Practice 1.

2020-07-02

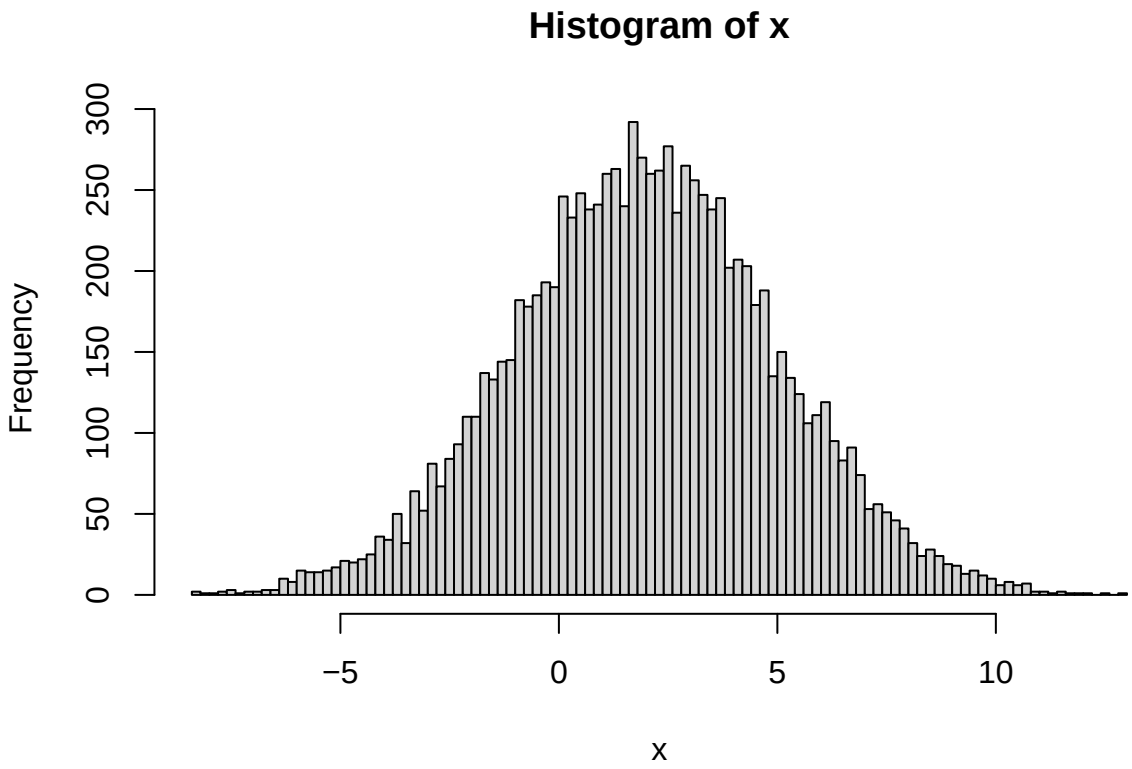
Question 1

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.

Solution:

```
x=rnorm(10000,2,3)
hist(x,breaks=100)
```



Question 1.2

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

Solution:

```
mean(x)
```

```
## [1] 1.991201
```

```
sd(x)
```

```
## [1] 3.002951
```

Question 2

Question 2.1

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

Solution:

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

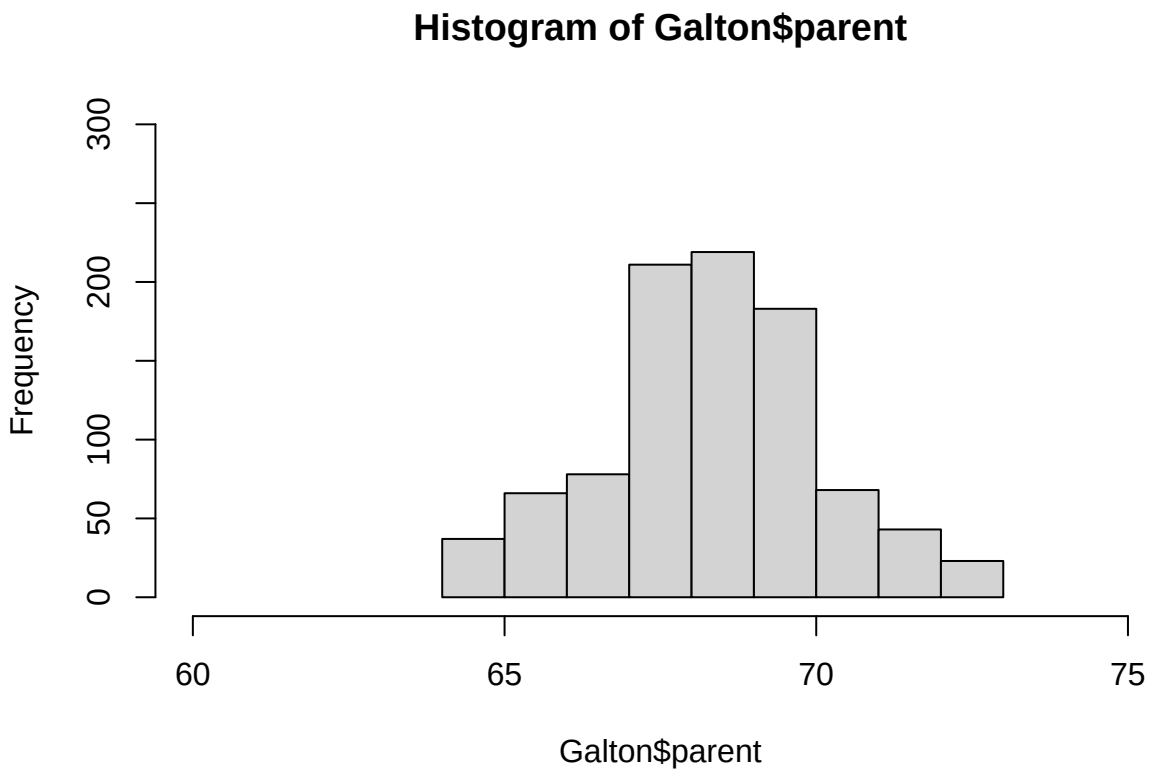
```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

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.

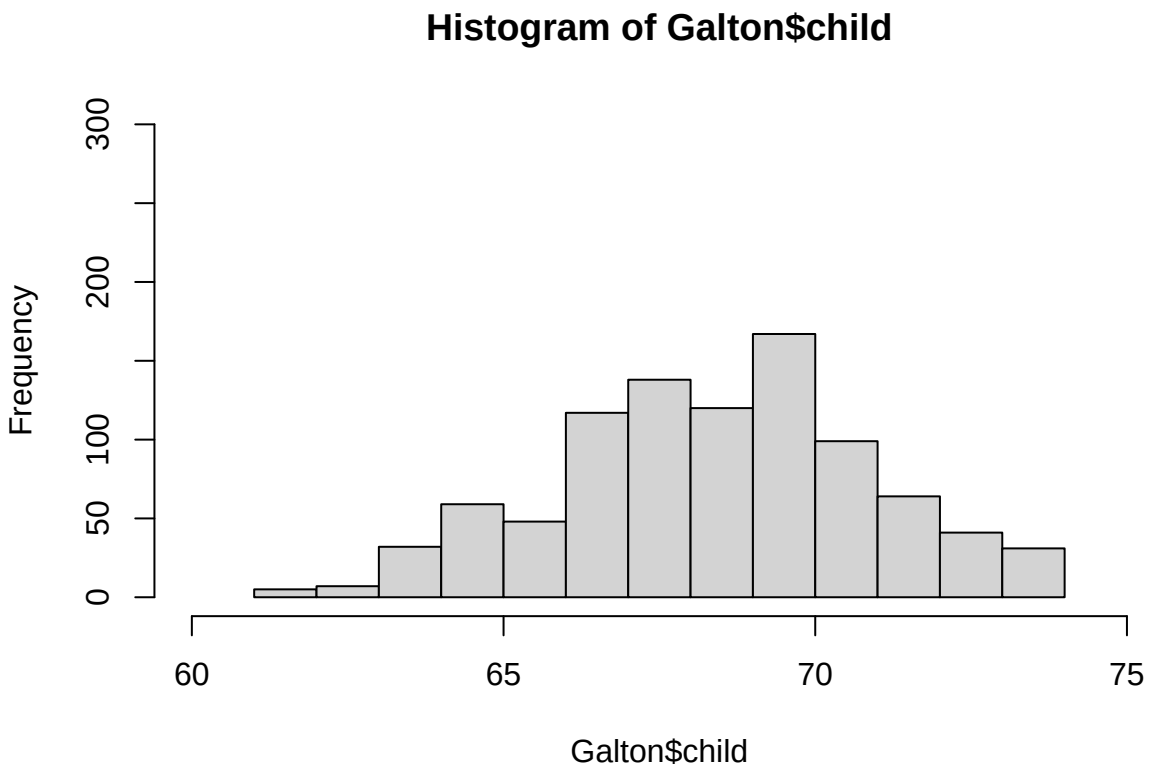
Plot one histogram of the heights of the children and one histogram of the heights of the parents. This histograms should use the same `x` and `y` scales.

Solution:

```
hist(Galton$parent,ylim = c(0,300),xlim = c(60,75))
```




```
hist(Galton$child,ylim = c(0,300),xlim = c(60,75))
```



Comment on the shapes of the histograms.

Solution:

From the histograms of child and parent, we can easily see that the shapes are more bell curve and normal distribution. Most parents have height from 67inches to 70inches. And most children have height from 66inches to 70inches.

Question 2.2

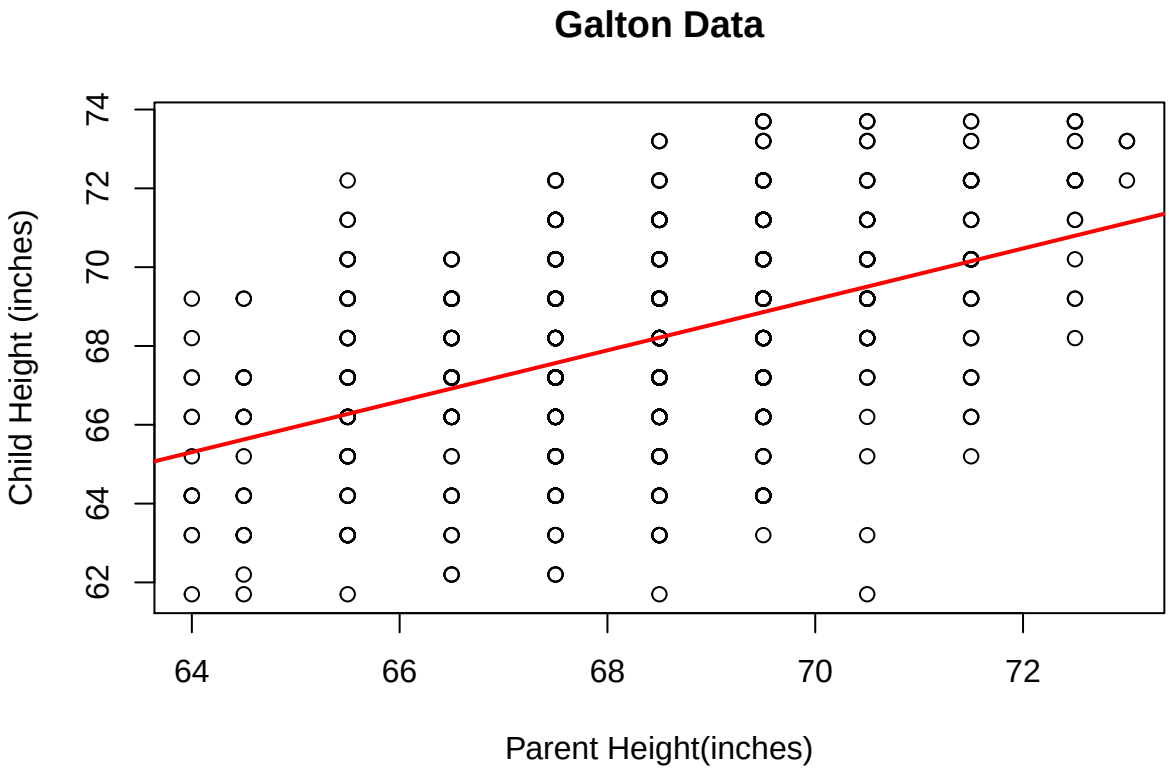
Make a scatterplot the height of the child as a function of the height of the parent. Label the `x`-axis "Parent Height (inches)," and label the `y`-axis "Child Height (inches)." Give the plot a main tile of "Galton Data."

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(child~parent,data=Galton,type='p',xlab="Parent Height(inches)",ylab="Child Height (inches)",main="
mod=lm(child~parent,data=Galton)
abline(mod,col='red',lwd=2)
```




```
summary(mod)
```

```
##
## 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 relating a child's height to the parent's height is 0.64629. The heights of children and parents tend to even out over time, which means short parents will tend to have higher kids and tall parents will tend to have shorter kids until the heights falls in to a certain interval–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:

The relationship between child's height to the parent's height is significant because the p-value is extremely small. However, the whole regression model is not very fit because the R-square is pretty low, which means this model can't explain the variability of response data around its mean well.

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)
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")`.

Solution:

```
bpc=cut(bp,br=c(80,120,240))
head(bpc)
```

```
## [1] (120,240] (120,240] (120,240] (120,240] (120,240] (120,240]
## Levels: (80,120] (120,240]
```

```
levels(bpc) <- c("low", "high")
```

Question 3.3

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

Solution:

```
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`.

Solution:

```
obesec=cut(obese,br=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
```

##Question 4

Using the Boston data in the MASS library, run a linear regression fit to determine a predictive model for the median value of a home using the indicators of rooms per dwelling and the property tax.

```
library(MASS)
mod=lm(medv~tax+age,data=Boston)
summary(mod)
```

```
##
## Call:
## lm(formula = medv ~ tax + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.638  -5.018  -2.181   2.765  34.648
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.060022   1.058433  33.124  < 2e-16 ***
## tax         -0.020377   0.002451  -8.315 8.66e-16 ***
## age         -0.061374   0.014673  -4.183 3.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.003 on 503 degrees of freedom
## Multiple R-squared:  0.2458, Adjusted R-squared:  0.2428
## F-statistic: 81.95 on 2 and 503 DF,  p-value: < 2.2e-16
```