Homework_1_LexBrunett_STAT408

Lex Miguel Brunett Chavez

Applied Regression Analysis

2023-09-15

Please provide detailed calculation and explanation in your solution. Points will be deducted for skimpily written answers. This homework will also require coding in R. On the coding part, the homework solutions should also include detailed description, R code, and output. Write your answers, scan them, and combine to a single pdf file. Name this file as yourname_hw1 and upload to Sakai.

QUESTION #1

1. (5 points) The PMF of the amount of memory X (GB) in a purchased flash drive is.

x	1	2	4	8	16
$p(x)$	.05	.10	.35	.40	.10

Compute the following:

a. E(X):
The expected value of x is determined by:

$$E[X] = \sum_x x\,p(x)$$

$$E[X] = 1(0.05) + 2(0.10) + 4\,(0.35) + 8(0.40) + 16(0.10) = 6.45$$

$$E[X] = 6.45$$

b. V(X) directly from the definition:
The variance of x is determined by following equation:

$$V[X] = E[(X - E[X])^2]$$

$$V[X] = \sum_x (X - 6.65)^2\,p(x)$$

$$V[X] = (1 - 6.65)^2 0.05 + (2 - 6.65)^2 0.10 + (4 - 6.65)^2 0.35 + (8 - 6.65)^2 0.40 + (16 - 6.65)^2 0.10$$

$$V[X] = 15.6875$$

$$V[X] = 15.6875$$

c. The standard deviation of X:

$$\sqrt{V[X]} = 3.957$$

QUESTION #2

2. (5 points) Consider the following sample of observations on coating thickness for low viscosity paint:

.83	.88	.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

a. Calculate a point estimate of the mean value of coating thickness, and state which estimator you used:
As it is a sample and not a population the estimator used is $\bar{X}$.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_I$$

$$n = 16$$

$$\bar{X} = \frac{\begin{array}{c}0.83 + 0.88 + 0.88 + 1.04 + 1.09 + 1.12 + 1.29 + 1.31 \\ +1.48 + 1.49 + 1.59 + 1.62 + 1.65 + 1.71 + 1.76 + 1.83\end{array}}{16}$$

$$\bar{X} = \frac{21.57}{16} = 1.35$$

b. Calculate a point estimate of the variance of coating thickness, and state which estimator you used.
As it is a sample the estimator used is s^2 for sample variance.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$As\ \bar{X} = 1.35$$

$$s^2 = \frac{\begin{array}{c}(0.83 - 1.35)^2 + (0.88 - 1.35)^2 + (0.88 - 1.35)^2 + (1.04 - 1.35)^2 \\ +(1.09 - 1.35)^2 + (1.12 - 1.35)^2 + (1.29 - 1.35)^2 + (1.31 - 1.35)^2 \\ +(1.48 - 1.35)^2 + (1.49 - 1.35)^2 + (1.59 - 1.35)^2 + (1.62 - 1.35)^2 \\ +(1.65 - 1.35)^2 + (1.71 - 1.35)^2 + (1.76 - 1.35)^2 + (1.83 - 1.35)^2\end{array}}{16 - 1}$$

$$s^2 = \frac{1.7191}{15} = 0.115$$

QUESTION #3

3. (5 points) A confidence interval is desired for the true average stray-load loss $\mu\mu$ (watts) for a certain type of induction motor. Assume that stray-load loss is normally distributed with $\sigma\sigma$ = 3.

   a. Compute a 95% CI for μ when n = 25 and $\bar{x}$= 58.3

   b. Compute a 95% CI for μ when n = 100 and $\bar{x}$= 58.3

   c. Compute a 99% CI for μ when n = 25 and $\bar{x}$= 58.3

```
#Knowing sigma and not mean, we can use this statement for the CI following the equation:

A confidence interval for the population mean, based on a simple random sample (SRS) of
size n, is determined  by where z* is the upper (1-C)/2 critical value for the
standard normal distribution.
```

