STAT390: Homework 1

Ken

2023-10-01

① 2.34, 2.43, 2.63, 2.74, 2.74, 2.74, 2.78, 3.00, 3.03, 3.45, 3.56, 3.66, 3.85, 3.89, 3.93, 4.21, 4.33, 4.52

a. $\frac{3.03 + 3.45}{2} = \boxed{3.24 = \text{median}}$

$Q_1 = .25 \cdot 18 = 4.5 = x_{(5)} = 2.74$

$Q_3 = .75 \cdot 18 = x_{(14)} = 3.89$

$\boxed{IQR = 1.15}$

b. $\boxed{\sum_{i=1}^{n} x_i = 59.83}$ $\boxed{\sum_{i=1}^{n} x_i^2 = 207.0041}$

c. $\frac{59.83}{18} = \boxed{3.324 = \text{mean}}$

$\sqrt{\frac{207.004 - 18(3.324^2)}{17}} = \boxed{.692 = \text{Std deviation}}$

d. $\frac{95.2}{18} = \boxed{5.2\bar{8} = \text{new mean}}$

Mean changes drastically, while median stays the same because 39.8 doesn't change the position of median

② a. iii

freq : $\frac{1}{150} = .0067$ $\frac{35}{150} = .2333$ $\frac{13}{150} = .0867$

$\frac{.0067}{.25}$ $\frac{.2333}{.25}$ $\frac{.0867}{.25}$

density: $(.0267 \times .25) + (.9333 \times .25) + (.3467 \times .25)$

$\boxed{\text{density} = .3267}$

④ Example data : [1, 3, 8]

Prove : $\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}_n^2$

$\bar{x}_n = \frac{12}{3} = 4$

$\sum_{i=1}^{3} x_i^2 = 1^2 + 3^2 + 8^2 = 74$

$74 - 3(4^2)$

$74 - 48 = \frac{26}{3-1} = \boxed{13}$

$\sum_{i=1}^{3} (x_i - \bar{x}_n)^2 = (1-4)^2 + (3-4)^2 + (8-4)^2$

$9 + 1 + 16 = \frac{26}{3-1} = \boxed{13}$

Both come up with 13, so
$(n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = (n-1)^{-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2\right)$

⑦ a. $\frac{3 \cdot 3 \cdot 7 \cdot 7}{12^3} = \frac{441}{1728} = \boxed{.2552 = 25.52\%}$

b.
2 black = {{B, B, x}, {x, B, B}, {B, x, B}}

3 possible positions so

$$\frac{3 \cdot 5 \cdot 7 \cdot 7}{12^3} = \frac{735}{1728} = \boxed{.4253 = 42.53\%}$$

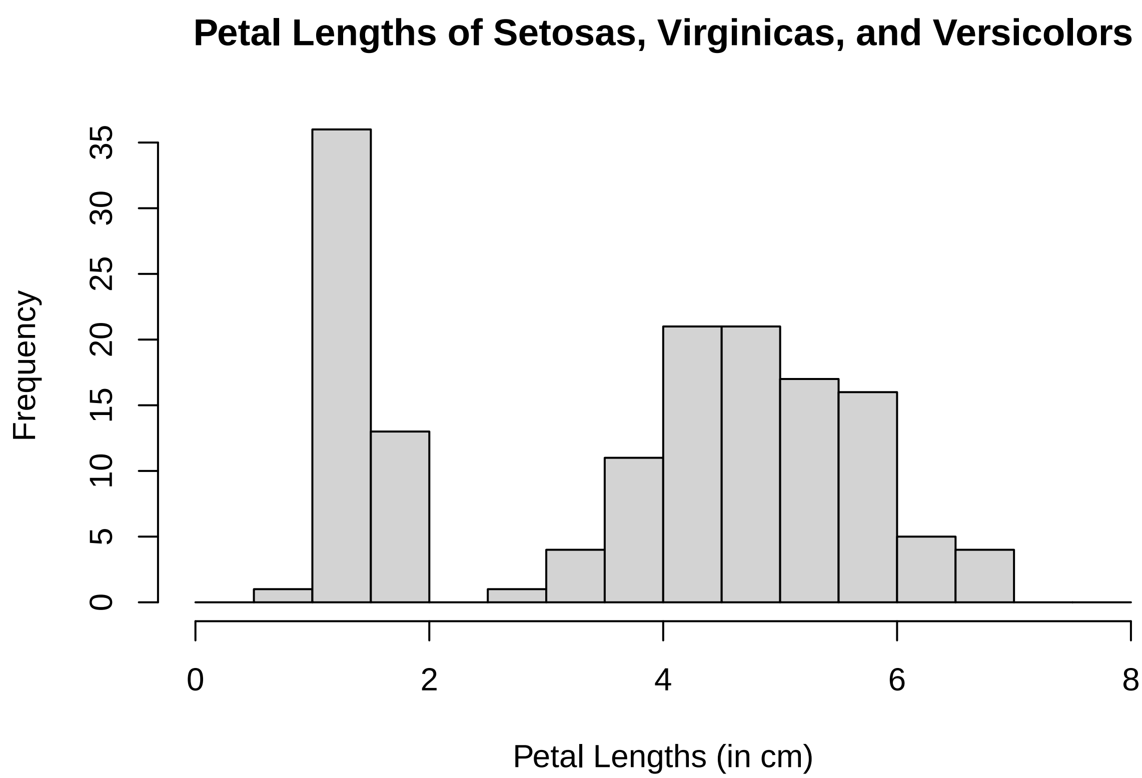
C. like last problem, $\{\{x, x, y\}, \{x, y, x\},$
$\{y, x, x\}\}$

