HW 1

Mike Lopez

September 2017

General instructions for homeworks:

- Make a new R Markdown file (.Rmd) referring to thea assignment on the course Github page
- Change the heading to include your author name
- Save the R Markdown file (named as: [MikeID]-[Homework01].Rmd – e.g. "mlopez-Lab01.Rmd") to somewhere where you'll be able to access it later (zip drive, My Documents, Dropbox, etc)
- Your file should contain the code/commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file
- **Each answer must be supported by written statements (unless otherwise specified) as well as any code used**: In other words, if the answer is 24, you should write "The answer is 24" (as opposed to just showing the code and output).
- Include the names of anyone you collaborated with at the top of the assignment

Part I:

Open Intro 1.1, 1.4, 1.6, 1.10, 1.14,1.15, 1.17, 1.26, 1.27 (Note: these are located in section 1.8 at the end of the chapter)

1.1:

a. 10/43 in the treatment group (23%) versus 2/46 in the control (4%)

b. This represents a 19% difference, which seems meaningful. At first glance, the treatment seems more effective than the control

c. Answers can vary, but should make sense. Ex: This is a relatively large difference, and I believe the difference to be significant. Ex: With a relatively small number of subjects, this difference could just be due to chance.

1.4

a. The cases are the 706 adults living near NYC (504 white, 202 black). Age (discrete, continuous), sex (categorical, regular/nominal), and ethnicity (categorical, regular/nominal) were measured, as were weight and height (both continuous). Body fat percentage was the outcome (continuous). The research question is how BMI correlates to body fat percentage, and whether or not that differs within genders and ethnicities.

b. The cases were the 129 undergrads. Categorical variables included class, education, and job respect. The number of candies taken was the outcome variable (continuous, discrete). The research question is whether or not individuals who associate themselves with higher social classes are unethical, as judged by taking of candy that was meant for children.

1.10

a. Explanatory: percent with bachelor's, per capita income
b. There is a positive link – roughly linear in shape – with a bit more variability in income among those with higher bachelor degree percents. There do not appear ato be any unusual observations
c. No: observational data

Part II:

Lab 1: intro to R questions 4 - 7

4. What years are included in this data set? What are the dimensions of the data frame? What are the variable (column) names?

```
library(oilabs)
data(present)
dim(present)
```

```
## [1] 74  3
```

```
glimpse(present)
```

```
## Observations: 74
## Variables: 3
## $ year  <dbl> 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 19...
## $ boys  <dbl> 1211684, 1289734, 1444365, 1508959, 1435301, 1404587, 16...
## $ girls <dbl> 1148715, 1223693, 1364631, 1427901, 1359499, 1330869, 15...
```

```
tail(present)
```

```
## # A tibble: 6 x 3
##    year    boys   girls
##   <dbl>   <dbl>   <dbl>
## 1  2008 2173625 2074069
## 2  2009 2113739 2016926
## 3  2010 2046561 1952825
## 4  2011 2024068 1929522
## 5  2012 2021800 1931041
## 6  2013 2013108 1919073
```