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

```
# First confidence Interval, with n = 25 and  x¯= 58.3
x_bar_first <- 58.3
n_first <- 25
sigma_first <- 3
confidence_first <-95
alpha_first <- 1 - confidence_first/100

z_firt <- alpha_first/2
error_first <- z_firt*(sigma_first/sqrt(n_first))


# Second confidence Interval, with n = 100 and  x¯= 58.3

x_bar_second <- 58.3
n_second <- 100
sigma_second <- 3
confidence_second <-95
alpha_second <- 1 - confidence_second/100

z_second <- alpha_first/2
error_second <- z_second *(sigma_second/sqrt(n_second))

# Third confidence Interval, with n = 25 and  x¯= 58.3

x_bar_third <- 58.3
n_third <- 25
sigma_third <- 3
confidence_third <-99
alpha_third <- 1 - confidence_third/100

z_third <- alpha_first/2
error_third <- z_third * (sigma_third/sqrt(n_third))
```

Answers

a) 58.3 +- 0.015

A 95% CI for μ = (57.7, 64.3)

Therefore, we have a 95% confidence that the mean stray-load loss $\mu\mu$ (watts) for a certain type of induction motor with n = 25 and $\bar{x}$= 58.3 is likely to reside between 57.7 to 64.3 watts.

b) 58.3 +- 0.0075

A 95% CI for μ = CI: (58, 58.6)

Therefore, we have a 95% confidence that the mean stray-load loss $\mu\mu$ (watts) for a certain type of induction motor with n = 100 and $\bar{x}$= 58.3 is likely to reside between 58 to 58.6 watts.

c) 58.3 +- 0.015

A 99% CI for μ = (57.7, 64.3)

Therefore, we have a 95% confidence that the mean stray-load loss $\mu\mu$ (watts) for a certain type of induction motor with n = 25 and $\bar{x}$= 58.3 is likely to reside between 57.7 to 64.3 watts.

QUESTION #4

4. (5 points) To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected, and tests are conducted on each weld in the sample. Suppose the specifications state that the mean strength of welds should exceed 100 lb/in2:

H0: The mean strength of the sample is > 100 lb/in2 (As the exercise details that the pipe should exceed 100 lb/In2 to meet specifications)

Ha: The mean strength of the selected welds is ≤ 100 lb/in2 (The alternative Hypothesis states that the mean of the sample doesn't meet the specifications of exceeding 100 lb/in2)

Therefore:

H0: μ>μ_0

Ha: μ<μ_0

In case the p-value is less or equal to the Significance level, we would have enough statistic evidence to reject the null hypothesis, meaning that the mean of the sample does not meet the specifications.

Type I error: Null hypothesis is rejected but the mean strength of the sample is > 100 lb/in2, meaning that is true.

Type II error: Null hypothesis is not rejected but the mean strength of the sample is< 100 lb/in2 meaning that is false that the pipes selected have meet the requirements.

QUESTION #5

(10 points) Dataset births.csv contains the information for 1992 newborns and their parents.

```
# a) Download the data set births.csv from Sakai, set your working directory, and import
it into RStudio. Name the data frame as NCbirths.

NCbirths <-
read.csv("https://raw.githubusercontent.com/LexoBrunett/STAT408_LexoBrunett/main/Datasets
/births.csv")


# b) Extract the weight variable as a vector from the data frame and name it as weights.

What units do you think the weights are in?


weights <- NCbirths$weight



# c) Create a new vector named weights_in_pounds which are the weights of the babies in

pounds. You can look up conversion factors on the internet.

weights_in_pounds <- weights/16

# d) Print the first 20 babies' weight in pounds.


weights_in_pounds[1:20]

[1] 7.7500

[2] 11.0625

[3] 6.6875

[4] 9.0000

[5] 7.3125

[6] 6.1250

[7] 9.1875

[8] 8.6250

[9] 6.5000
[10] 7.6875

[11] 9.5625

[12] 8.0625
```

```
[13] 7.4375

[14] 6.7500

[15] 6.6250

[16] 7.8125

[17] 7.1875

[18] 8.0000
[19] 8.2500

[20] 5.1875

# e) What is the mean weight of all babies in pounds?

mean(weights_in_pounds)

The mean weight of all babies is 7.2532 pounds

# f) The habit variable records the smoking status for mothers of each baby. What

percentage of the mothers in the sample smoke? Hint: consider table() function.

prop.table(table(NCbirths$Habit))*100


 NonSmoker     Smoker
  90.61245%   9.38755%


Therefore, 90.61% of the 1992 mothers of the sample weren't smokers, while an approximate
of 9.38% of mothers were smokers.


# g) According to the Centers for Disease Control, approximately 14% of adult Americans

are smokers. How far off is the percentage you found in (b) from the CDC's report?
```