3 possible positions per color so

$$\frac{(\overset{\text{2 blacks}}{3 \cdot 5 \cdot 7 \cdot 7}) + (\overset{\text{2 blues}}{3 \cdot 10 \cdot 2 \cdot 2}) + (\overset{\text{2 reds}}{3 \cdot 9 \cdot 3 \cdot 3})}{12^3} = \boxed{\begin{array}{c} .6354 \\ 63.54\% \end{array}}$$

Question 2

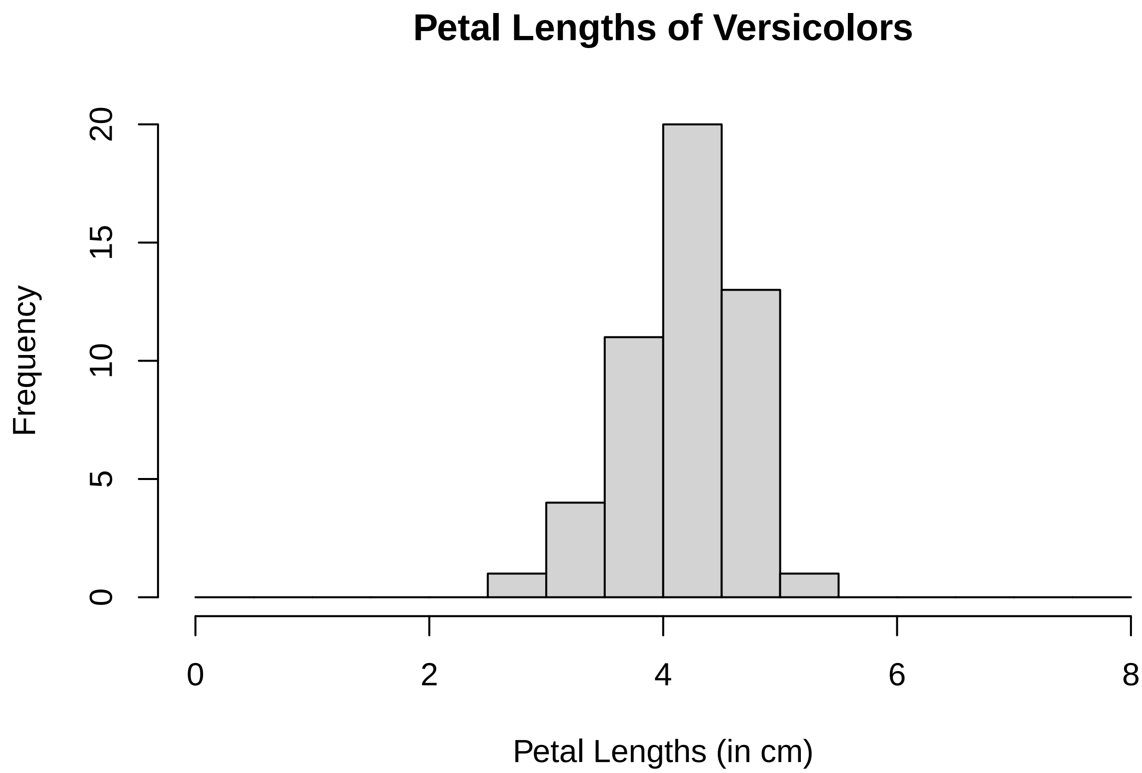
```
iris <- read.csv("iris.csv", header=TRUE)
histogram <- hist(iris$Petal.Length,
                  breaks=seq(0,8,l=17),
                  main="Petal Lengths of Setosas, Virginicas, and Versicolors",
                  xlab="Petal Lengths (in cm)",
                  ylab="Frequency")
```



