Central Limit Theorem

GROUP 9

M5 ICA2

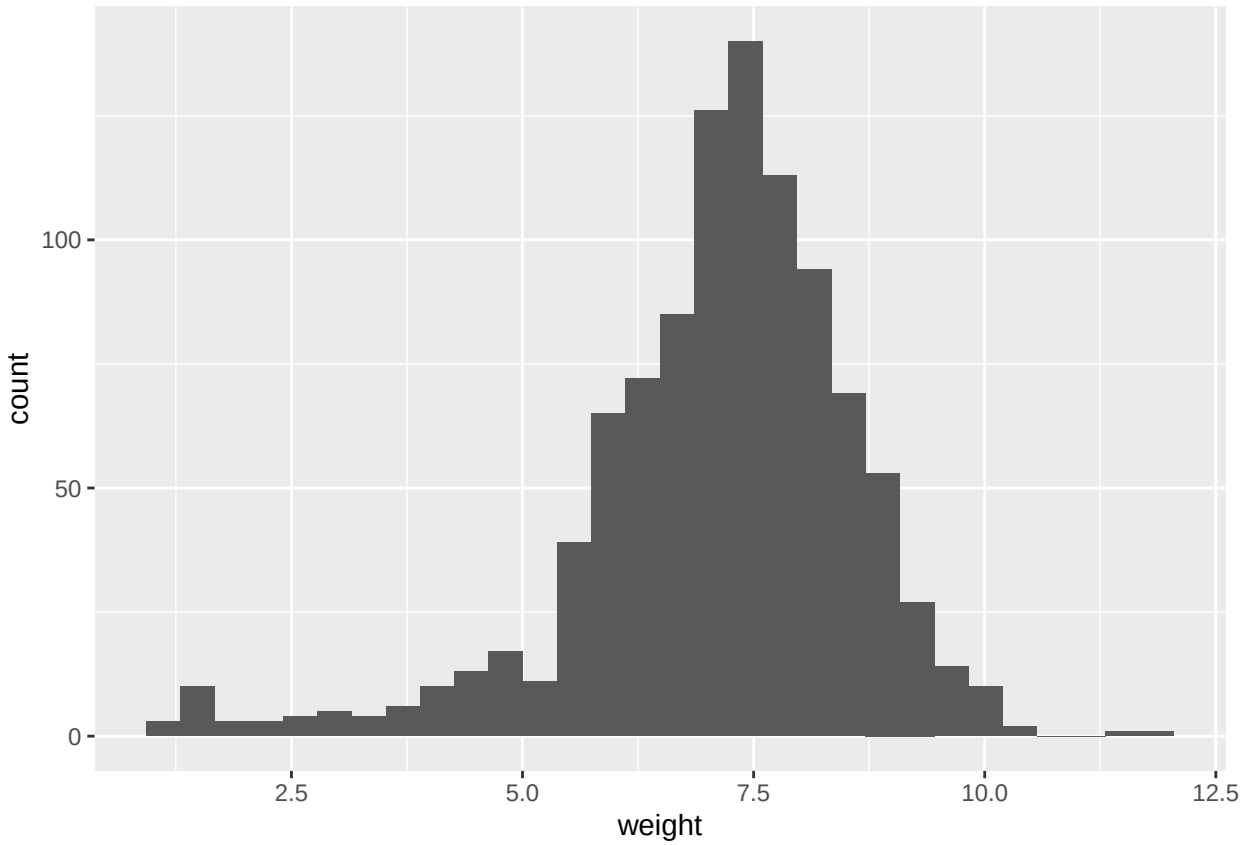
Load the required packages.