I found that there were an approximate of 9% of smoker moms in 1996, if the CDC's report dates from 1996 too, it shows that only a 5% of American men and women with no kids are smokers. If the sample of mothers was a representative sample instead of a simple random sample, these results may indicate a dangerous smoking culture between 1992 mothers.

QUESTION #6

(10 points) The dataset flint.csv records the water pollution levels in different locations at Flint, Michigan.

```
# a) Download the flint.csv from Sakai and read it into R. When you read in the data,
name your object "flint".

flint <-
read.csv("https://raw.githubusercontent.com/LexoBrunett/STAT408_LexoBrunett/main/Datasets
/flint.csv")
```

```
# b) The EPA states a water source is especially dangerous if the lead level (Pb) is 15
PPB or greater. What proportion of the locations tested were found to have dangerous lead
levels?

#Making the for loop to capture the value of the percentage
loc_PB_Great = 0
for (x in flint$Pb) {
  if (!is.na(x) & (x >= 15)){
    loc_PB_Great  <- loc_PB_Great + 1
  }
}


#Once obtained the value, compare it with the total amount of regions

locations = length(flint$Region)
percentage = 100*(loc_PB_Great/locations)

## [1] 4.436229%
```

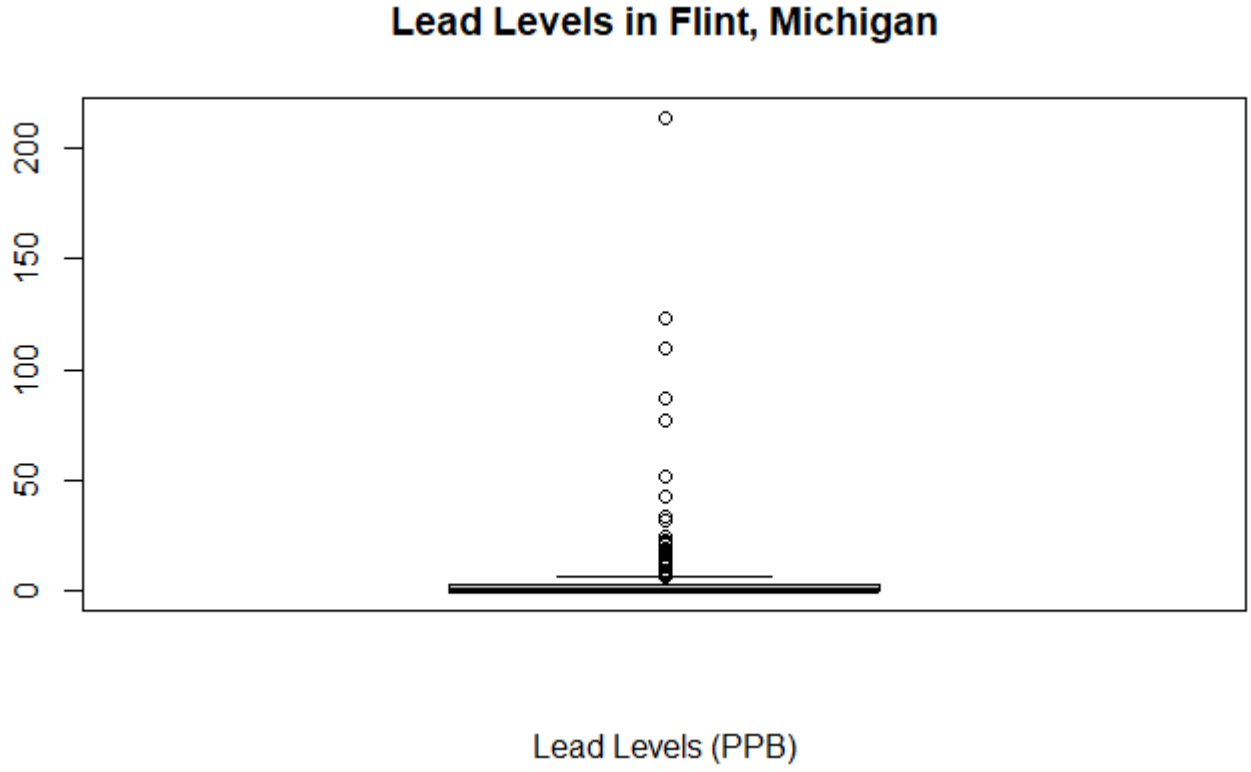
Approximately 4.43% of locations in Flint, Michigan, have dangerous lead levels in water.

```
#c) Report the mean copper level for only test sites in the North region


copper_regions <- flint[flint$Region == "North", ]

mean(copper_regions$Cu)

## [1] 44.6424

The mean value of copper levels only in the sites of the north region is 44.6424.

#d) Report the mean copper level for only test sites with dangerous lead levels (at least
15 PPB).

dangerous_Pb_Region <- flint[flint$Pb >= 15, ]

mean(dangerous_Pb_Region$Cu)

## [1] 305.8333

The mean copper levels of the locations with dangerous lead levels is approximately
305.83



#e) Report the mean lead and copper levels for all locations.

mean(flint$Pb)

## [1] 3.383272

The mean lead level for all locations is 3,38 PPB, which indicates that the average
location in Flint, Michigan doesn't have dangerous lead levels.

mean(flint$Cu)
```