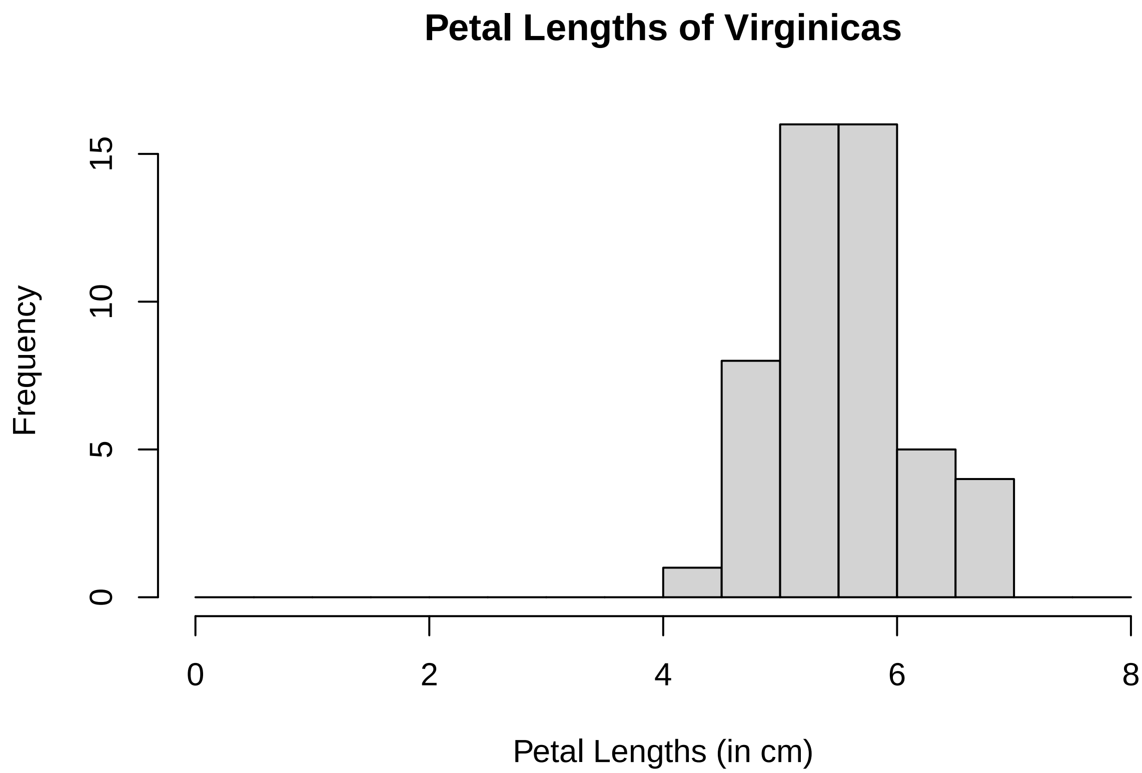
The shape of the histogram is bimodal. The percentage of iris flowers with a petal length less than or equal to 2cm is .3267 or 32.67%.

```
setosas <- iris %>% filter(iris$Species == "setosa")
setosa_hist <- hist(setosas$Petal.Length,
                  breaks=seq(0,8,l=17),
                  main="Petal Lengths of Setosas",
                  xlab="Petal Lengths (in cm)",
                  ylab="Frequency")
```


```
versicolors <- iris %>% filter(iris$Species == "versicolor")
versicolor_hist <- hist(versicolors$Petal.Length,
                  breaks=seq(0,8,l=17),
                  main="Petal Lengths of Versicolors",
                  xlab="Petal Lengths (in cm)",
                  ylab="Frequency")
```




```
virginicas <- iris %>% filter(iris$Species == "virginica")
virginica_hist <- hist(virginicas$Petal.Length,
                  breaks=seq(0,8,l=17),
                  main="Petal Lengths of Virginicas",
                  xlab="Petal Lengths (in cm)",
                  ylab="Frequency")
```



The most striking difference between the three is that each are in a different segment of the histogram. For example, the setosa petal lengths are on the lower end, versicolor petal lengths are in the middle, and virginica petal lengths are on the higher end.

