Stat 123 Assignment 3

Eric Huber

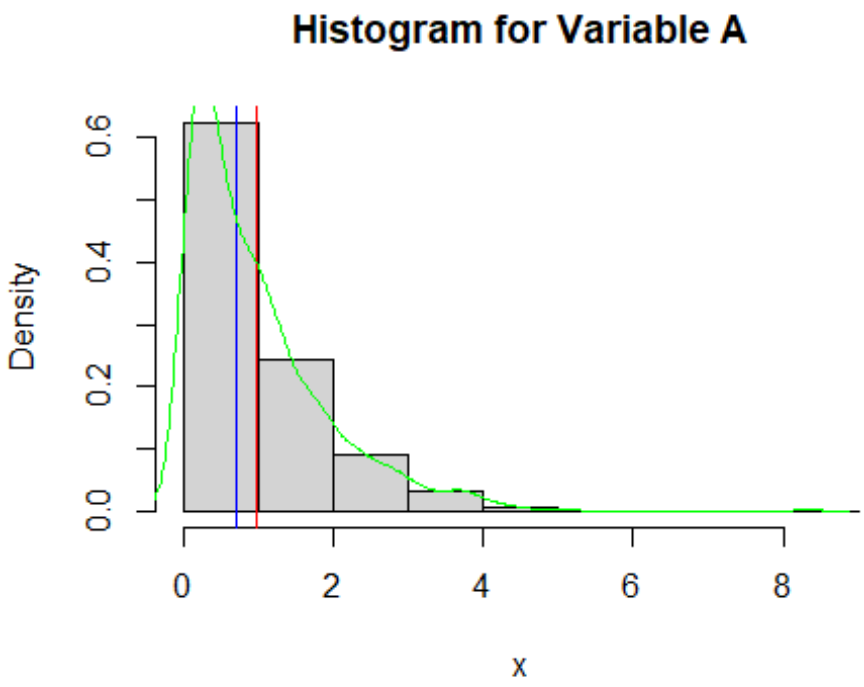
14/03/2022

```
HmwData = read.csv("homework3Data.csv")
```

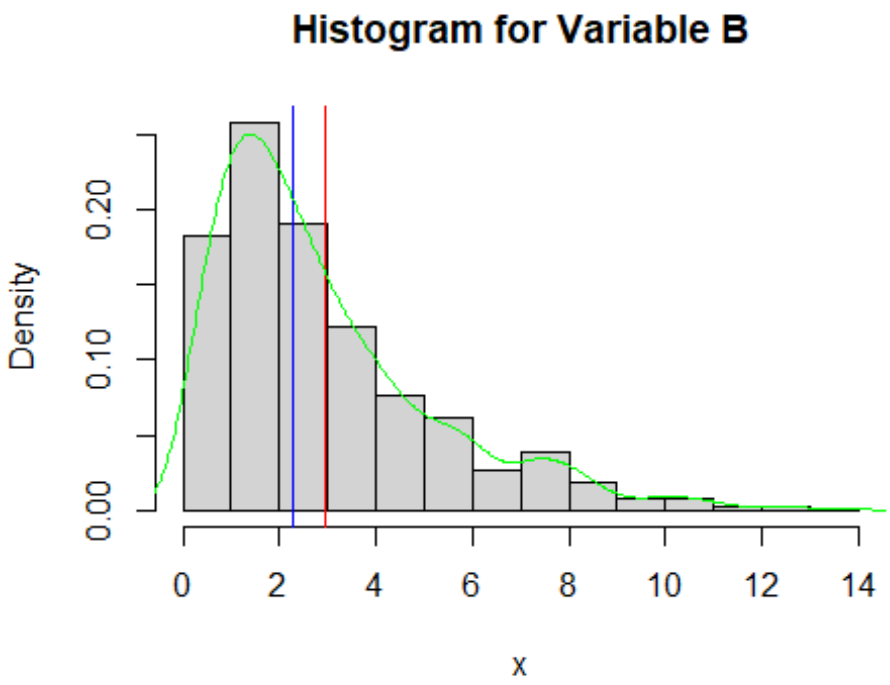
Question 1:

(1.a)

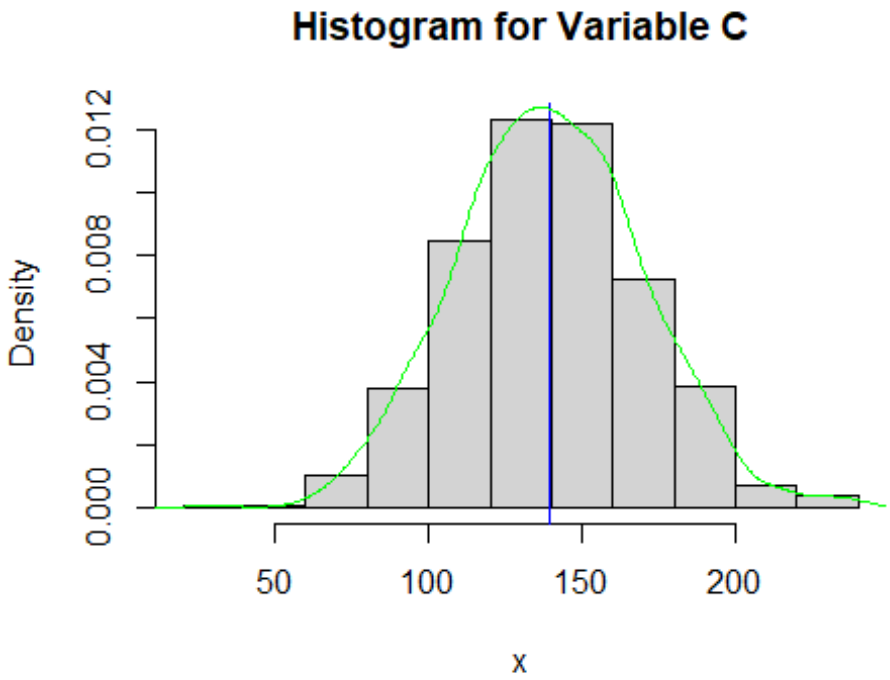
```
hist(HmwData$A, xlab = "x", main = "Histogram for Variable A", prob = TRUE)
abline(v =mean(HmwData$A), col="red")
abline(v = median(HmwData$A), col = "blue")
lines(density(HmwData$A), col = "green")
```



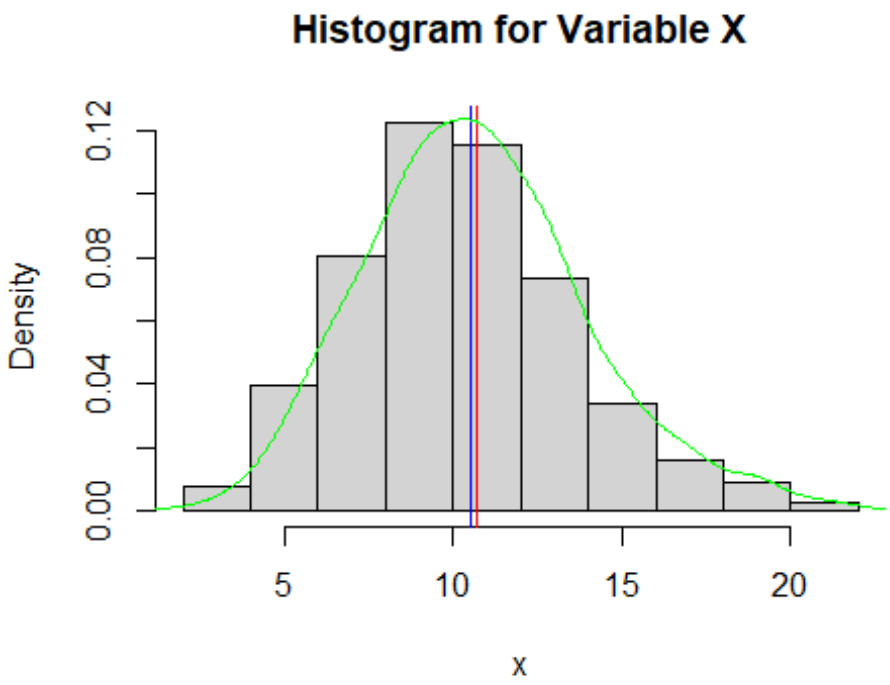

```
hist(HmwData$B, xlab = "x", main = "Histogram for Variable B", prob = TRUE)
abline(v =mean(HmwData$B), col="red")
abline(v = median(HmwData$B), col = "blue")
lines(density(HmwData$B), col = "green")
```



