FE590. Assignment #1.

2017-02-01

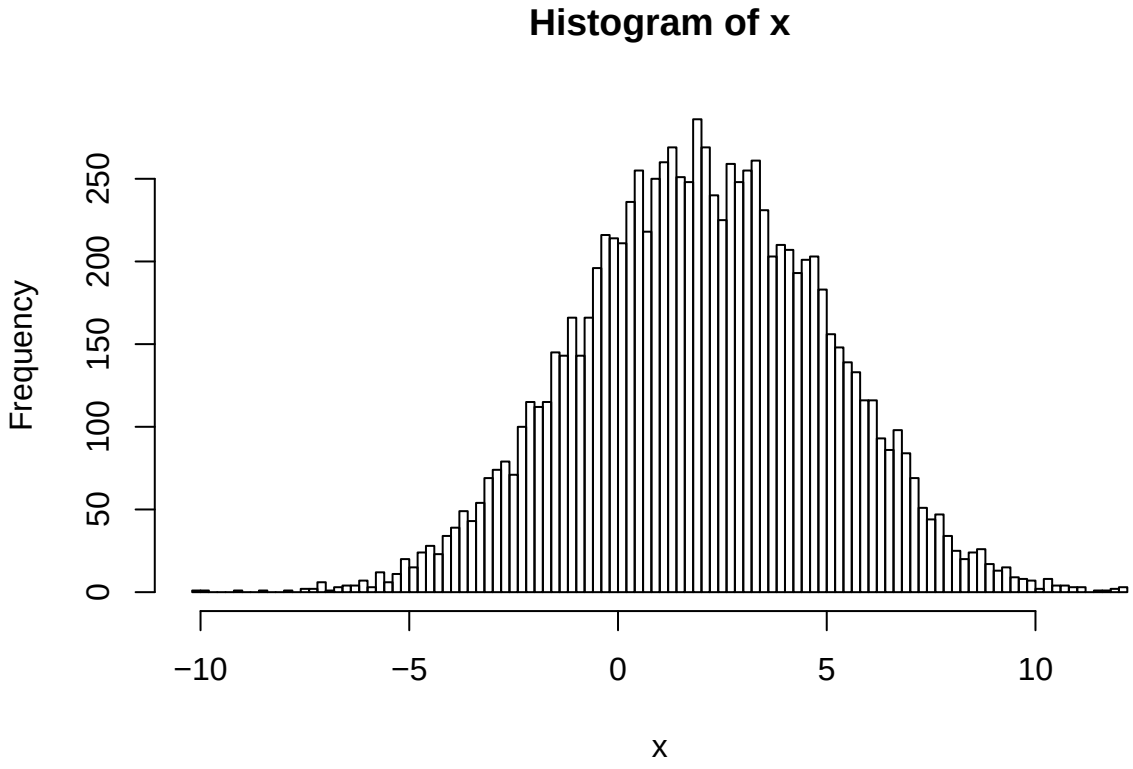
Question 1

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.

Solution:

```
set.seed(10000)
x <- rnorm(10000, mean = 2, sd = 3)
hist(x, nclass = 100)
```



Question 1.2

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

```
# the mean
m <- mean(x)
# [1] 2.016439
```

```r
# the dtandard deviation
s <- sd(x)
# [1] 2.978108
```

The mean is 2.016 and the standard deviation is 2.978. Compared to true value of 2 and 3 respectively, these values are what I expected.

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:

```r
ms <- rep(0, 10)
for (i in 1:10) {
    ms[i] <- mean(sample(x, 500))
}
mm <- mean(ms)
# [1] 2.064826

ss <- sd(ms)
# [1] 0.1507299
```

Solution:

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

According to what we discussed in cluss, the mean of the sample means should be close to the true mean. While the standard deviation of the sample means should be around the true standard deviation divided by the square root of the sample numbers (500), which is:

```r
sr <- 3/sqrt(500)
# [1] 0.1341641
```

so we have 0.134 compared to the standard deviation of the sample means 0.151. They are close.

