned a sex of boy or girl at birth in the United States from 1940-2002.

1) First, let's load the present data. Note that the code below is written in a "chunk", which surrounds your code with two lines that each contain three tickbacks. After the first set of backticks, the code {r} is needed to convey that the subsequent lines contain a chunk of code. You can run this code in a few ways. One way is to click on the right-pointing arrow on the right side of the chunk. Another way is to place your cursor anywhere within the line that you want to run and click Ctrl+Enter for Windows or Cmd+Return for Macs.

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

You should now see the present dataset listed in the Environment tab in the top right window of RStudio. It is a datafile with 63 observations and 3 variables. Note that it is labeled as present in R, and because it has a label, we can refer to it in our code.

2) Let's look at the first six rows of the dataset. We'll use the head() function.

```
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
```

Q: How many columns are there? What do the columns represent? What do the rows represent?

A: There are three columns: year, boys, and girls. Each row includes data for a different year.

3) Now let's look at the last six rows in the dataset using the tail() function. Replace the ??? with the appropriate code before running.

```
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
```

4) Let's explore the range of years included in this dataset. We will be referring to a specific column within the present dataset, which we can do using the $ sign with the following format:

dataset_name$variable_name

Use the min(), max(), and range() functions to examine the range of years present in the dataset.

```
min(present$year)
```

```
## [1] 1940
```

```
max(present$year)
```

```
## [1] 2002
```

```
range(present$year)
```

```
## [1] 1940 2002
```

Q: Why do you think that min(), max(), and range() are more reliable for determining the span of years present in the dataset than head() and tail()?

A: A dataset may not include data in order by year, which means the head() and tail() functions might now display the first and last year in the dataset. The range() function will return the minimum and maximum values regardless of what order they are in.

5) Let's see if the ratio of babies assigned as boys or girls at birth has changed over time.

We can easily create and name new variables in R. For example, let's make the variables A and B and save their sum as C.

```
A<-5

B<-3

C<-A+B
```

We can also do this with vectors of numbers. Note that R requires that vectors be identified using the c() notation.

```
X<-c(A, B)
X
```

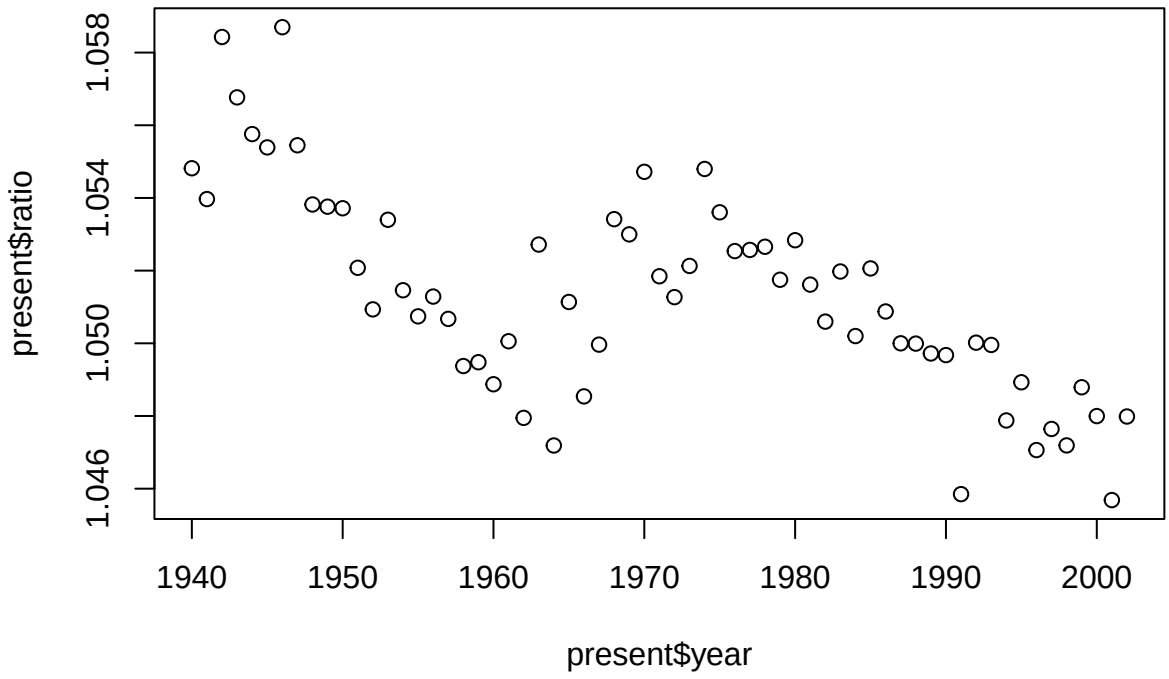
```
## [1] 5 3
```

To calculate the ratio of boys and girls in the present datset, we can divide one column of data (boys) by another column of data (girls) and saving the output as a new variable in present (which we will call ratio).

```
present$ratio = present$boys / present$girls
```

6) Let's make a plot of the ratio over time. We can do this with the plot() function using the notation plot(x, y), where the x and y arguments are where you will specify which variable you want displayed along the x and y axes.

```
plot(present$year, present$ratio)
```



Q: Has the ratio of babies assigned boy vs. girl at birth changed over time?