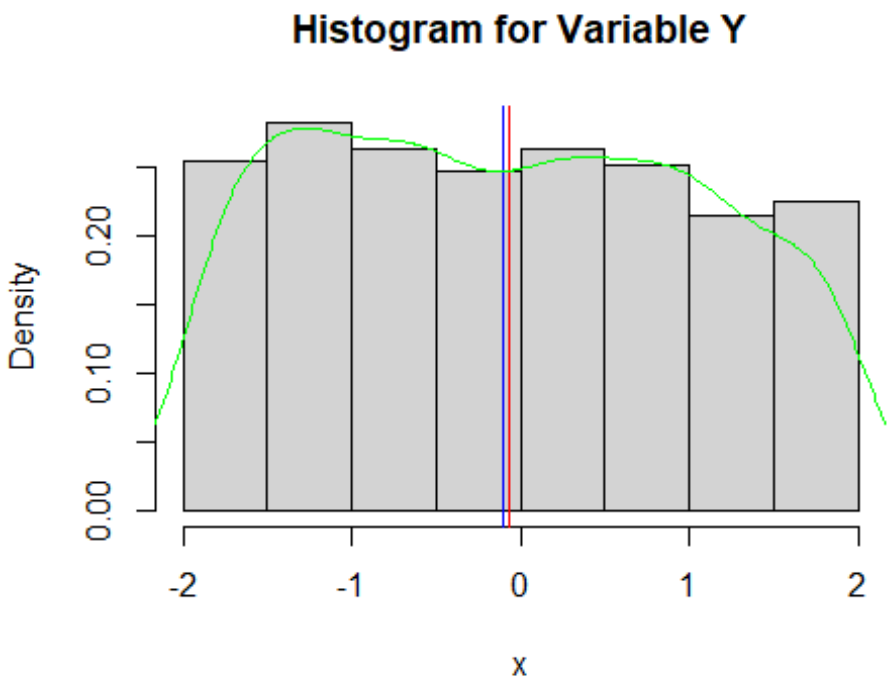

```
hist(HmwData$C, xlab = "x", main = "Histogram for Variable C", prob = TRUE)
abline(v =mean(HmwData$C), col="red")
abline(v = median(HmwData$C), col = "blue")
lines(density(HmwData$C), col = "green")
```