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

Solution:

```
galton <- as.data.frame(Galton)
min(galton)
```

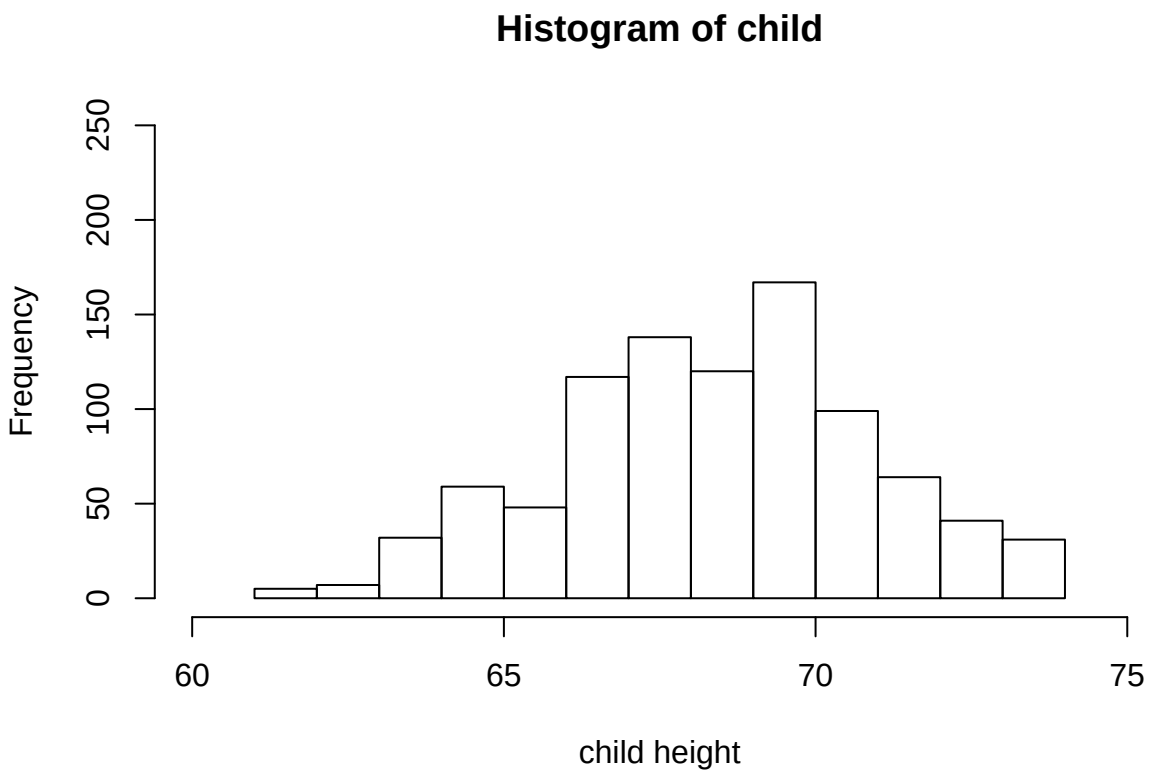
```
## [1] 61.7
```

```
max(galton)
```

