FE590. Assignment #1.

2019-02-05

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature _____Yifu He____________________ Date:___02/08/2019__________

Question 1

Question 1.1

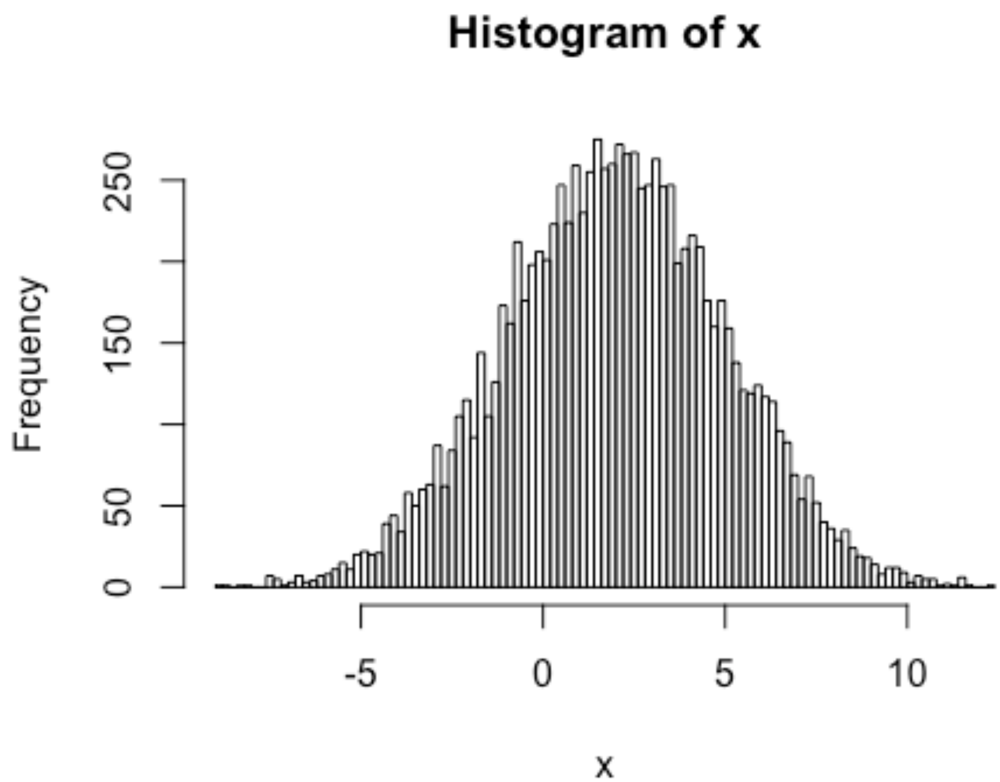
```
CWID = 10442277 #Place here your Campus wide ID number, this will personalize
#your results, but still maintain the reproduceable nature of using seeds.
#If you ever need to reset the seed in this assignment, use this as your seed
#Papers that use -1 as this CWID variable will earn 0's so make sure 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.

```
#response
CWID

## [1] 10442277

x = rnorm(10000,
          mean = 2.0,
          sd = 3.0)
hist(x,breaks = 100 )
```



Question 1.2

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

```
# my response
mean1 = mean(x)
sd1 = sd(x)
mean1

## [1] 1.982528

sd1

## [1] 3.021309
```

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.

```r
# Enter your R code here!
sample.mean <- c()
for(i in 1:(length(x)/500)){
  sample.mean[i] <- mean(sample(x[(500*i-499):(500*i)],size = 10))
}
mean2 <- mean(sample.mean)
sd2 <- sd(sample.mean)
mean2

## [1] 2.004516

sd2

## [1] 0.7575051
```

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)`.

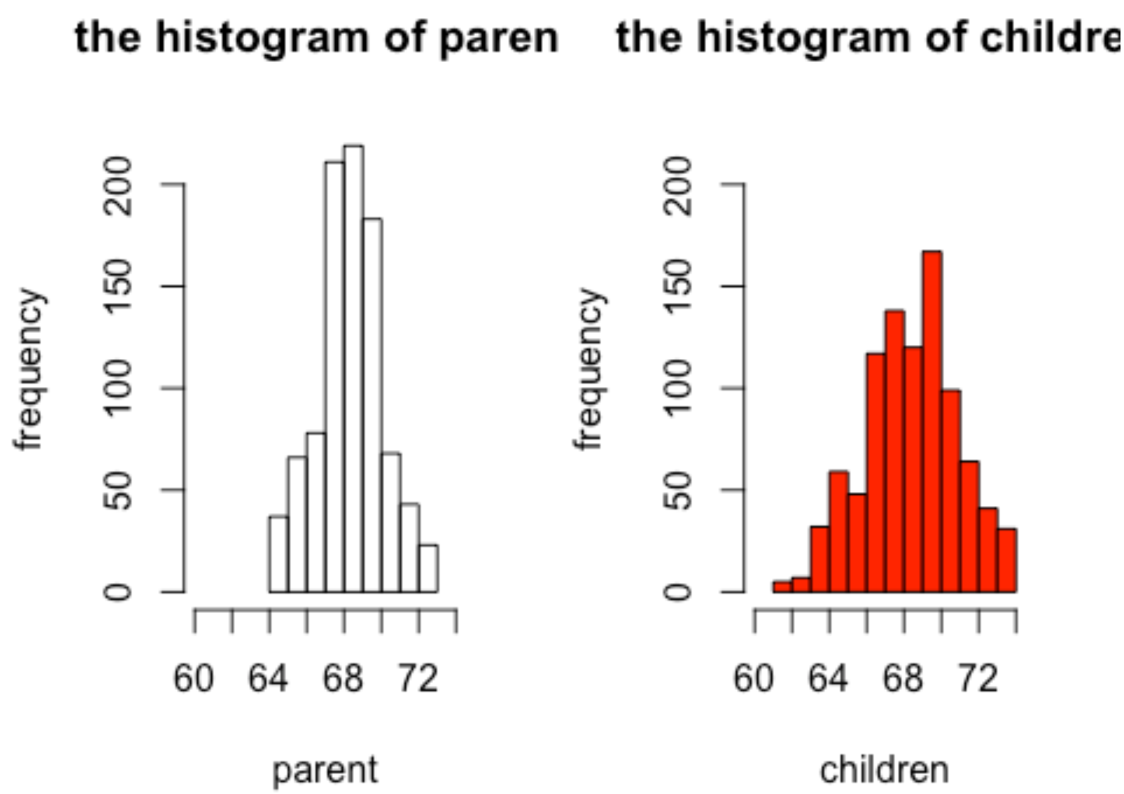
```r
# Enter your R code here!
library("HistData")
data("Galton")
head(Galton)

