FE590. Assignment #1.

John-Craig Borman (10402229)

9/12/18

2018-09-17

Question 1

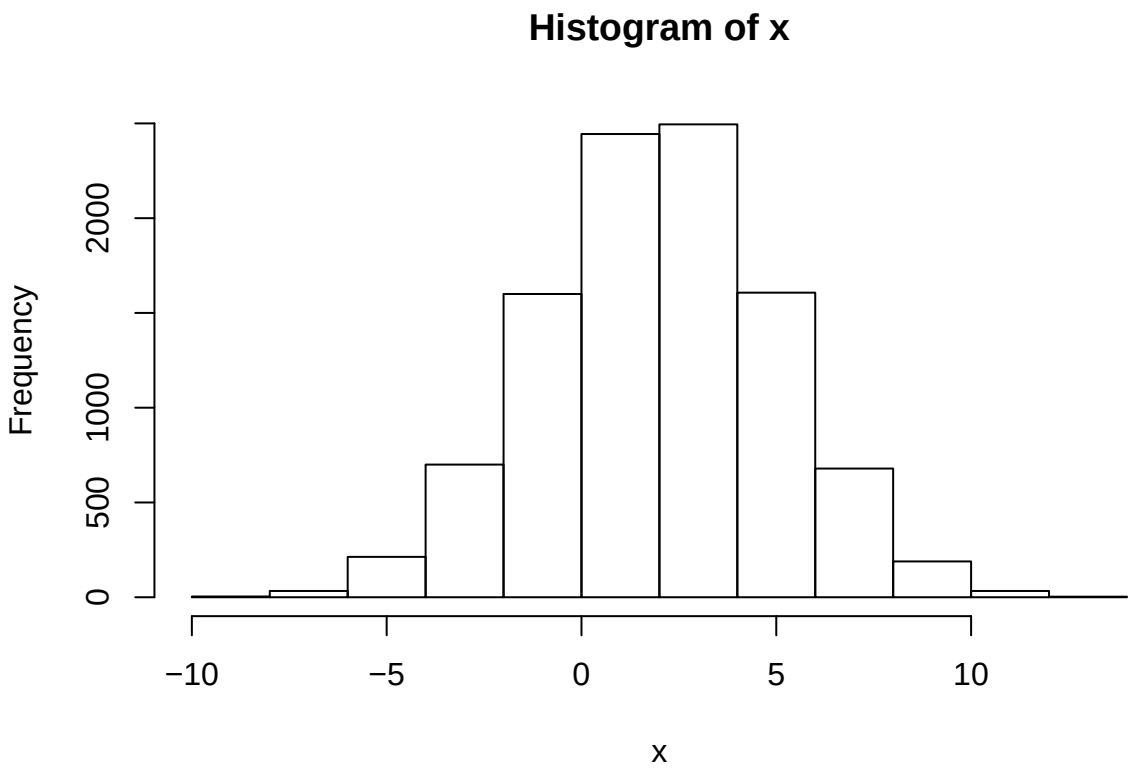
Question 1.1

```
CWID = 10402229 #Place here your Campus wide ID number, this will personalize
#your results, but still maintain the reproduceable nature of using seeds.
#If you ever need to reset the seed in this assignment, use this as your seed
#Papers that use -1 as this CWID variable will earn 0's so make sure you change
#this value before you submit your work.
personal = CWID %% 10000
set.seed(personal)
```

Generate a vector `x` containing 10,000 realizations of a random normal variable with mean 2.0 and standard deviation 3.0, and plot a histogram of `x` using 100 bins. To get help generating the data, you can type `?rnorm` at the R prompt, and to get help with the histogram function, type `?hist` at the R prompt.

Solution:

```
x <- rnorm(n = 10000, mean = 2.0, sd = 3.0)
hist(x)
```



Question 1.2

Confirm that the mean and standard deviation are what you expected using the commands `mean` and `sd`.

