Stat408_HW1_Ghosh

Akhil Ghosh

2022-09-18

Q1:

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)

b. V(X) directly from the definition

c. The standard deviation of X

Figure 1: A caption

a)

```
#E(x) = sum of (xi * p(xi)) for every xi (for discrete pmf)
#E(x) = 1 * 0.05 + 2 * 0.10 + 4 * 0.35 + 8 * 0.4 + 16 * .01
E_x <- 1 * 0.05 + 2 * 0.10 + 4 * 0.35 + 8 * 0.4 + 16 * .01
print(E_x)
```

```
## [1] 5.01
```

E(X) = 5.01

b)

```
#V(x) = E(X-mu)^2 = E(X^2) - (E(x))^2
#V(x) = 1 * 0.05 + 4 * 0.10 + 16 * 0.35 + 64 * 0.4 + 256 * .01 - (5.01^2)
V_x <- 1 * 0.05 + 4 * 0.10 + 16 * 0.35 + 64 * 0.4 + 256 * .01 - (5.01^2)
print(V_x)
```

```
## [1] 9.1099
```

$\mathrm{V(X)} = 9.1099$

c)

```
#SD = sqrt(variance)
sd <- sqrt(V_x)
sd
```

```
## [1] 3.018261
```

$\mathrm{SD} = 3.018$

Q2:

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
b. Calculate a point estimate of the variance of coating thickness, and state which estimator you used

Figure 2: A caption

a)

```r
#Point estimate for mean value of coating thickness using sample
sample <- 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)
X_bar <- mean(sample)
X_bar
```

```
## [1] 1.348125
```

point estimate used is X_bar, which is unbiased. X_bar is 1.348125

b)

```r
S <- var(sample)
S
```

```
## [1] 0.1146029
```

point estimate used is S, which is unbiased point estimator for population variance. S is 0.1146029

Q3:

> 3. (5 points) A confidence interval is desired for the true average stray-load loss μ (watts) for a certain type of induction motor. Assume that stray-load loss is normally distributed with $\sigma = 3$.
>
> 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 $\bar{x} = 58.3$

Formula for finding a 95% CI interval with normal distribution given by below image

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

\begin{figure}
\caption{95% CI formula} \end{figure}

a)

```
x_bar_q3 <- 58.3
sd_q3 <- 3
n1 <- 25
z_score95 <- round(qnorm(0.025,lower.tail=FALSE),2)
CI_1 <- c((x_bar_q3 - (z_score95*sd_q3/sqrt(n1))), (x_bar_q3 + (z_score95*sd_q3)/sqrt(n1)))
CI_1
```

```
## [1] 57.124 59.476
```

```
#Can also use function
```

95% CI for mu when n=25 and X_bar = 58.3 is (57.124, 59.476)

b)

```
n2 <- 100

CI_2 <- c((x_bar_q3 - (z_score95*sd_q3/sqrt(n2))), (x_bar_q3 + (z_score95*sd_q3)/sqrt(n2)))
CI_2
```

```
## [1] 57.712 58.888
```

95% CI for mu when n=100 and X_bar = 58.3 is (57.712, 58.888)

c)

```
#99% CI is the same as in image above, but instead of 1.96 the z-score value is 2.58
z_score99 <- round(qnorm(0.005,lower.tail=FALSE),2)
CI_3 <- c((x_bar_q3 - (z_score99*sd_q3/sqrt(n1))), (x_bar_q3 + (z_score99*sd_q3)/sqrt(n1)))
CI_3
```

```
## [1] 56.752 59.848
```

99% CI for mu when n=25 and x_bar = 58.3 is (56.752, 59.848)

Q4:

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/in^2

a. What hypotheses should be tested? Write down H_0 and H_a and explain your reason.
b. Describe type I and II errors in the context of this problem situation.

a)

$$H_0 : \mu = 100 \ lb/in^2$$

$$H_a : \mu > 100 \ lb/in^2$$

Under this set of hypotheses, we are testing if the mean strength of welds exceed 100. I wrote it this way because under this, we assume that the null is true and that mean weld strength is equal to 100. The alternative in this scenario would be that mean strength exceeds 100, and we would look at the distribution of the test stat in order to determine the probability that mean strength is 100 or less. If lower than a designated significance level, α, we would reject the null.

b)

Type 1 error, represented by α, is the probability that given the null is actually true, we end up falsely rejecting the null hypothesis. In the context of this problem, it is the probability of rejecting the null when the actual mean strength of welds is equal to 100 or lower.