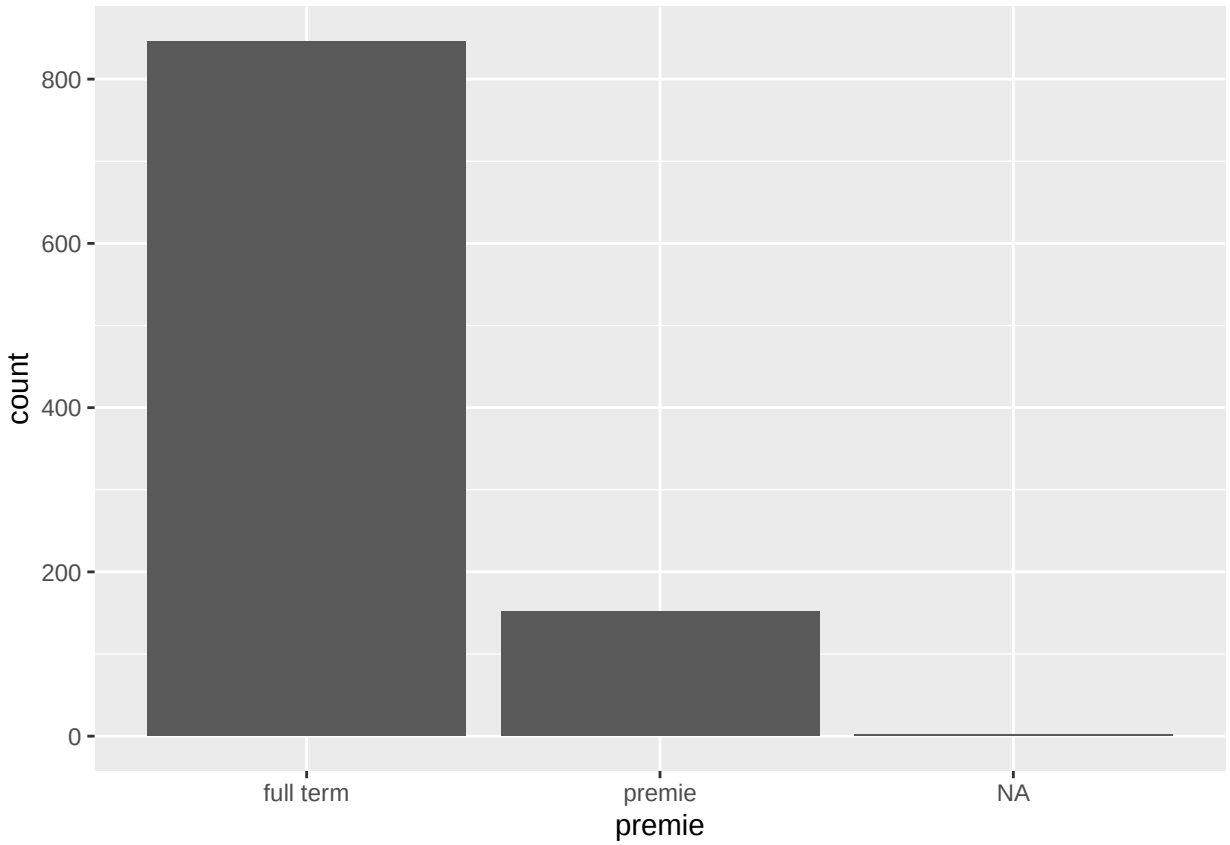
```
library(tidyverse)
library(statsr)
```