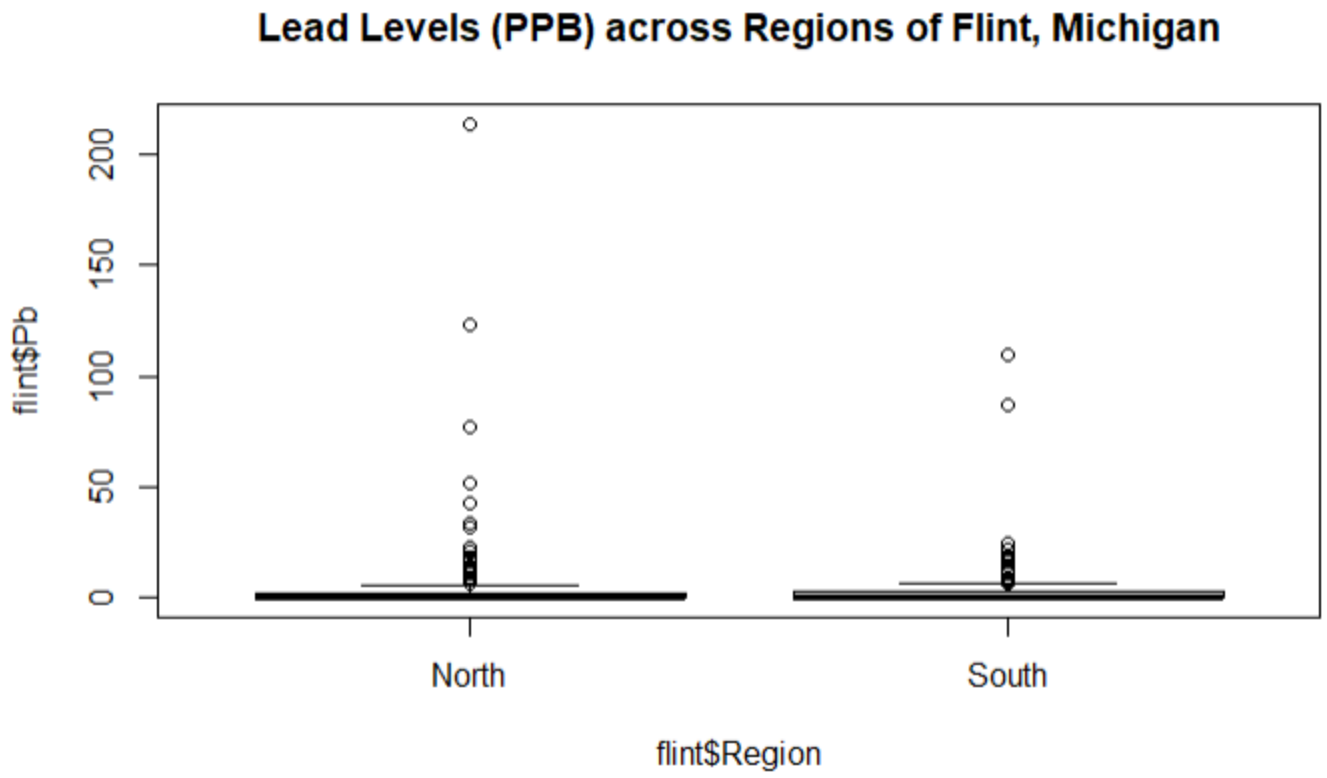
```
## [1] 54.58102
```

The mean copper level for all locations is 54.58

```
#f) Create a box plot with a good title for the lead levels. Hint: consider boxplot()
function.

boxplot(x=flint$Pb, , main=" Lead Levels in Flint, Michigan", xlab="Lead Levels (PPB)")
```




```
#g) Based on what you see in part (f), does the mean seem to be a good measure of center