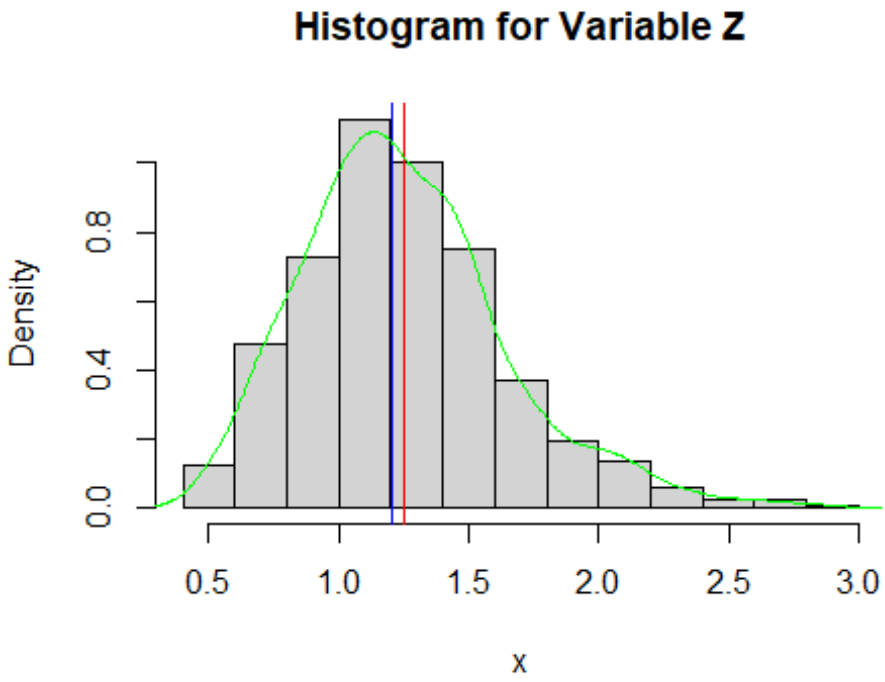

```
hist(HmwData$X, xlab = "x", main = "Histogram for Variable X", prob = TRUE)
abline(v =mean(HmwData$X), col="red")
abline(v = median(HmwData$X), col = "blue")
lines(density(HmwData$X), col = "green")
```




```
hist(HmwData$Y, xlab = "x", main = "Histogram for Variable Y", prob = TRUE)
abline(v =mean(HmwData$Y), col="red")
abline(v = median(HmwData$Y), col = "blue")
lines(density(HmwData$Y), col = "green")
```




```
hist(HmwData$Z, xlab = "x", main = "Histogram for Variable Z", prob = TRUE)
abline(v =mean(HmwData$Z), col="red")
abline(v = median(HmwData$Z), col = "blue")
lines(density(HmwData$Z), col = "green")
```



(1.b)

The variable C appears to be normally distributed. We can see that our histogram is symmetric and our mean and median have the same value. The mean and median is also located in the centre of our density curve which is bell-shaped.

(1.c)

```
std = sd(HmwData$C)

sixtyeight = c((mean(HmwData$C)-std), (mean(HmwData$C)+std))
sixtyeight

## [1] 109.3024 169.9447

ninetyfive = c((mean(HmwData$C)-(2*std)), (mean(HmwData$C)+(2*std)))
ninetyfive

## [1]  78.98131 200.26580

ninetynine = c((mean(HmwData$C)-(3*std)), (mean(HmwData$C)+(3*std)))
ninetynine

## [1]  48.66019 230.58693
```

Approximately 68% of the data falls between ~109.30-169.94, 95% of the data falls between ~78.91-200.27 and 99.7% falls between ~48.660-230.59.

(1.d)

```
quantile(HmwData$C, c(.16,.84))

##      16%      84%
## 110.6479 169.5428

quantile(HmwData$C, c(0.025, 0.975))

##      2.5%     97.5%
##  80.89612 197.69088

quantile(HmwData$C, c(0.0015, 0.9985))

##     0.15%    99.85%
##  61.86316 236.73577
```

We can see that the intervals are close, but definitely not the exact same. The first two (68% & 95%) are very close but the last one (99.7%) is a little further apart

(1.e)

```
xbar = mean(HmwData$C)
qnorm(c(0.16, 0.84), xbar, std )

## [1] 109.4705 169.7766

qnorm(c(0.025, 0.975), xbar, std )

## [1]  80.19525 199.05187

qnorm(c(0.0015, 0.9985), xbar, std )

## [1]  49.63841 229.60870
```

We can see that the use of the qnorm() function gives us very similar output to the method used in part c and relatively similar to the method used in part d.