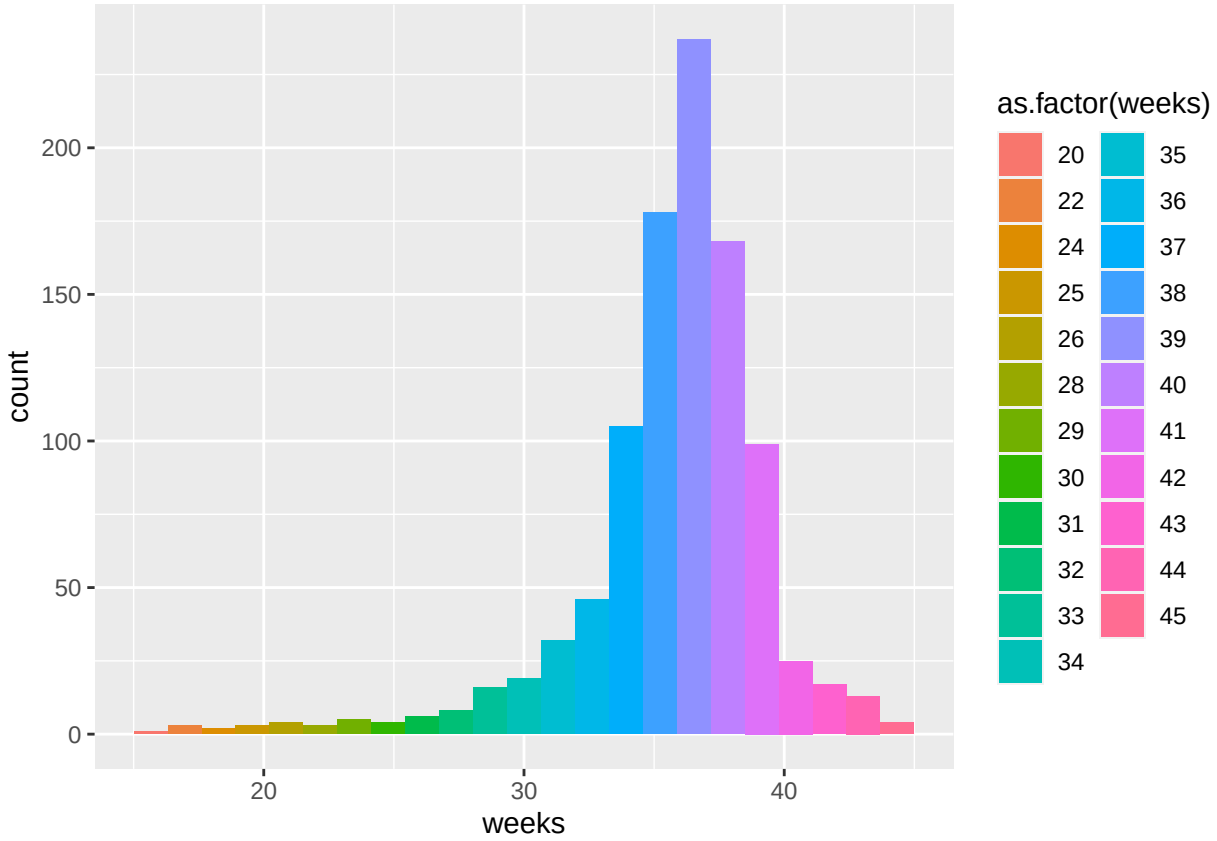
Question 1

Load a data set `nc` which contain information on births recorded in the state of North Carolina. Plot the histograms for the variables `weight`, `premie` and `weeks`

```
Rows: 1,000
Columns: 13
$ fage           <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA, NA, NA,~
$ mage           <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16,~
$ mature         <fct> younger mom, younger mom, younger mom, younger mom, you~
$ weeks          <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42, 40, 38,~
$ premie         <fct> full term, full term, full term, full term, full term, ~
$ visits         <int> 10, 15, 11, 6, 9, 19, 12, 5, 9, 13, 9, 8, 4, 12, 15, 7,~
$ marital        <fct> married, married, married, married, married, married, m~
$ gained         <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34, 12, 30,~
$ weight         <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69, 8.81, 6~
$ lowbirthweight <fct> not low, not low, not low, not low, not low, low, not l~
$ gender         <fct> male, male, female, male, female, male, male, male, mal~
$ habit          <fct> nonsmoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, ~
$ whitemom       <fct> not white, not white, white, white, not white, not whit~
```