Type 2 error, represented by $1 - \alpha$, or β, is the probability that given the null is actually false but we fail to reject it. In the context of this problem, it is the probability that we don't reject the null given that the strength of welds is actually exceeds 100.

Q5:

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.

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?

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.

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

e. What is the mean weight of all babies in 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.

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?

a)

```
NCbirths <- read_csv("births.csv")
```

```
## Rows: 1992 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): Gender, Premie, Marital, Racemom, Racedad, Hispmom, Hispdad, Habit...
## dbl  (9): weight, Apgar1, Fage, Mage, Feduc, Meduc, TotPreg, Visits, Gained
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

b)

```
weights <-  NCbirths$weight
head(weights)
```

```
## [1] 124 177 107 144 117  98
```

I think weights are most likely in measured in ounces, considering there are many 3 digit weight values.

c)

```
#Conversion: 16 oz in 1 pound
oztolb <- 16
weights_in_pounds = weights/oztolb
```

d)

```
head(weights_in_pounds, 20)
```

```
##  [1]  7.7500 11.0625  6.6875  9.0000  7.3125  6.1250  9.1875  8.6250  6.5000
## [10]  7.6875  9.5625  8.0625  7.4375  6.7500  6.6250  7.8125  7.1875  8.0000
## [19]  8.2500  5.1875
```

e)

```
mean(weights_in_pounds)
```

```
## [1] 7.2532
```

Mean weight is 7.25 pounds

f)

```
smoke_stat <- table(NCbirths$Habit)
smoke_percent <- smoke_stat[2]/sum(smoke_stat)
smoke_percent
```

```
##    Smoker
## 0.0938755
```

About 9.4% of all mothers in the sample reported smoking during pregnancy

g)

```
popsmoke <- 0.14
smoke_diff <- as.vector(popsmoke - smoke_percent)
smoke_diff
```

```
## [1] 0.0461245
```

The difference b/w population smoke percentage and smoke percent in the sample was about 4.6%

Q6:

a)

```
flint <- read_csv("flint.csv")
```

```
## Rows: 541 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Region
## dbl (4): Latitude, Longitude, Pb, Cu
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(flint)
```

```
## # A tibble: 6 x 5
##   Latitude Longitude    Pb    Cu Region
##      <dbl>     <dbl> <dbl> <dbl> <chr>
## 1     43.1     -83.6     0     0 North
## 2     43.1     -83.7     0   130 North
## 3     43.1     -83.7     4   170 North
## 4     43.1     -83.8     0     0 North
## 5     43.1     -83.7     0     0 North
## 6     43.1     -83.7     0     0 North
```

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”.

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?

c. Report the mean copper level for only test sites in the North region.

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

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

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

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.

Figure 3: A caption

b)

```
mean(flint$Pb>=15)
```

```
## [1] 0.04436229
```

About 4.44% of the locations tested had dangerous lead levels of at least 15 PPB

c)

```
clr <- flint %>%
  group_by(Region) %>%
  summarize("Average Copper Levels by Region" = mean(Cu))
clr
```

```
## # A tibble: 2 x 2
##   Region `Average Copper Levels by Region`
##   <chr>                               <dbl>
## 1 North                                44.6
## 2 South                                63.8
```

Average copper levels for Northern region is 44.642

d)