Based on the histogram, the setosas have the least variability in petal length because the setosas are not as widely spread out across multiple bins like versicolors and viriginicas are. As shown below, the setosas do indeed have the least variability.

```
sd(setosas$Petal.Length)
```

```
## [1] 0.173664
```

```
sd(versicolors$Petal.Length)
```

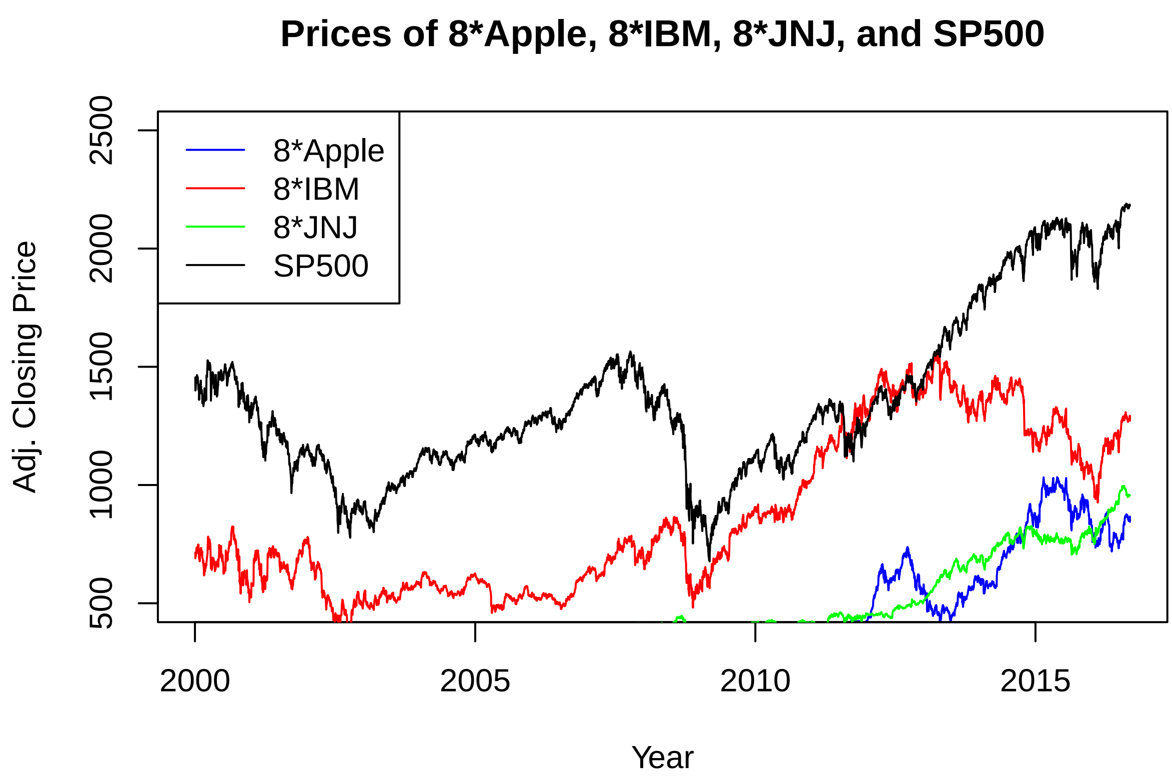
```
## [1] 0.469911
```

```
sd(virginicas$Petal.Length)
```

