FE590. Assignment #1.

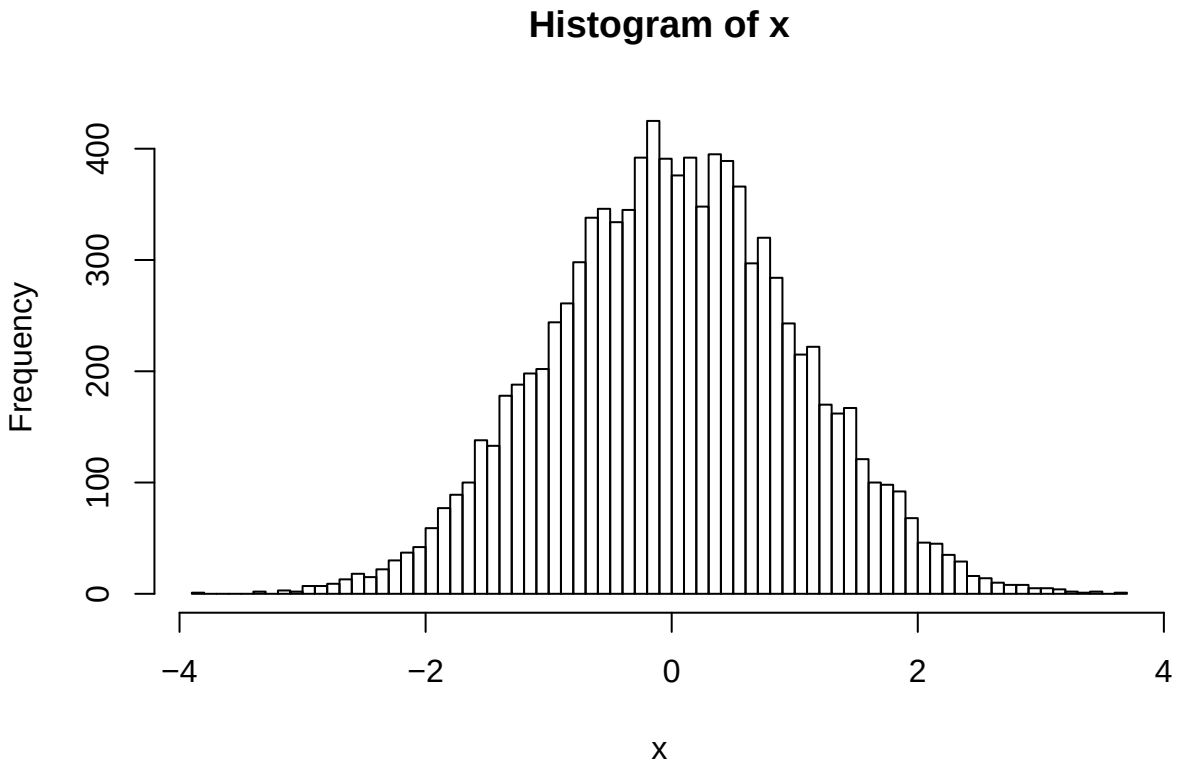
Kejia Huang

2016-04-10

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

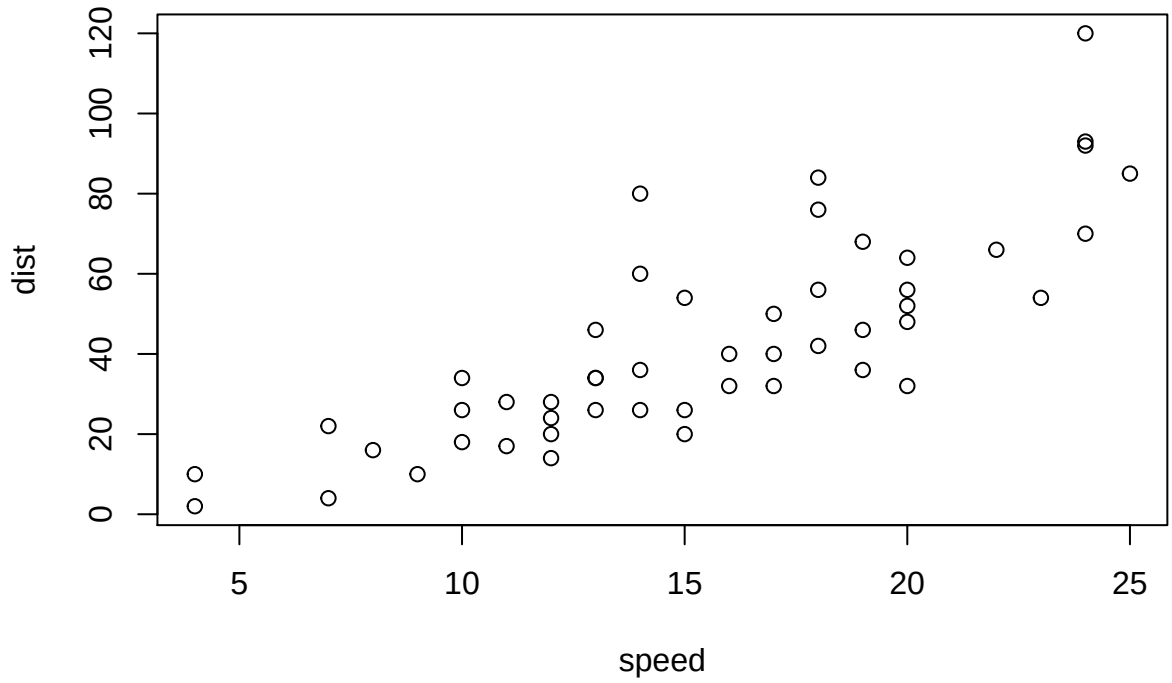
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
x <- rnorm(10000)
hist(x, 100)
```



You can also embed plots, for example:

```
plot(cars)
```



Instructions