```
clp <- flint %>%
  group_by(Pb>=15) %>%
  summarize("Average Copper Levels by Region" = mean(Cu))

clp
```

```
## # A tibble: 2 x 2
##   `Pb >= 15` `Average Copper Levels by Region`
##   <lgl>                                  <dbl>
## 1 FALSE                                   42.9
## 2 TRUE                                   306.
```

Average copper levels for areas with lead levels higher than 15 PPB is 305.833

e)

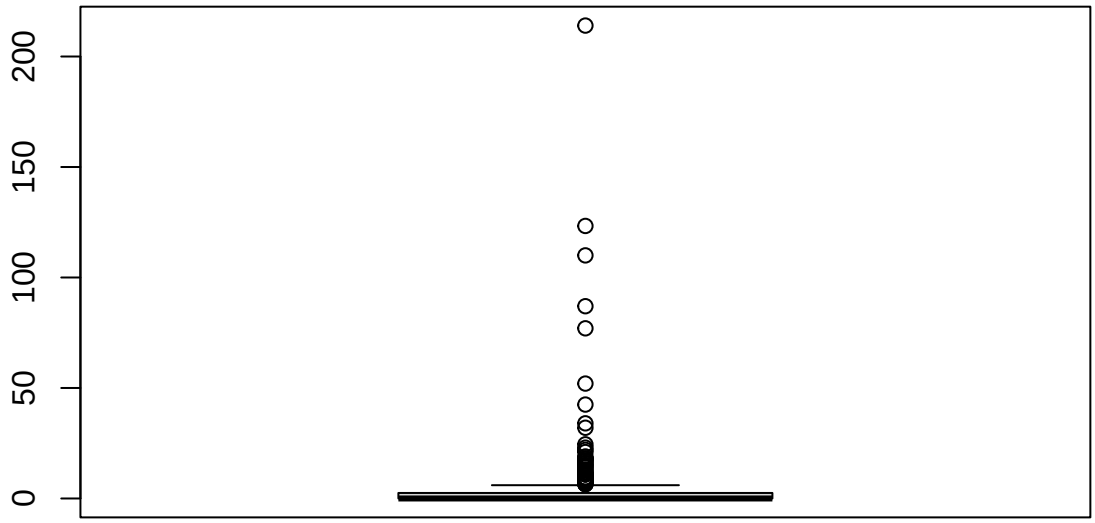
```
flint %>%
  summarize('Average Lead levels'= mean(Pb),"Average Copper Levels" = mean(Cu))
```

```
## # A tibble: 1 x 2
##   `Average Lead levels` `Average Copper Levels`
##                   <dbl>                   <dbl>
## 1                  3.38                    54.6
```

Across all locations, the average lead levels is 3.383 PPB and average copper levels is 54.581

f)

```
boxplot(flint$Pb, Title = "Boxplot of lead levels")
```



g)

Given the results of the boxplot, mean is not a very useful statistic for measuring the center of lead levels for the dataset This is because the average is pulled up by a few outlier values that are well above the center. A more useful stat for the measuring the center of lead levels would be median