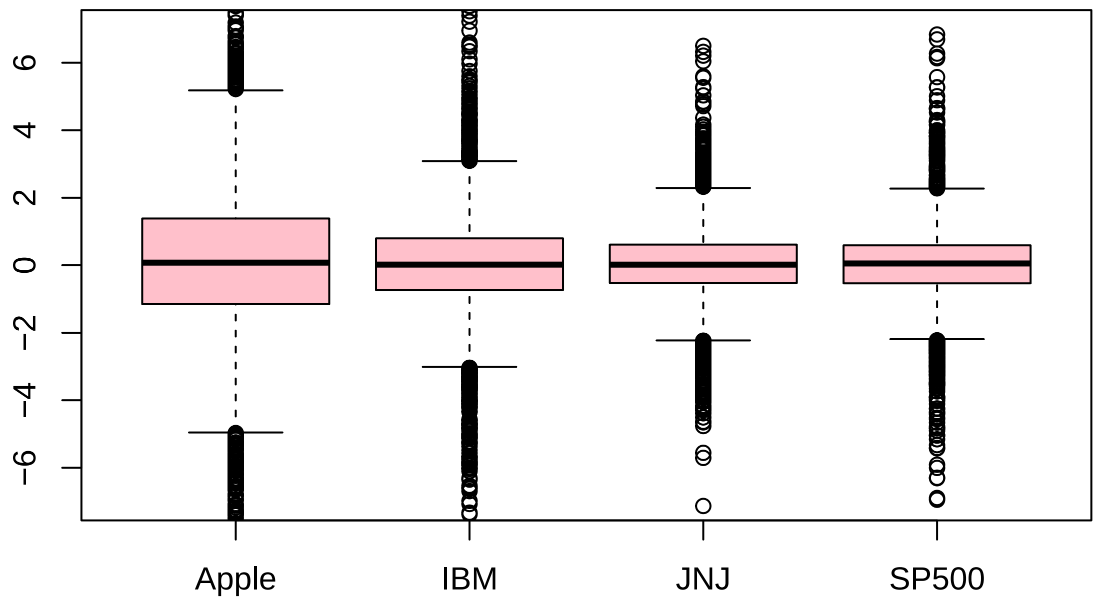
```
## [1] 0.5518947
```

Question 3

```
apple <- read.csv("APPL.csv", header=TRUE)
ibm <- read.csv("IBM.csv", header=TRUE)
jnj <- read.csv("JNJ.csv", header=TRUE)
sp500 <- read.csv("SP500.csv", header=TRUE)

dates <- apple[,1]
dates <- strptime(dates, "%Y-%m-%d")
dates <- rev(dates)

pApple <- apple[,7]
pApple <- rev(pApple)
pIBM <- ibm[,7]
pIBM <- rev(pIBM)
pJNJ <- jnj[,7]
pJNJ <- rev(pJNJ)
pSP500 <- sp500[,7]
pSP500 <- rev(pSP500)

plot(dates, 8*pApple, ylim = c(500, 2500), col="blue", type="l", xlab = "Year", ylab = "Adj. Closing Pri
lines(dates, 8*pIBM, col="red")
lines(dates, 8*pJNJ, col="green")
lines(dates, pSP500, col="black")

title("Prices of 8*Apple, 8*IBM, 8*JNJ, and SP500")
legend(x="topleft", legend=c("8*Apple", "8*IBM", "8*JNJ", "SP500"), lty=c(1, 1), col = c("blue", "red",
```



```r
logApple <- diff(log(pApple))*100
logIBM <- diff(log(pIBM))*100
logJNJ <- diff(log(pJNJ))*100
logSP500 <- diff(log(pSP500))*100

bp <- boxplot(list(logApple, logIBM, logJNJ, logSP500), names = c("Apple", "IBM", "JNJ", "SP500"), ylim=
```




```r
summary(logApple)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -73.12469  -1.15228   0.07750   0.08009   1.38442  13.01943
```

```r
summary(logIBM)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -16.89162  -0.73623   0.01864   0.01376   0.79594  11.35364
```

```r
summary(logJNJ)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -17.25166  -0.52485   0.01784   0.03274   0.61147  11.53729
```

```r
summary(logSP500)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -9.469512 -0.536993  0.052386  0.009644  0.588574 10.957197
```

```r
IQR(logApple)
```

```
## [1] 2.536701
```

```r
sd(logApple)
```

```
## [1] 2.81097
```

```r
IQR(logIBM)
```

```
## [1] 1.532177
```

```r
sd(logIBM)
```

```
## [1] 1.670554
```

```r
IQR(logJNJ)
```

```
## [1] 1.13632
```

```r
sd(logJNJ)
```

```
## [1] 1.222967
```

```r
IQR(logSP500)
```

```
## [1] 1.125568
```

```r
sd(logSP500)
```

```
## [1] 1.253355
```

Ranked from highest risk to lowest: Apple, IBM, SP500, JNJ.

Question 5

The first statement is more probable because the second statement is a subset of the first, which means P(second statement) <= P(first statement). Also, from a non-statistics standpoint, LaGrange, Georgia is a lot smaller and more specific than North America. Georgia is not even a tornado hotspot compared to other places in North America, like Kansas, Oklahoma, etc.

Question 6

a.

Sample space of flipping a coin three times: {HHH, HHT, HTT, HTH, THH, THT, TTT, TTH}

b.

i. A = {HTT, TTH, THT}
ii. B = {TTT, TTH, HTT, THT}
iii. C = {HTT, HTH, HHT, HHH}

c.

P(A) = 3/8, P(B) = 1/2, P(C) = 1/2