##   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. This histograms should use the same `x` and `y` scales.

```r
# Enter your R code here!
par(mfrow=c(1,2))
hist(Galton[,1],
     xlab = "parent",
     ylab = "frequency",
     main = "the histogram of parent",
     xlim = c(60,74),
     ylim = c(0,220))
hist(Galton[,2],
     xlab = "children",
     ylab = "frequency",
     main = "the histogram of children",
     col="red",
     xlim = c(60,74),
     ylim = c(0,220))
```



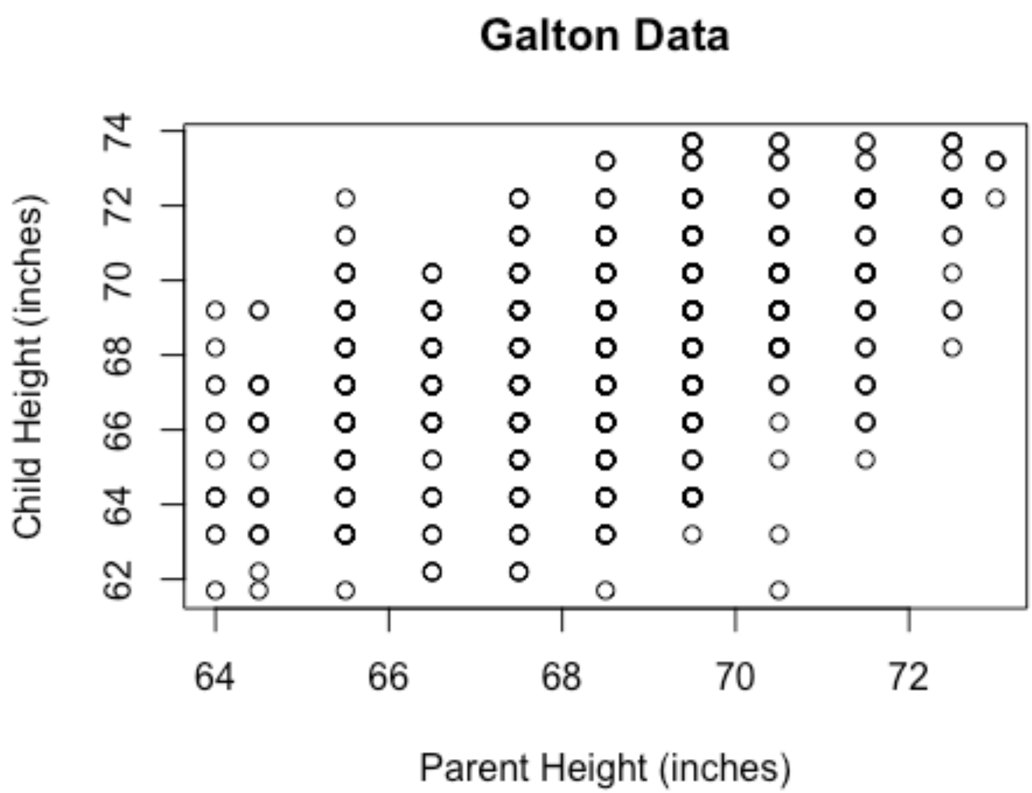
Comment on the shapes of the histograms.

It looks like normal distribution. The approximate mean of each are 68 and 69.

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."

```
# Enter your R code here!
plot(Galton[,1],
     Galton[,2],
     xlab = "Parent Height (inches)",
     ylab = "Child Height (inches)",
     main = "Galton Data"
     )
```



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.

```
# Enter your R code here!
library("ISwR")
attach(bp.obese)
table(sex)

## 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")`.

```
# Enter your R code here!
breakpoint <- c(80, 120, 240)
bpc <- cut(bp,breakpoint)
levels(bpc) <- c("low", "high")
```

Question 3.3

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

```
# Enter your R code here!
table(sex,bpc)

##    bpc
## sex 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`.

```
# Enter your R code here!
breakpoint2 <- c(0,1.25,2.5)
obesec <- cut(obese, breakpoint2)
levels(obesec) <- c("low", "high")
table <- ftable(sex,bpc,obesec)
```

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

Conclusion： female group with high blood pressure Compare the ratios of high obesity to low obesity among each group. We can find that the female group of high blood pressure owns the highest ratio, which means it is most at risk.

```
# Enter your R code here!
ratios <- c()
for (i in 1:4){
  ratios[i] <- table[i,2]/table[i,1]
}
ratios

## [1] 0.3333333 0.8666667 1.0000000 6.5000000

detach(bp.obese)
```