```
median(flint$Pb)
```

```
## [1] 0
```

Median value for this data set is 0, which makes sense since a vast majority of locations don't have any lead levels.

Q7:

7. (10 points) We will use a simulation study to show central limit theorem.

a. Set random seed to 2022. Use hist() function to plot a histogram on the weight variable in the NCbirths. Do you think weight follows a normal distribution? Why?

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

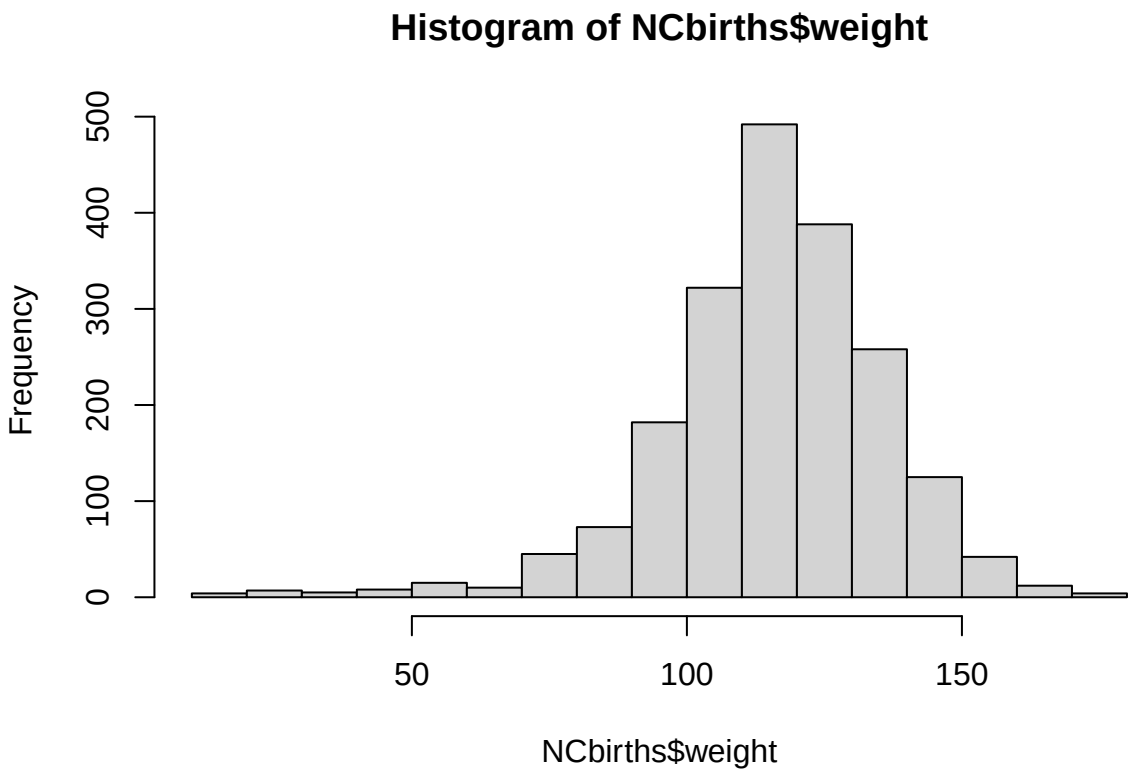
c. Use a for loop to repeat the (b) 1000 times. Save 1000 means in a vector. Show the histogram for 1000 means. Is this distribution close to normal?

d. Change the sample size 10 in (b) to 30 and 100, Repeat (c) for these two sample sizes. Are these two distributions close to normal? Interpret your reason.

Figure 4: A caption

a)

```
set.seed(2022)
hist(NCbirths$weight)
```



Weight of newborns does not follow a normal distribution. Instead the distribution seems to be left skewed.

b)

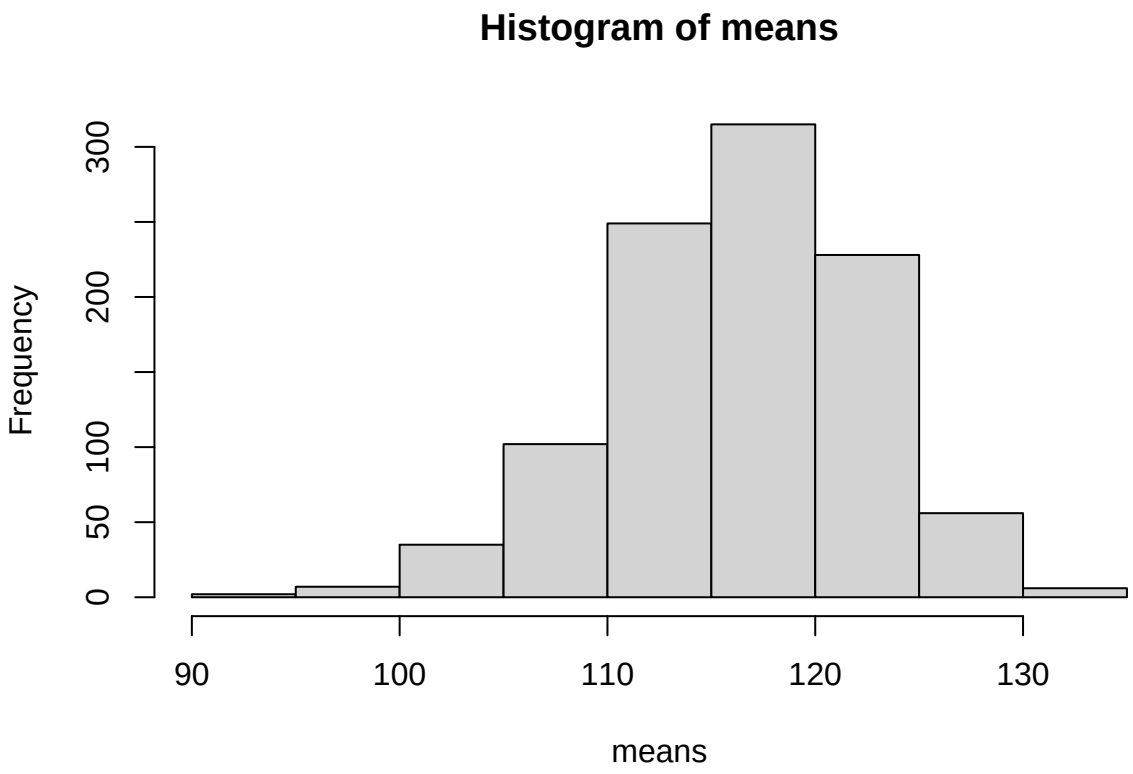
```
set.seed(2022)
mean(sample(NCbirths$weight,10))
```