for the data? Report a more useful statistic for this data.

boxplot(flint$Pb~flint$Region, , main = "Lead Levels (PPB) across Regions of Flint,
Michigan")
```

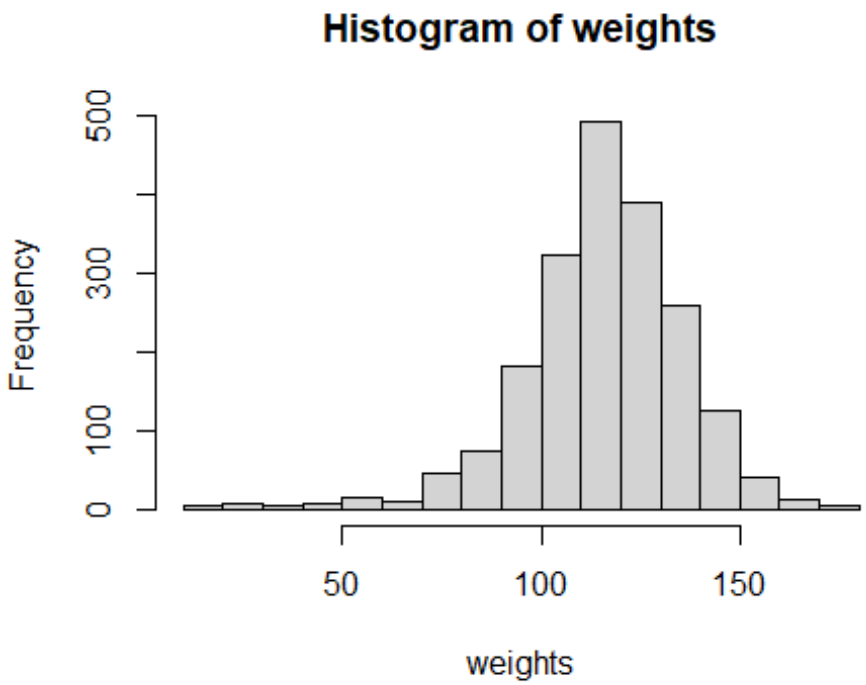


Based on the average, we can observe most of the values are approximately below 5 PPB of lead level, we can assure that the mean value is a good measure of data.