In this assignment, you should use R markdown to answer the questions below. Simply type your R code into embedded chunks as shown above.

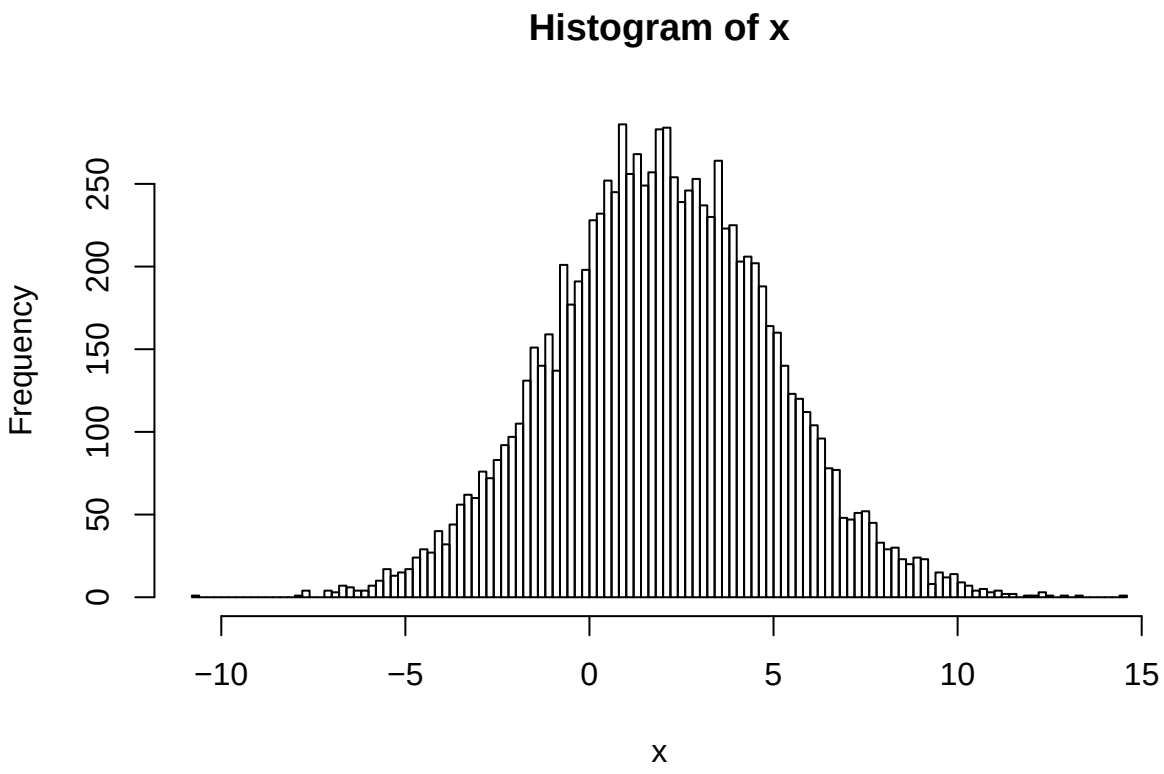
When you have completed the assignment, knit the document into a PDF file, and upload *both* the .pdf and .Rmd files to Canvas.

Question 1

Question 1.1

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.