```
## [1] 119
```

Mean of this sample is 119 ounces.

c)

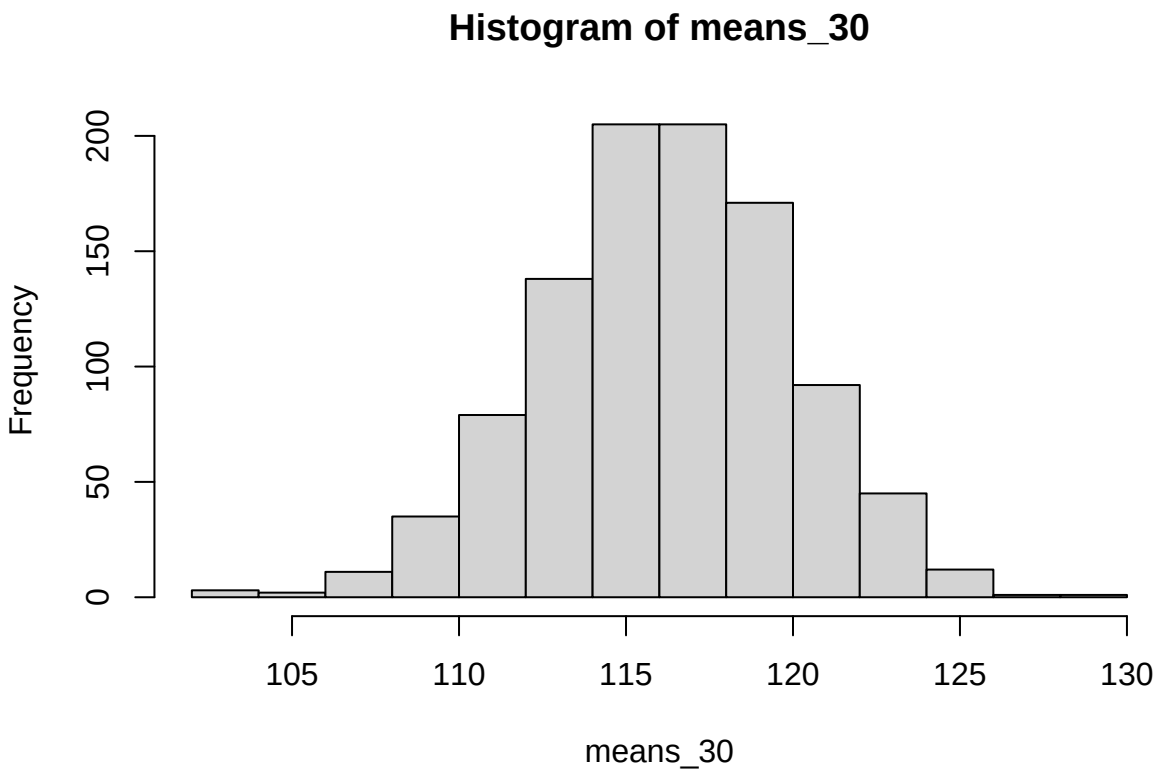
```
set.seed(2022)
means <- c()
for(i in 1:1000){
  # sample_mean <- mean(sample(NCbirths$weight,10))
  means <- append(means,mean(sample(NCbirths$weight,10)))
}
hist(means)
```



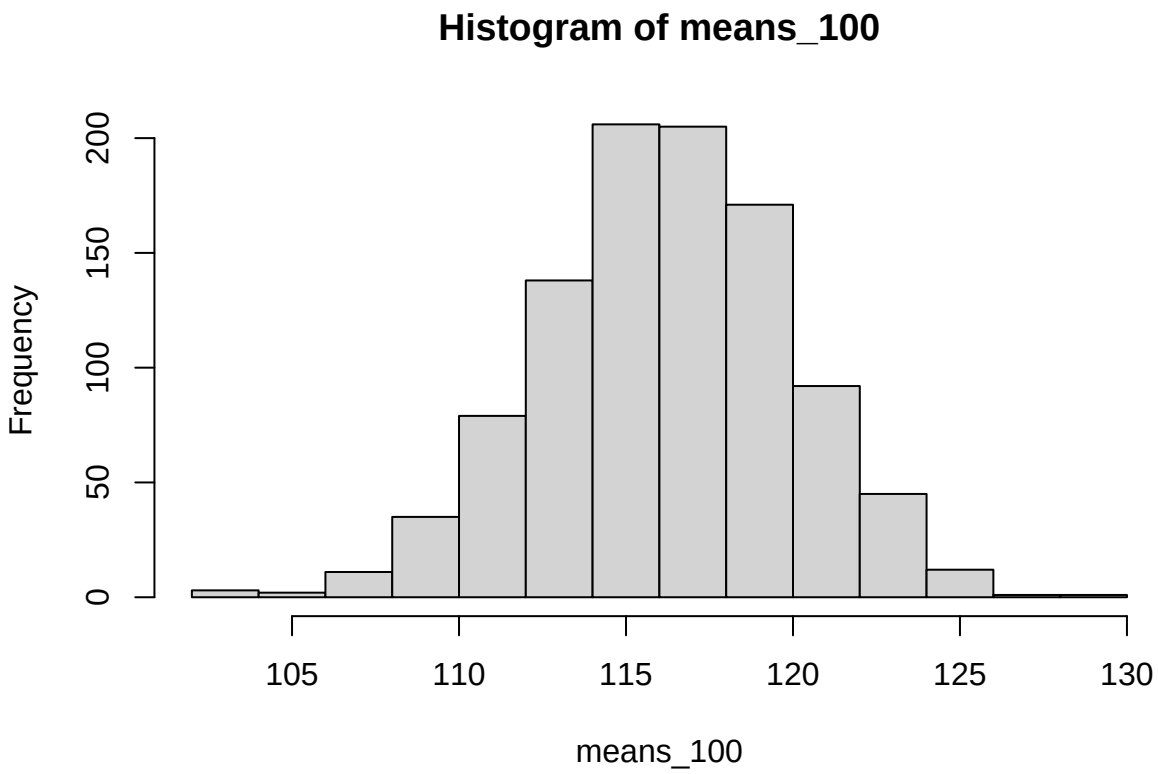
The means of the weight is doesn't seem to follow a normal distribution, as the means of the weights seem to still be left skewed

d)

```
set.seed(2022)
means_30 <- c()
for(i in 1:1000){
  # sample_mean <- mean(sample(NCbirths$weight,10))
  means_30 <- append(means_30,mean(sample(NCbirths$weight,30)))
}
hist(means_30)
```




```
means_100 <- c()
for(i in 1:1000){
  means_100 <- append(means_30,mean(sample(NCbirths$weight,100)))
}
hist(means_100)
```



After increasing the sample sizes to 30 and 100 respectively, the distribution of the weights becomes normal or approximately normal. This is because as stated by CLT, since the sample size of weights is now more than or equal to 30, the sample mean is normally distributed with mean mu and variance of sdˆ2/n