count
100
50
0
2.5
5.0
7.5
10.0
12.5
weight

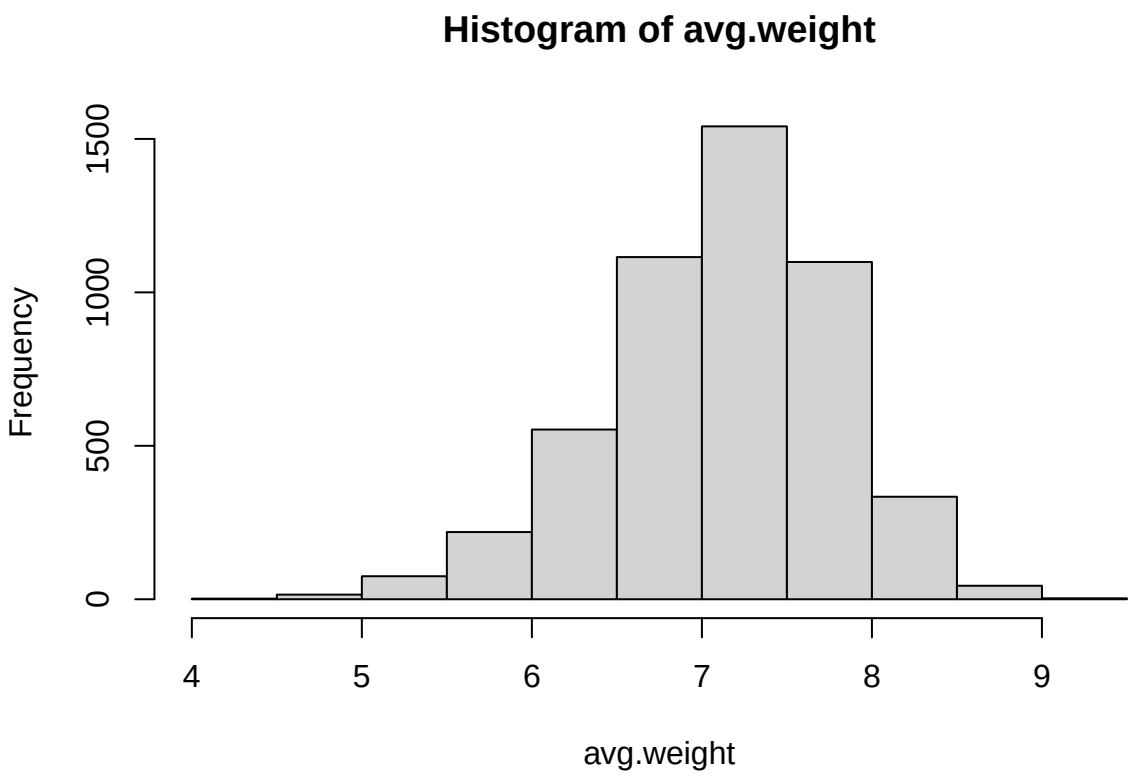

800
600
400
200
0
count
full term
premie
NA
premie

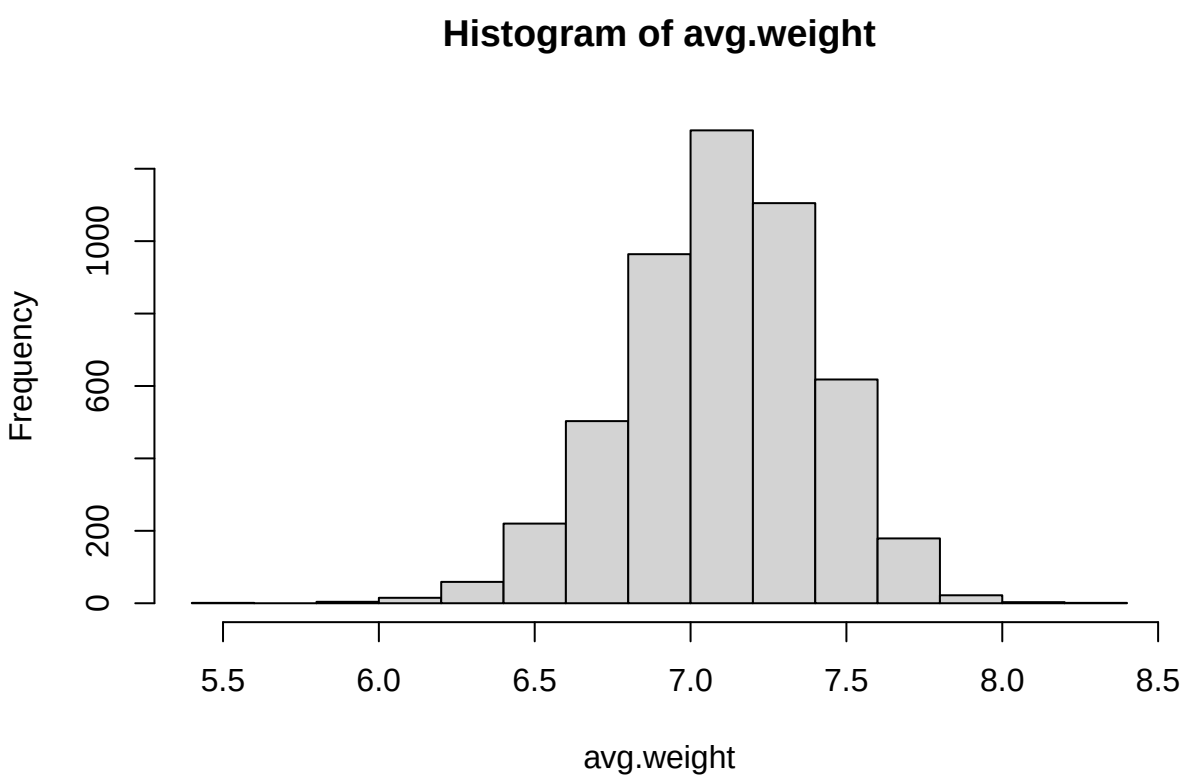


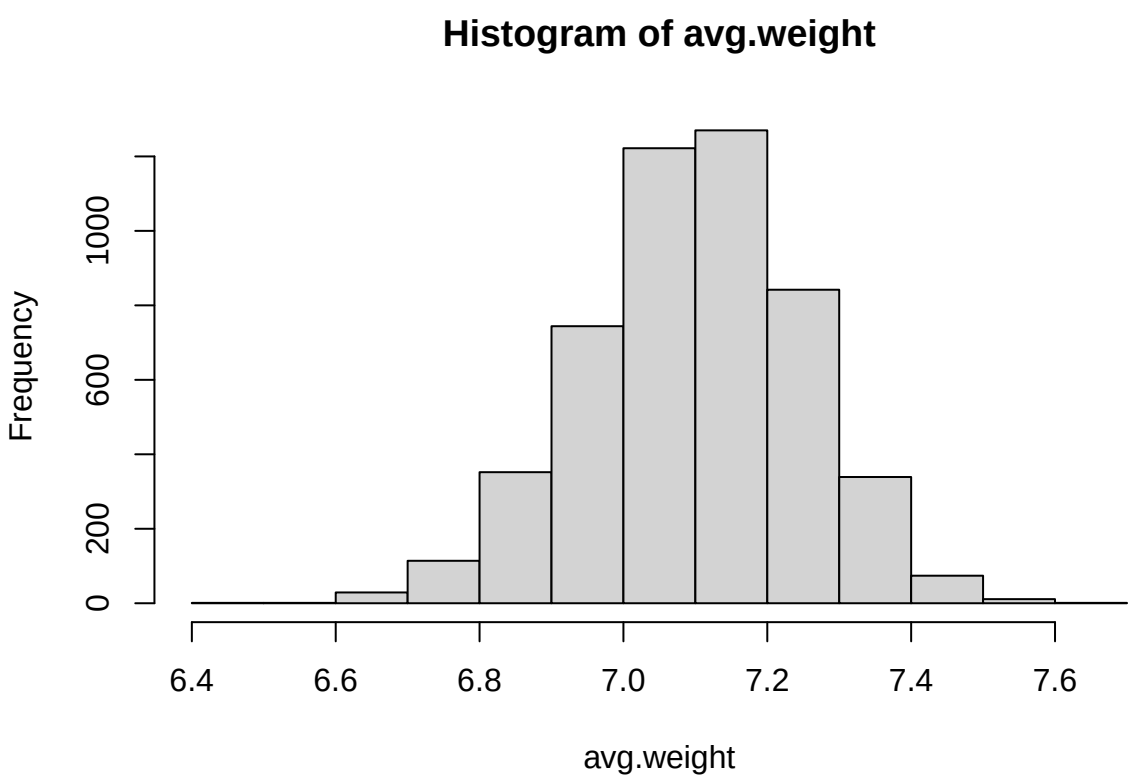
Comment on the distribution of the histograms. Each histogram is not evenly distributed. However, the weeks and the premie histograms are the most skewed. The histogram based on weight is the closest to the center.

Question 2

Let us take into consideration the variable `weight`. Take repeated sample (5000 samples) of size 5,25,100 and find the average `weight` for each sample. Thus write a for-loop to sample $\bar{x}$ for n = 5, 25, 100 where $\bar{x}$ is the average weight.