```
# Enter your R code here!
x <- rnorm( 10000, mean = 2, sd = 3 )
hist( x, 100)
```



Question 1.2

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

```
# Enter your R code here!
mean( x )
```

```
## [1] 1.983402
```

```
sd ( x )
```

```
## [1] 2.995915
```

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.

```
# Enter your R code here!
n <- vector()
for( i in 1:10 ){
    n[i] <- mean( sample( 500, replace = TRUE) )
}
mean( n )
```

```
## [1] 249.5702
```

```
sd( n )
```

```
## [1] 6.312639
```

Do your results correspond approximately to the analytic expression that we discussed in class? **yes**

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)`.

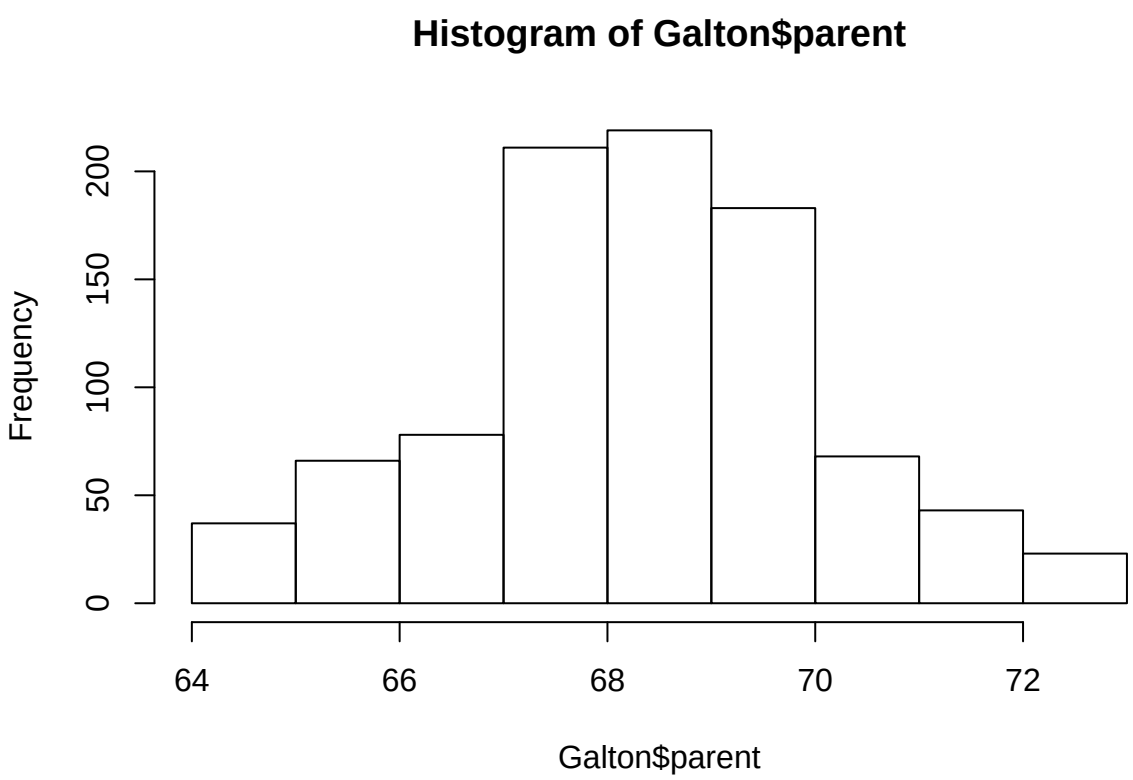
```
# Enter your R code here!
library(HistData)
data(Galton)
head(Galton)
```