A: All values are slightly above 1, which shows that slightly more boys are born than girls each year. However, the ratio seems to be decreasing over time, although it may be cyclical. It would be interesting to see the past 20 years of data.

7) Let's customize our graph of ration over time. Here is an explanation of the arguments:

xlim=c(): set the range of the x-axis. Within the c(), list the min and max of the range separated by a comma.

ylim=c(): set the range of the y-axis. Within the c(), list the min and max of the range separated by a comma.

xlab=" ": specify the x-axis label. This must be in quotes.

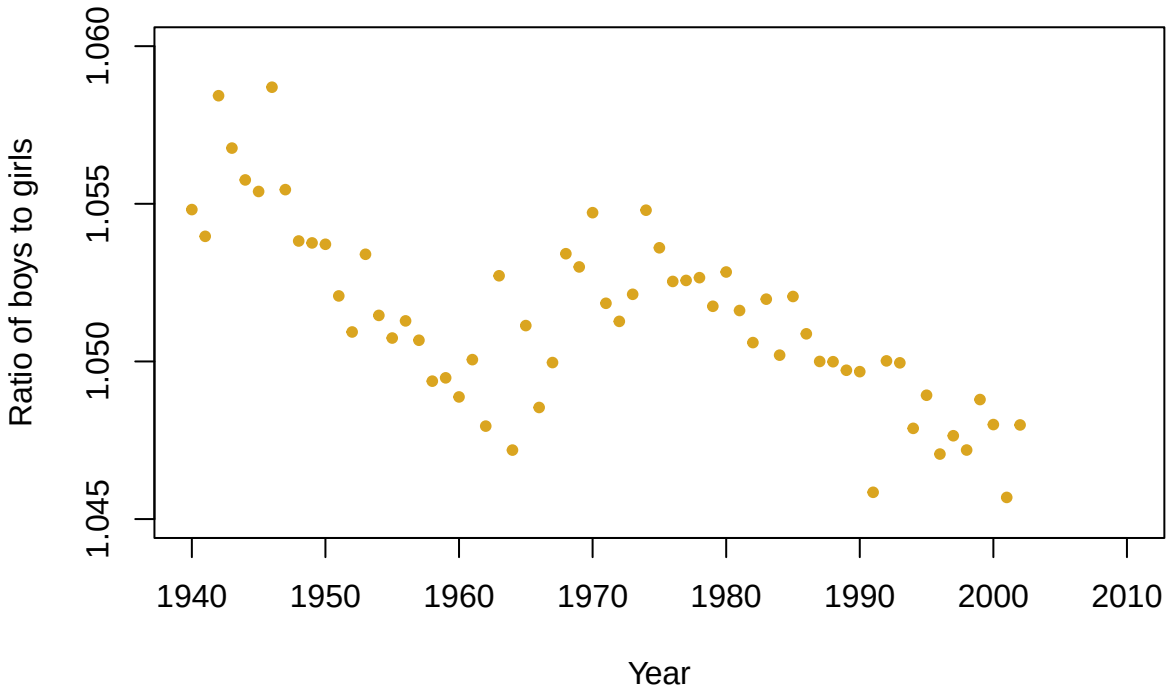
ylab=" ": specify the y-axis label. This must be in quotes.

main=" ": specify the title of the graph. This must be in quotes. To remove a title that is automatically displayed, leave an empty space between the quotes.

col=" ": specify the color of the points, lines, bars, boxplots, etc. If you are naming a color, it must be in quotes. To list more than one color, you need to list them within c() and separated by a comma, with each color within quotes. To see a list of some (but not all) of colors you can specify in R, see this image: https://derekogle.com/NCGraphing/img/colorbynames.png

pch=: only for scatterplots, you can specify the symbol plotted with a number or symbol as you see in this image: https://www.statmethods.net/advgraphs/images/points.png

```
plot(present$year, present$ratio, xlim=c(1940, 2010), ylim=c(1.045, 1.06),
     xlab="Year", ylab="Ratio of boys to girls", main=" ", col="goldenrod", pch=20)
```



8) Let's practice turning this R markdown document into a PDF file as if you were submitting it as an assignment. Click on the drop down arrow next to the Knit button above, which has a symbol of yarn and a knitting needle. Select "Knit to PDF". Note that this won't work if you have any problematic code in your R markdown document. To check for issues, scroll up your R markdown file and make sure that you don't see any red Xs next to any of the code chunks.