Solution:

```
mean(x) # ~2
```

```
## [1] 1.982375
```

```
sd(x)   # ~3
```

```
## [1] 3.018047
```

Question 1.3

Using the `sample` function, take out 10 random samples of 500 observations each. Calculate the mean of each sample. Then calculate the mean of the sample means and the standard deviation of the sample means.

Solution:

```
means <- c()

for(i in 1:10){
  sample_vec <- sample(x = x, size = 500)
  print(paste0("Iteration: ", i))
```

```
  means <- c(means, mean(sample_vec))
  print(paste0("mean: ", means[i]))
}
```

```
## [1] "Iteration: 1"
## [1] "mean: 1.89762886557256"
## [1] "Iteration: 2"
## [1] "mean: 1.93741578107116"
## [1] "Iteration: 3"
## [1] "mean: 1.9120817890946"
## [1] "Iteration: 4"
## [1] "mean: 1.96472386868012"
## [1] "Iteration: 5"
## [1] "mean: 1.98143301600266"
## [1] "Iteration: 6"
## [1] "mean: 2.00828097868113"
## [1] "Iteration: 7"
## [1] "mean: 2.0442036923372"
## [1] "Iteration: 8"
## [1] "mean: 1.94493969423164"
## [1] "Iteration: 9"
## [1] "mean: 2.02687931113431"
## [1] "Iteration: 10"
## [1] "mean: 2.01943668875613"
```

Mean and Std. Dev of sample means:

```
print(paste0("Mean of means: ", mean(means)))
```

```
## [1] "Mean of means: 1.97370236855615"
```

```
print(paste0("Std Dev of means: ", sd(means)))
```

```
## [1] "Std Dev of means: 0.0505021342828562"
```

Question 2

Sir Francis Galton was a controversial genius who discovered the phenomenon of "Regression to the Mean." In this problem, we will examine some of the data that illustrates the principle.

Question 2.1

First, install and load the library `HistData` that contains many famous historical data sets. Then load the Galton data using the command `data(Galton)`. Take a look at the first few rows of `Galton` data using the command `head(Galton)`.

Solution:

```
library("HistData")
data("Galton")
head(Galton)
```

```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
```