QUESTION #7

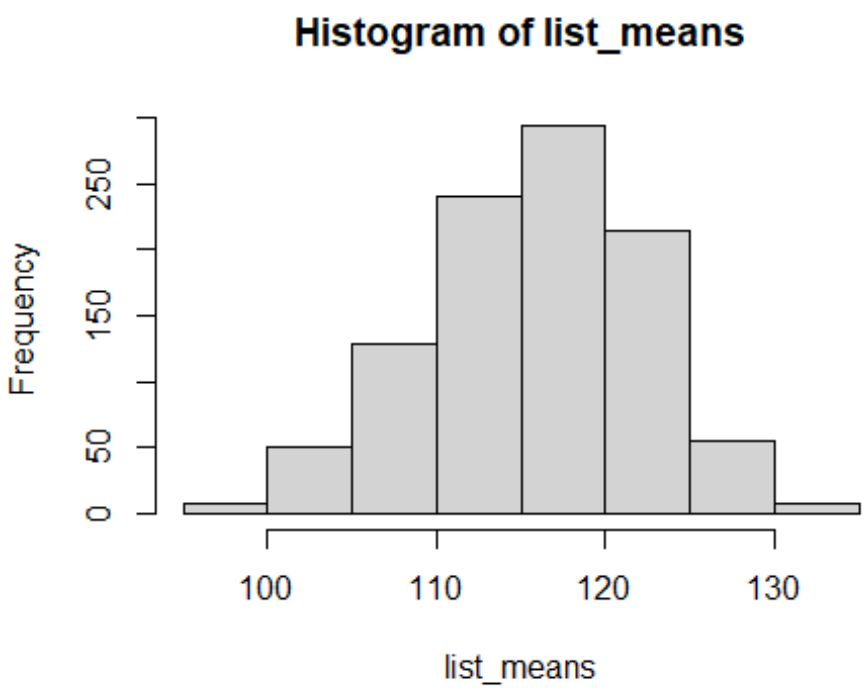
(10 points) We will use a simulation study to show central limit theorem.

```
#a) Use hist() function to plot a histogram on the weight variable in the NCbirths. Do
you think weight follows a normal distribution? Why?

hist(weights)
```



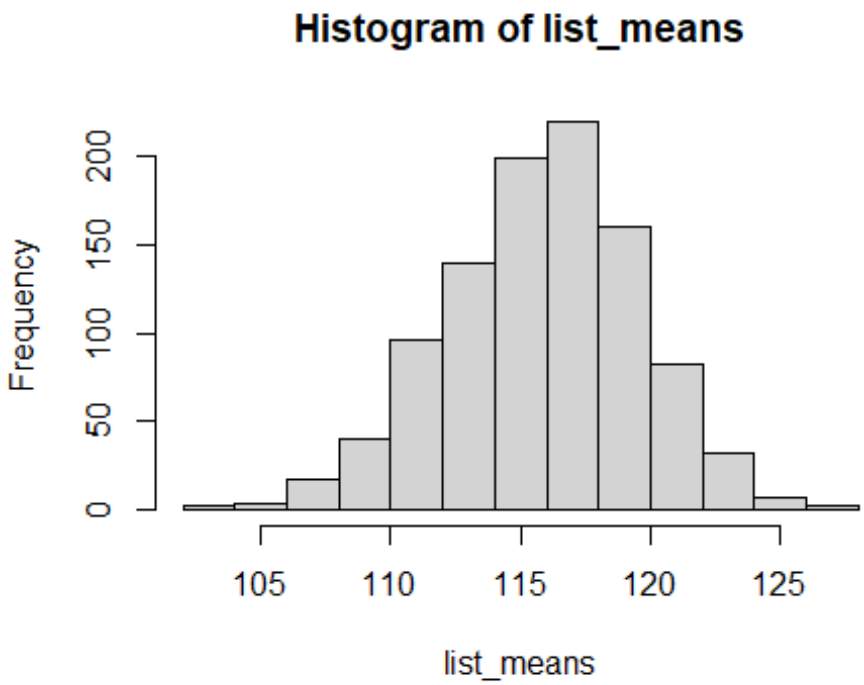
The Histogram doesn't follow a normal distribution, as it doesn't have a bell shape because it is left skewed, meaning that the mean is closer to the left, therefore, there is a greater number of babies with a weight between 90 and 140 ounces.

```
#b) Use sample() function to randomly select 10 observations from weight. Show the mean

of these 10 observations

sample_weight <- sample(weights, 10)

mean(sample_weight)

## [1] 119.5

The mean weight of newborns of a 10 size random sample is 119.5 ounces.

#c) Use a for loop to repeat the (b) 1000 times. Save 1000 means in a vector. Show the

histogram for 1000 means. Is this distribution close to normal?

list_means <- c()
for (x in 1:1000) {
  sample_loop <- sample(weights, 10)
  sample_loop <- mean(sample_loop)
  list_means = c(list_means, sample_loop)
}

hist(list_means)
```