Histogram of avg.weight
Frequency
0
500
1000
1500
4
5
6
7
8
9
avg.weight

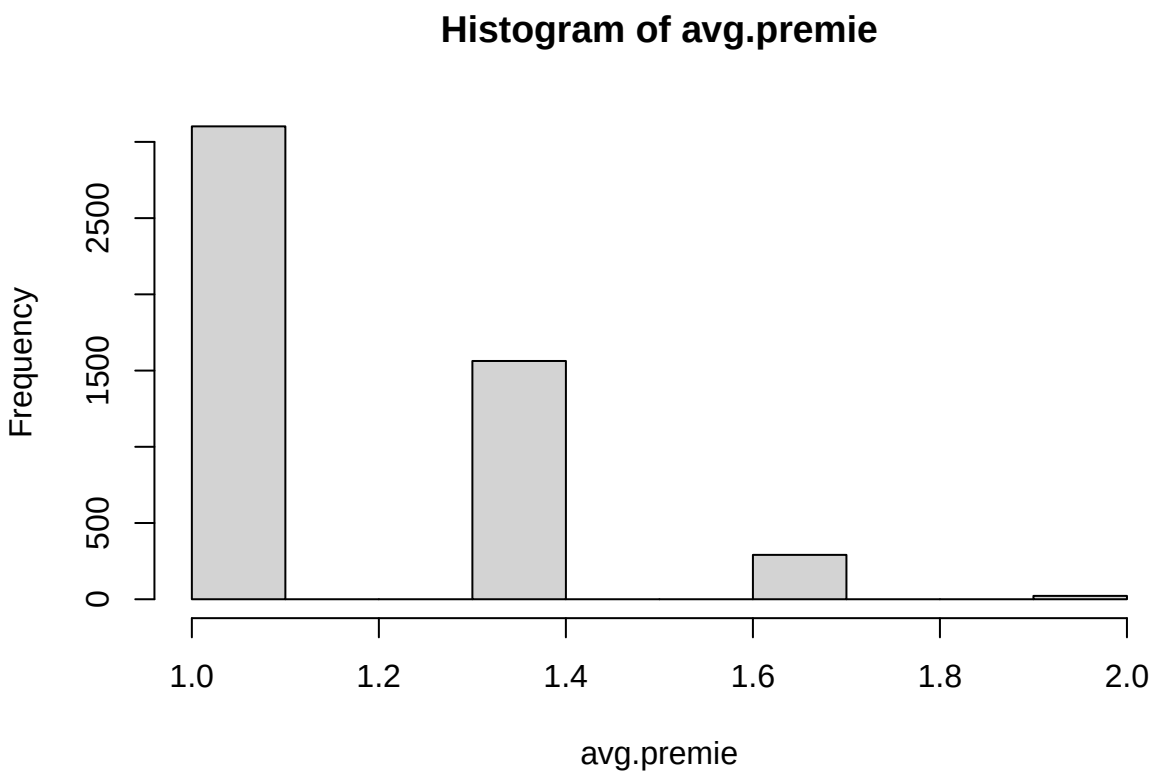

Histogram of avg.weight
Frequency
0
200
600
1000
5.5
6.0
6.5
7.0
7.5
8.0
8.5
avg.weight