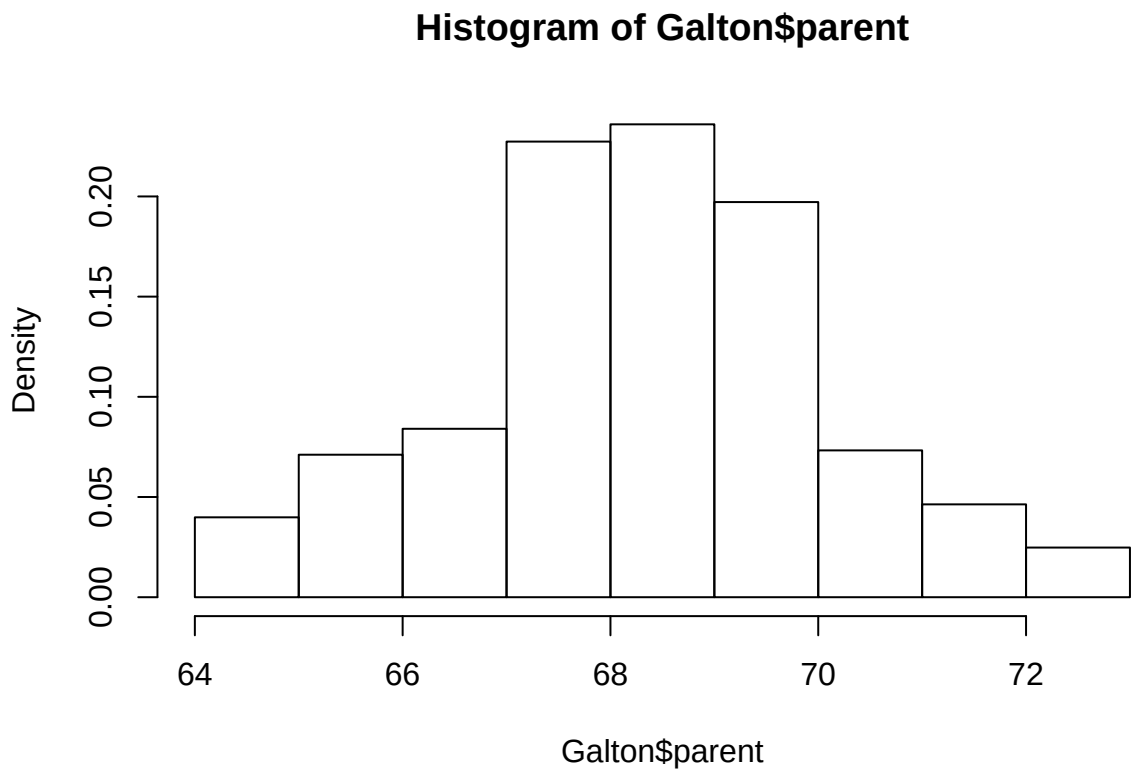
```
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

As you can see, the data consist of two columns. One is the height of a parent, and the second is the height of a child. Both heights are measured in inches.

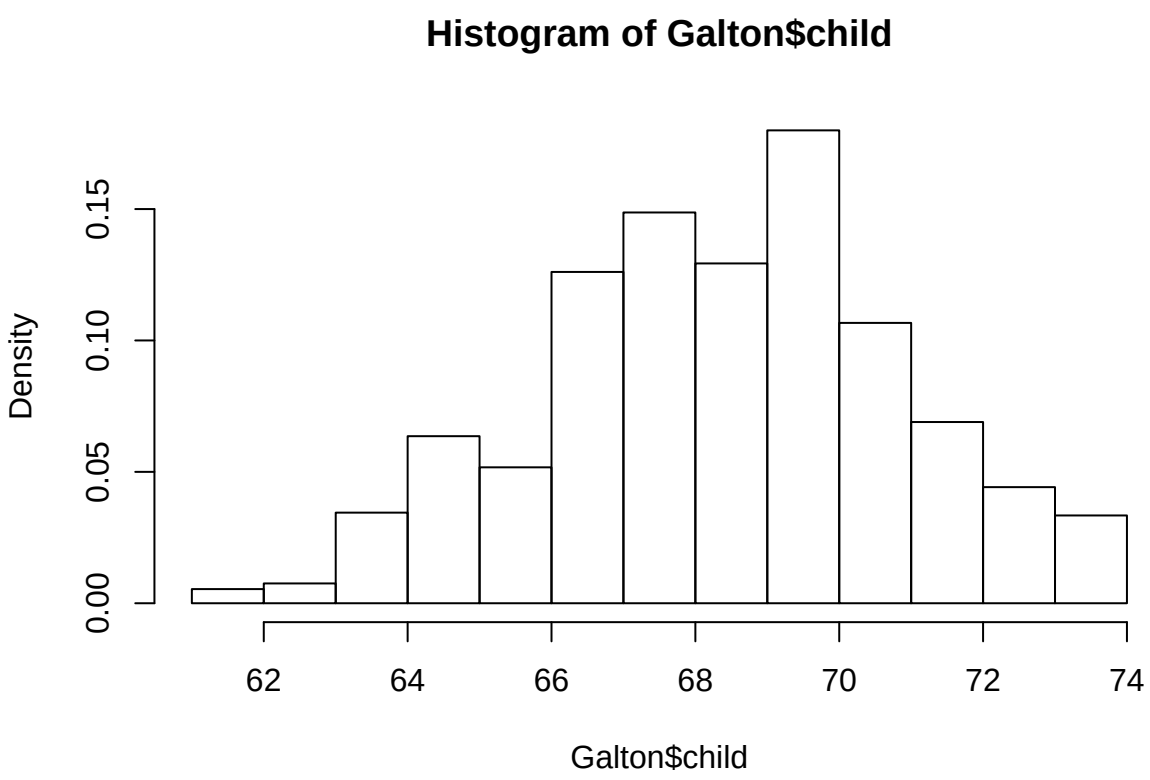
Plot one histogram of the heights of the children and one histogram of the heights of the parents. These histograms should use the same `x` and `y` scales.

Solution:

```
hist(Galton$parent, freq = FALSE)
```




```
hist(Galton$child, freq = FALSE)
```



Comment on the shapes of the histograms.

Solution:

The parent histogram is strongly centered around 67 to 70 inches. In all, the data ranges from 64 to 73 inches.

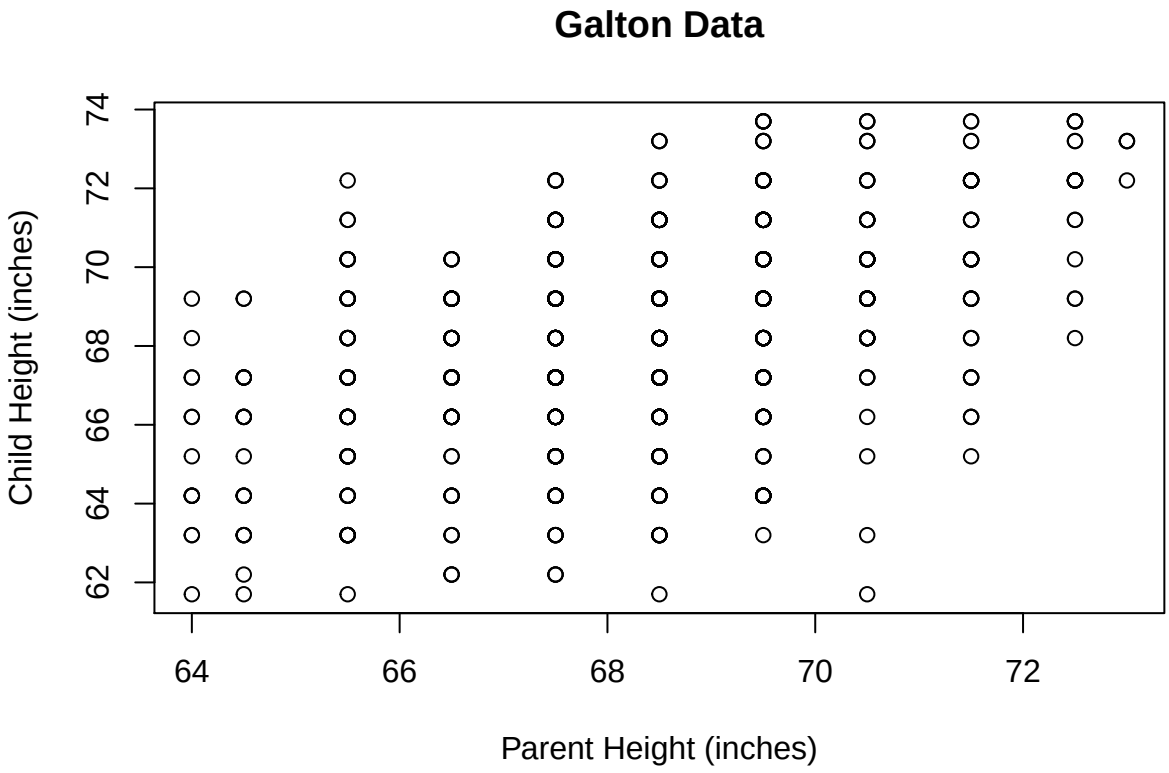
Conversely, there is greater variation in the distribution of children's heights. The data spans from 61 to 74 inches with a concentration between 66 and 71 inches. This histogram notably has a flatter distribution relative to the parent's height distribution.

Question 2.2

Make a scatterplot the height of the child as a function of the height of the parent. Label the x-axis "Parent Height (inches)," and label the y-axis "Child Height (inches)." Give the plot a main tile of "Galton Data."

Solution:

```
plot(y = Galton$child, x = Galton$parent, xlab = "Parent Height (inches)", ylab = "Child Height (inches)
```



Question 3

If necessary, install the `ISwR` package, and then `attach` the `bp.obese` data from the package. The data frame has 102 rows and 3 columns. It contains data from a random sample of Mexican-American adults in a small California town.

Question 3.1

The variable `sex` is an integer code with 0 representing male and 1 representing female. Use the `table` function operation on the variable 'sex' to display how many men and women are represented in the sample.

Solution:

```
library("ISwR")
data("bp.obese")
table(bp.obese$sex)
```

```
##
##  0  1
## 44 58
```

Question 3.2

The `cut` function can convert a continuous variable into a categorical one. Convert the blood pressure variable `bp` into a categorical variable called `bpc` with break points at 80, 120, and 240. Rename the levels of `bpc`

using the command `levels(bpc) <- c("low", "high")`.

Solution:

```
bpc <- cut(bp.obese$bp, breaks = c(80, 120, 240))
levels(bpc) <- c("low", "high")
```

Question 3.3

Use the `table` function to display a relationship between `sex` and `bpc`.

Solution:

```
table(bp.obese$sex, bpc)
```

```
##    bpc
##     low high
##   0  16   28
##   1  28   30
```

Question 3.4

Now cut the `obese` variable into a categorical variable `obesec` with break points 0, 1.25, and 2.5. Rename the levels of `obesec` using the command `levels(obesec) <- c("low", "high")`.

Use the `ftable` function to display a 3-way relationship between `sex`, `bpc`, and `obesec`.

Solution:

```
obesec <- cut(bp.obese$obese, breaks = c(0, 1.25, 2.5))
levels(obesec) <- c("low", "high")

ftable(bp.obese$sex, bpc, obesec)
```

```
##        obesec low high
##   bpc
## 0 low          12    4
##   high         15   13
## 1 low          14   14
##   high          4   26
```

Which group do you think is most at risk of suffering from obesity?

Solution:

Proportions of Obese Men and Women in Sample:

```
table(bp.obese$sex, obesec)
```

```
##    obesec
##     low high
##   0  27   17
##   1  18   40
```

```
17 / (17 + 27) # % of Obese Men
```

```
## [1] 0.3863636
```

```
40 / (40 + 18) # % of Obese Men
```

```
## [1] 0.6896552
```

Regardless of blood pressure, women are more likely to be obese than men (based on the sample data proportions: 68% vs 38%).

Proportions of Obese High and Low blood pressure individuals in Sample:

```
table(bpc, obesec)
```

```
##       obesec
## bpc    low high
##   low   26   18
##   high  19   39
```

```
18 / (18 + 26)  # % of Obese Individuals with Low blood pressure
```

```
## [1] 0.4090909
```

```
39 / (39 + 19)  # % of Obese Individuals with High blood pressure
```

```
## [1] 0.6724138
```

Regardless of sex, individuals with high blood pressure are more likely to be obese than individuals with low blood pressure (based on the sample data proportions: 67% vs 41%).

Proportions of Obese Men/Women by blood pressure in Sample:

```
ftable(bp.obese$sex, bpc, obesec)
```

```
##        obesec low high
##   bpc
## 0 low          12    4
##   high         15   13
## 1 low          14   14
##   high          4   26
```

```
# Proportion of Obese Men
4 / (4 + 12)   # Men w/ Low Blood Pressure
```

```
## [1] 0.25
```

```
13 / (13 + 15) # Men w/ High Blood Pressure
```

```
## [1] 0.4642857
```

```
# Proportion of Obese Women
14 / (14 + 14)# Women w/ Low Blood Pressure
```

```
## [1] 0.5
```

```
26 / (26 + 4) # Women w/ High Blood Pressure
```

```
## [1] 0.8666667
```

For Men and Women of comparable Blood Pressure classifications, the sample proportion of obese women is always higher (for both high and low blood pressure categories).

Given that the sample was taken from a small town of Mexican-American adults in southern California, these results most closely apply to individuals of this geographical location and ethnic background. With this

context in mind, the female population (regardless of blood pressure category) is most at risk of suffering from obesity because of the comparative proportions analyzed above.