```
##   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
```

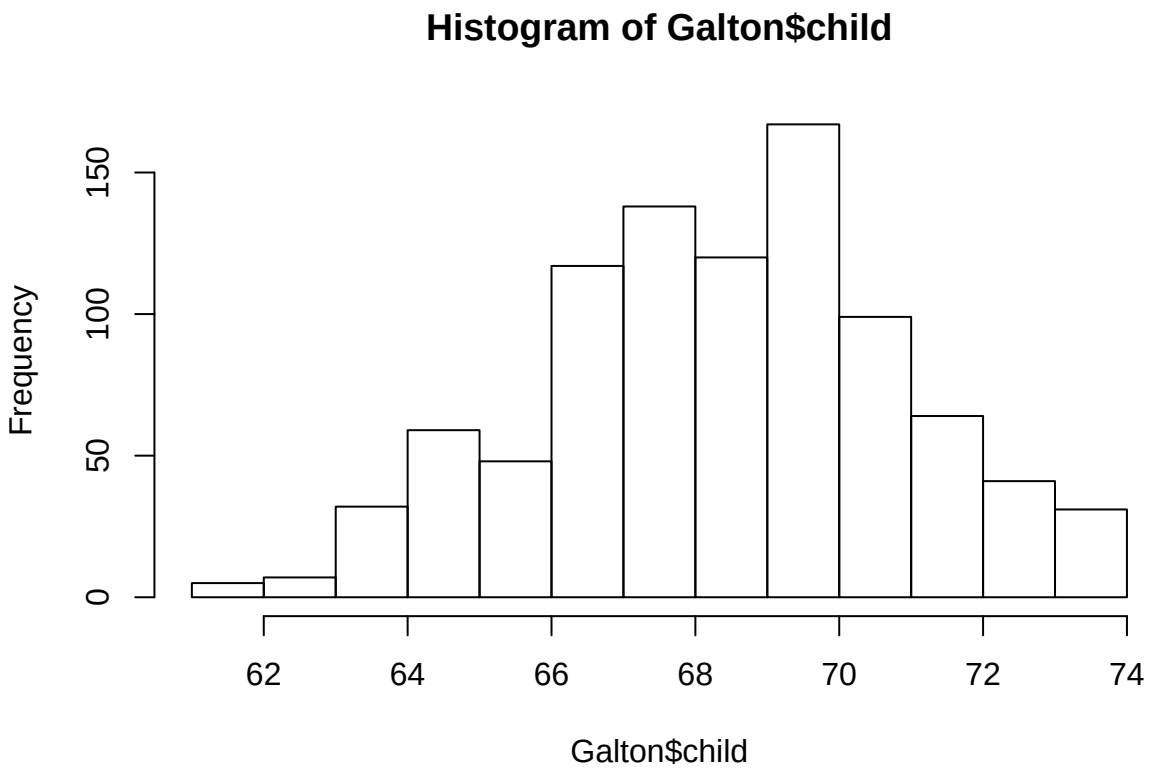
As you can see, the data consist of two columns. One is the height of a parent, and the second is the height of a child. Both heights are measured in inches.

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