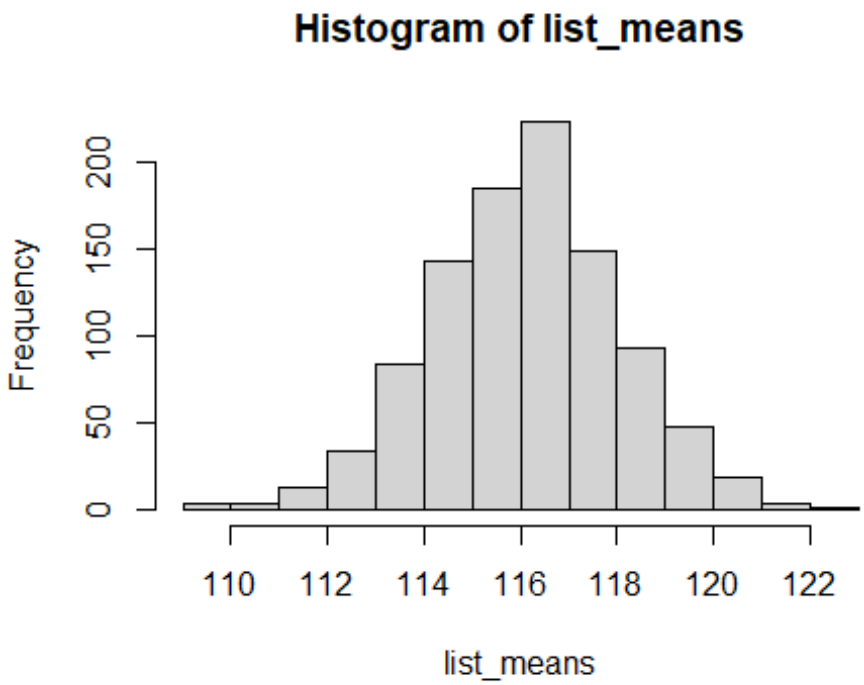


This graphics is closer to the normal distribution, comparing from the last histogram, the mean is more centered, therefore the mean of the weight on 1000 times 10 samples are from 115 to 120 ounces.

```
#c) Change the sample size 10 in (b) to 30 and 100, Repeat (c) for these two sample
sizes.

# sample with 30 size



sample_weight <- sample(weights, 30)

mean(sample_weight)

## [1] 116.1667

list_means <- c()
for (x in 1:1000) {
  sample_loop <- sample(weights, 30)
  sample_loop <- mean(sample_loop)
  list_means = c(list_means, sample_loop)
}

hist(list_means)
```




```r
# sample with 100 size



sample_weight <- sample(weights, 100)

mean(sample_weight)

## [1] 117.71

list_means <- c()
for (x in 1:1000) {
  sample_loop <- sample(weights, 100)
  sample_loop <- mean(sample_loop)
  list_means = c(list_means, sample_loop)
}

hist(list_means)
```




```
#Are these two distributions close to normal? Interpret your reason.
```

Yes, these two distributions (list of means with 30 and 100 samples respectively) are closest to a normal distribution, giving a gauss bell with the average of 115 to 118 ounces for the 30 random samples repeated 1000 times. In the other hand we have 116 to 117 ounces for the 100 random samples repeated 1000 times.