(1.f)

```
mean(HmwData$C)

## [1] 139.6236

n = length(HmwData$C)

std/sqrt(n)

## [1] 0.9837471

qnorm(c(.06,0.94))

## [1] -1.554774  1.554774

xbar - qnorm(.94)*std/sqrt(n)

## [1] 138.0941
```

```
xbar + qnorm(0.94)*std/sqrt(n)
```

```
## [1] 141.1531
```

Our estimate of the population mean is ~139.6. The estimated standard error of the statistic is ~0.98. Our critical value for an 88% confidence interval is given by ~1.55 and the 88% confidence interval is ~(138.1 - 141.2)

Question 2:

(2.a)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
names = starwars$name
```

(2.b)

```
nchar(names[5])
```

```
## [1] 11
```

```
nchar(names[20])
```

```
## [1] 9
```

```
nchar(names[34])
```

```
## [1] 13
```

(2.c)

```
num_char = numeric()
i = 0
n1 = length(names)

for(i in 1:n1){
  num_char[i] = nchar(names[i])
  i = i+1

}
```

(2.d)

```
num_char = sapply(names, nchar)
```

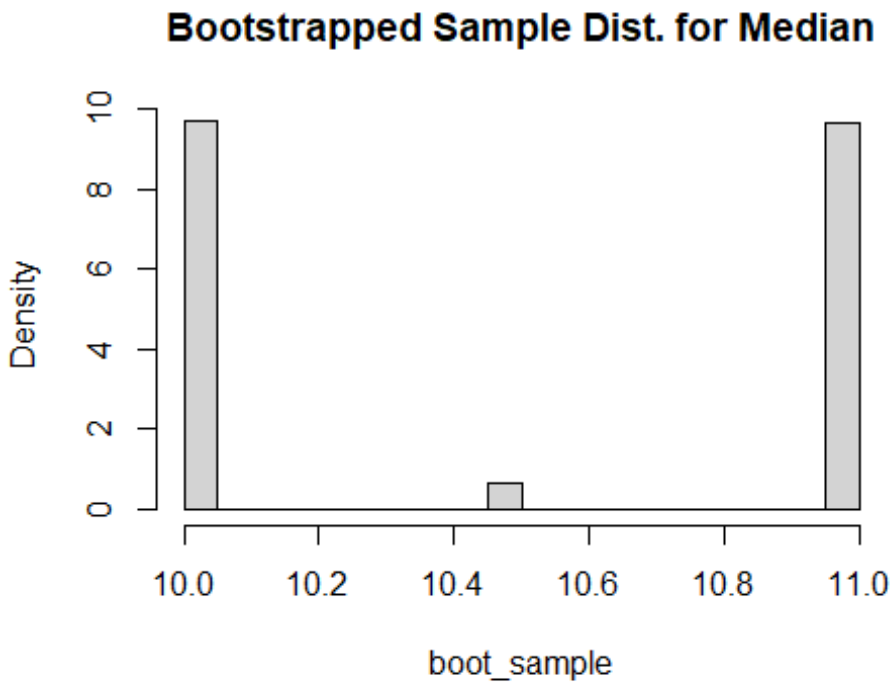
Question 3:

(3.a)

```
boot_sample = numeric()
i = 0
for(i in 1:10000){
boot_sample[i] = median(sample(HmwData$X, 600, replace = TRUE))
i = i+1
}
```

(3.b)

```
hist(boot_sample, main = "Bootstrapped Sample Dist. for Median", prob = TRUE)
```



(3.c)

```
mean(boot_sample)

## [1] 10.498
```

(3.d)

```
quantile(boot_sample, c(0.025,0.975))

##  2.5% 97.5%
##    10    11
```

Our confidence interval is (10,11)