```
# Enter your R code here!
hist( Galton$parent )
```




```
hist( Galton$child )
```




```
mean(Galton$parent)
```

```
## [1] 68.30819
```

```
mean(Galton$child)
```

```
## [1] 68.08847
```

```
sd(Galton$parent)
```

```
## [1] 1.787333
```

```
sd(Galton$child)
```

```
## [1] 2.517941
```

Comment on the shapes of the histograms.

For two histograms, the histogram of Galton'child is more scatter than the histogram of Galton' parent. And it seems that the mean of heights of parents is bit of higher than childrens

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."

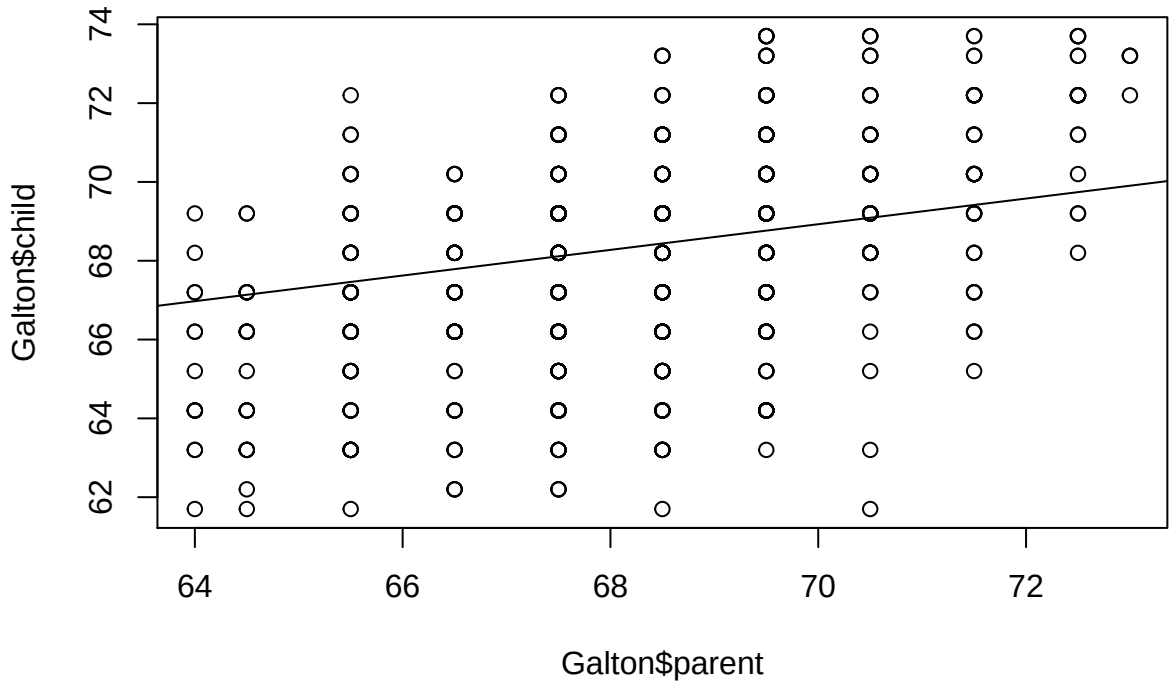
Perform a linear regression of the child's height onto the parent's height. Add the regression line to the scatter plot.

Using the `summary` command, print a summary of the linear regression results.

```
# Enter your R code here!
plot ( x = Galton$parent, y = Galton$child )
lm( Galton$parent ~ Galton$child )
```

```
##
## Call:
## lm(formula = Galton$parent ~ Galton$child)
##
## Coefficients:
##  (Intercept)  Galton$child
##      46.1353        0.3256
```

```
abline( a = 46.1353, b = 0.3256 )
```



What is the slope of the line relating a child's height to the parent's height? Can you guess why Galton says that there is a "regression to the mean"?

The slope is 0.3256. I think the reason is that the average regression of the offspring is a constant fraction of their respective mid-parental deviations because of the slope is less than 1

Is there a significant relationship a child's height to the parent's height? If so, how can you tell from the regression summary?

If the parents' height is higher, then their children's heights will be higher But because of the slope of the line is less than 1, so the heights difference between children will be less than parents'

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!
bpc <- cut( bp, breaks = 40*c(2, 3, 6))
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`.

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

```
##          obesec low high
## sex 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?

female with high blood pressure