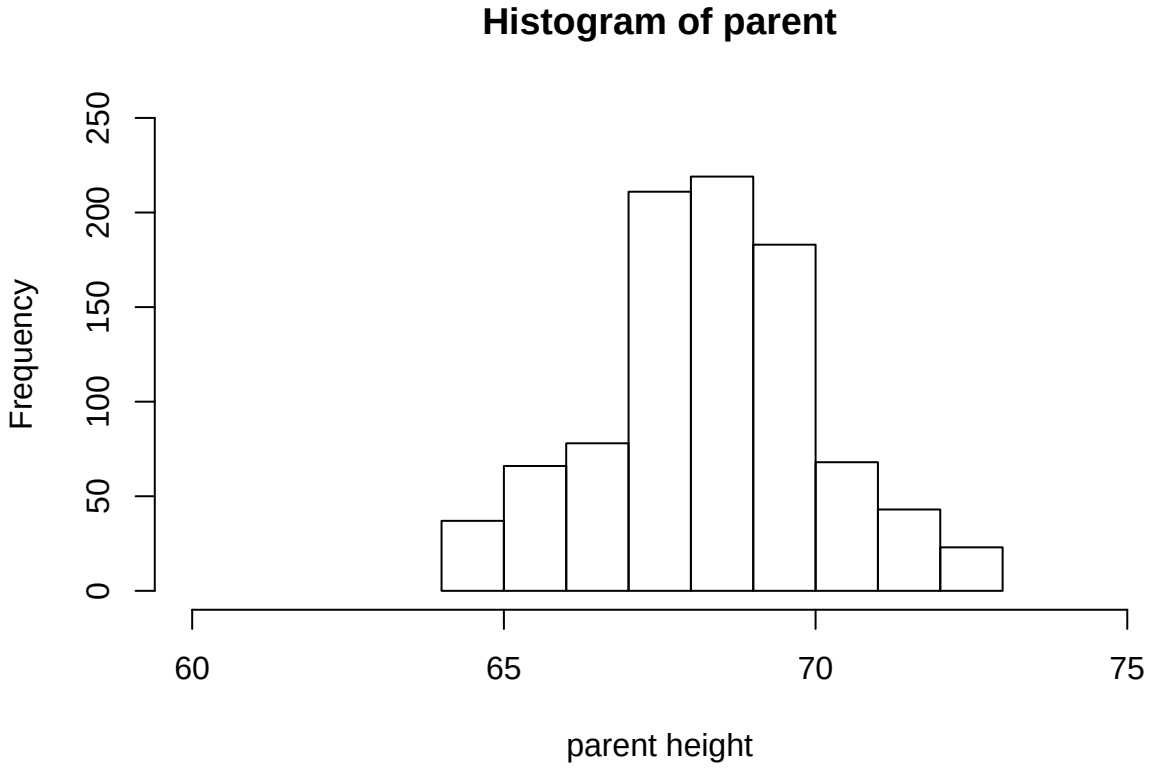
```
## [1] 73.7
```

```
# so we choose [60, 75] as the range of heights.

child <- as.numeric(unlist(galton[2]))
hist(child, xlim = c(60, 75), ylim = c(0, 250),
     xlab = "child height")
```




```
parent <- as.numeric(unlist(galton[1]))
hist(parent, xlim = c(60, 75), ylim = c(0, 250),
     xlab = "parent height")
```



Comment on the shapes of the histograms.

Solution:

The histogram of children heights includes a wider range of heights and it is flatter than parents' which indicate the childrens' heights have a larger variance.

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."