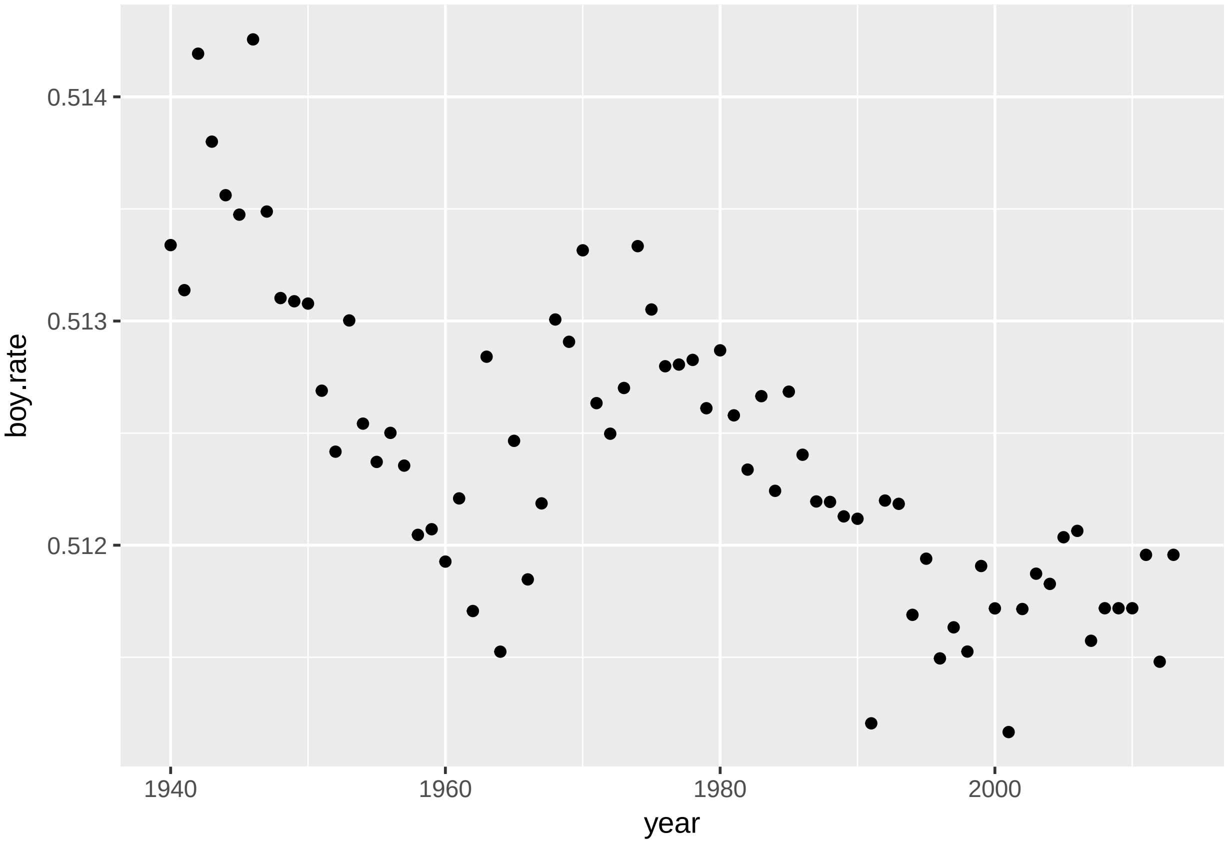
The `present` data set includes 1940-2013 births, and has 74 rows/3 columns. The variables are `year`, `boys`, and `girls`

5. How do these counts compare to Arbuthnot's? Are they of a similar magnitude?

These are much larger counts than Arbuthnots, but of similar ratio.

6. Make a plot that displays the proportion of boys born over time. 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. *Hint:* You should be able to reuse your code from Ex 3 above, just replace the dataframe name.

```
present <- present %>% mutate(boy.rate = (boys)/(boys + girls))
qplot(year, boy.rate, data = present)
```



Ratio of boys is greater than 0.5, but is slowly descreasing with time. A slight spike in boys in 1970

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

```
present %>% mutate(total = boys + girls) %>% arrange(-total) %>% head()
```

```
## # A tibble: 6 x 5
##    year    boys   girls boy.rate   total
##   <dbl>   <dbl>   <dbl>    <dbl>   <dbl>
## 1  2007 2208071 2108162 0.5115736 4316233
## 2  1961 2186274 2082052 0.5122088 4268326
## 3  2006 2184237 2081318 0.5120640 4265555
## 4  1960 2179708 2078142 0.5119269 4257850
## 5  1957 2179960 2074824 0.5123550 4254784
## 6  2008 2173625 2074069 0.5117188 4247694
```

2007 had the most births