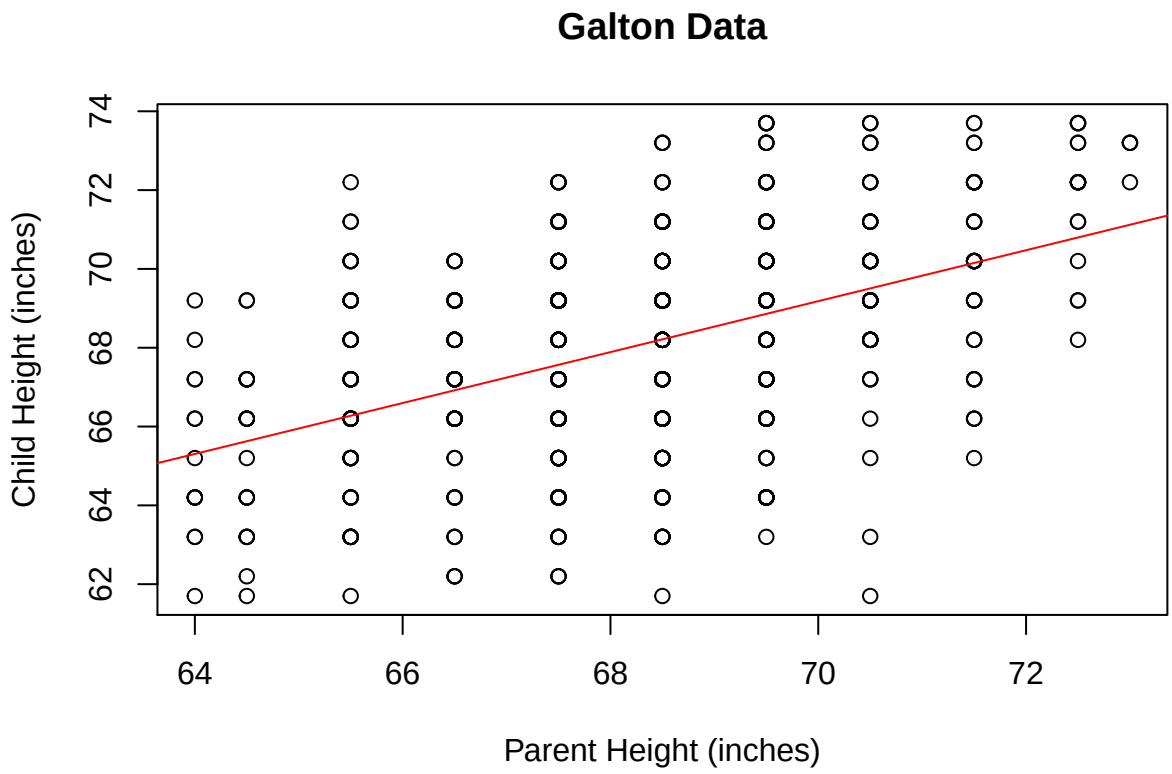
Perform a linear regression of the child's height onto the parent's height. Add the regression line to the scatter plot.

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

Solution:

```
plot(parent, child,
     xlab = "Parent Height (inches)", ylab = "Child Height (inches)",
     main = "Galton Data")

lg <- lm(child ~ parent, data = galton)
abline(lg, col = "red")
```




```
summary(lg)
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

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"?

Solution:

The slope of the line is 0.64629 which is less than one. Due to this, a child's height might be closer to the mean of childrens's heights than his parents' heights to the mean of parents' heights. That's what Galton says about "regression to the mean".

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

Solution:

Yes. Because the p-values of both coefficients are less than 2e-16 which are significant.

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:

```r
library(ISwR)
attach(bp.obese)
sexfm <- rep("male", length(sex))
```

```
sexfm[sex == 1] <- "female"
table(sexfm)
```

```
## sexfm
## female   male
##     58     44
```

There are 44 men and 58 women represented in the sample.

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, breaks = c(80, 120, 240))
levels(bpc) <- c("low", "high")
bpc
```

```
##   [1] high high high high high high high low  high high low  low  high low
##  [15] high low  high low  high high high high high low  low  low  low  low
##  [29] high high high high high low  high high high low  high high low  low
##  [43] low  high high high high high high low  high low  high low  low  low
##  [57] high high high high low  low  low  high high high low  low  low  low
##  [71] low  low  low  high high high low  low  high high high low  low  low
##  [85] high low  low  low  low  low  high high high low  low  high low  high
##  [99] high high high high
## Levels: low high
```

Question 3.3

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

Solution:

```
table(sexfm, bpc)
```

```
##         bpc
## sexfm    low high
##   female  28   30
##   male    16   28
```

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(obese, breaks = c(0, 1.25, 2.5))
levels(obesec) <- c("low", "high")
ftable(sexfm, bpc, obesec)
```

```
##             obesec low high
## sexfm  bpc
## female low          14   14
##        high          4   26
## male   low          12    4
##        high         15   13
```

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

Solution:

From the first table, we can see that the women with low blood pressure (bp) have the same amount of those with high bp. While the men with high bp are much more than men with low bp.

And from the second table, it shows that these females at low bp and males at high bp have the same chance to be low or high obese. However, females with higher bp have a much larger probability to be obese and males with lower bp have less chance to be high obese.

In sum, women with high blood pressure are most at risk of suffering from obesity.