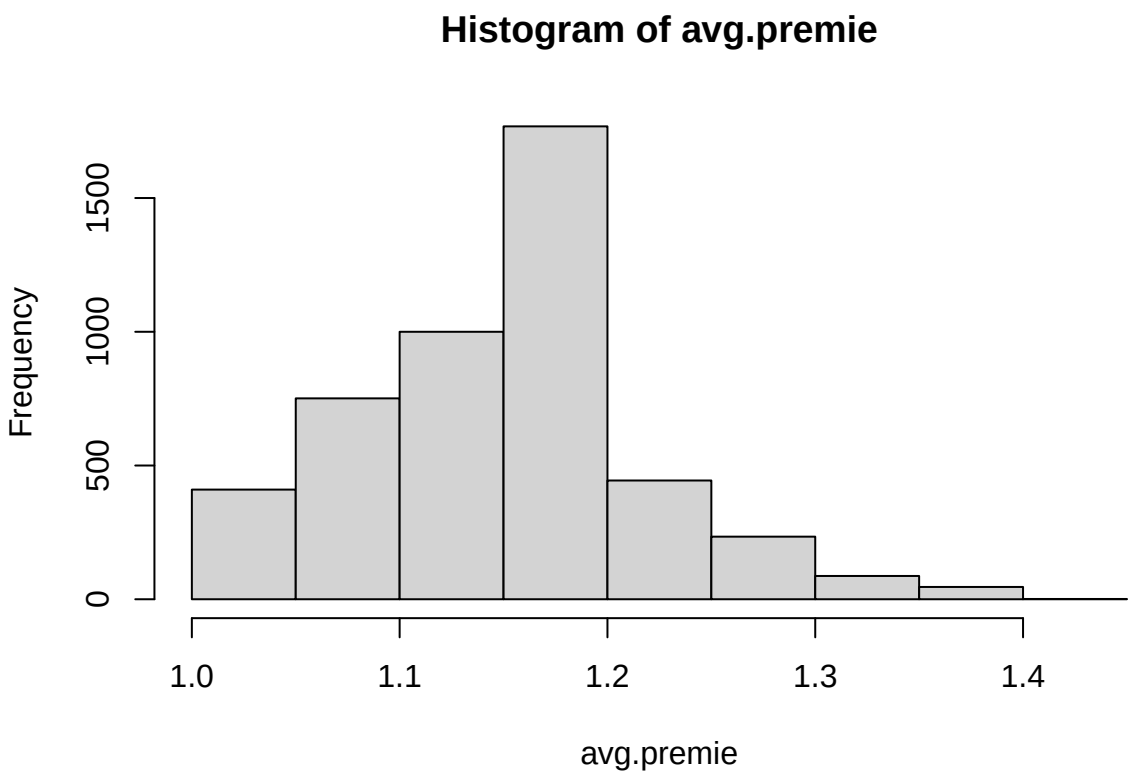


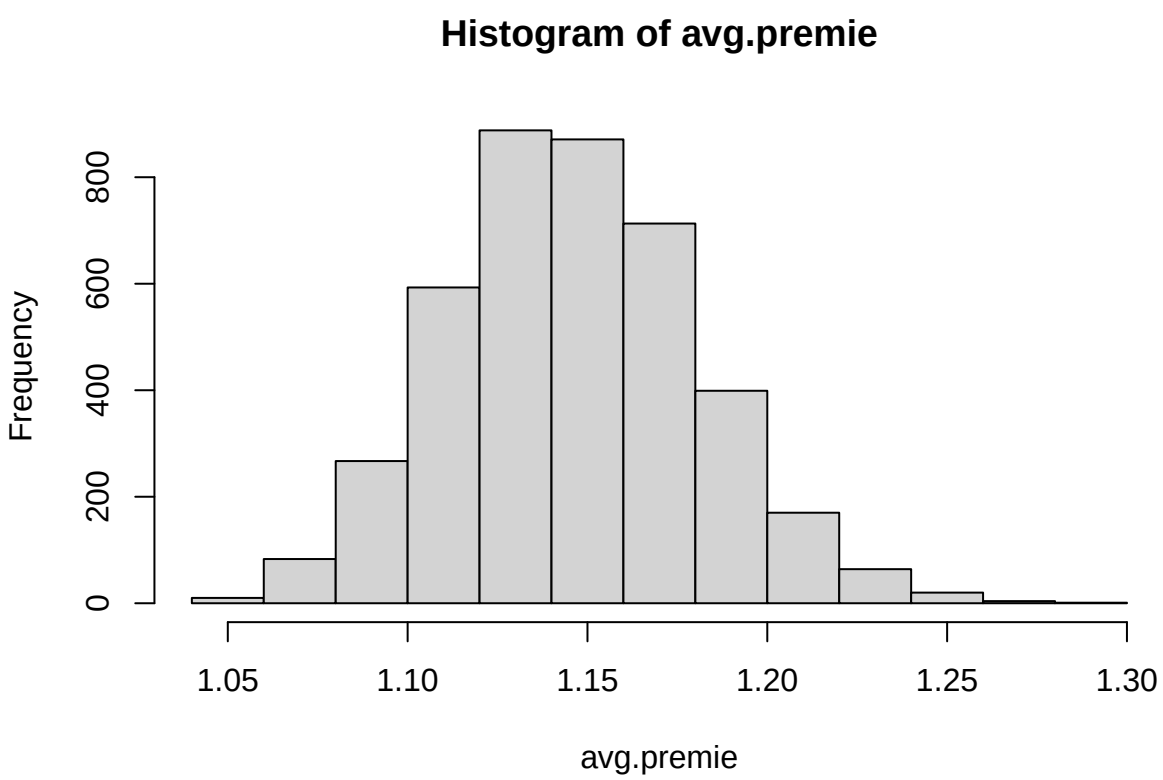
REFERENCES: Rpubs https://rstudio-pubs-static.s3.amazonaws.com Accesssed on November 8, 2021

Question 3

Let us take into consideration the variable `premie`. We will take 5000 repeated samples and find the average `premie`. Thus write a for-loop to sample $\hat{p}$ for n = 3,25,100 where $\hat{p}$ is the proportion of premature births.Write a for-loop to sample $\hat{p}$ for n = 3, 25, 100

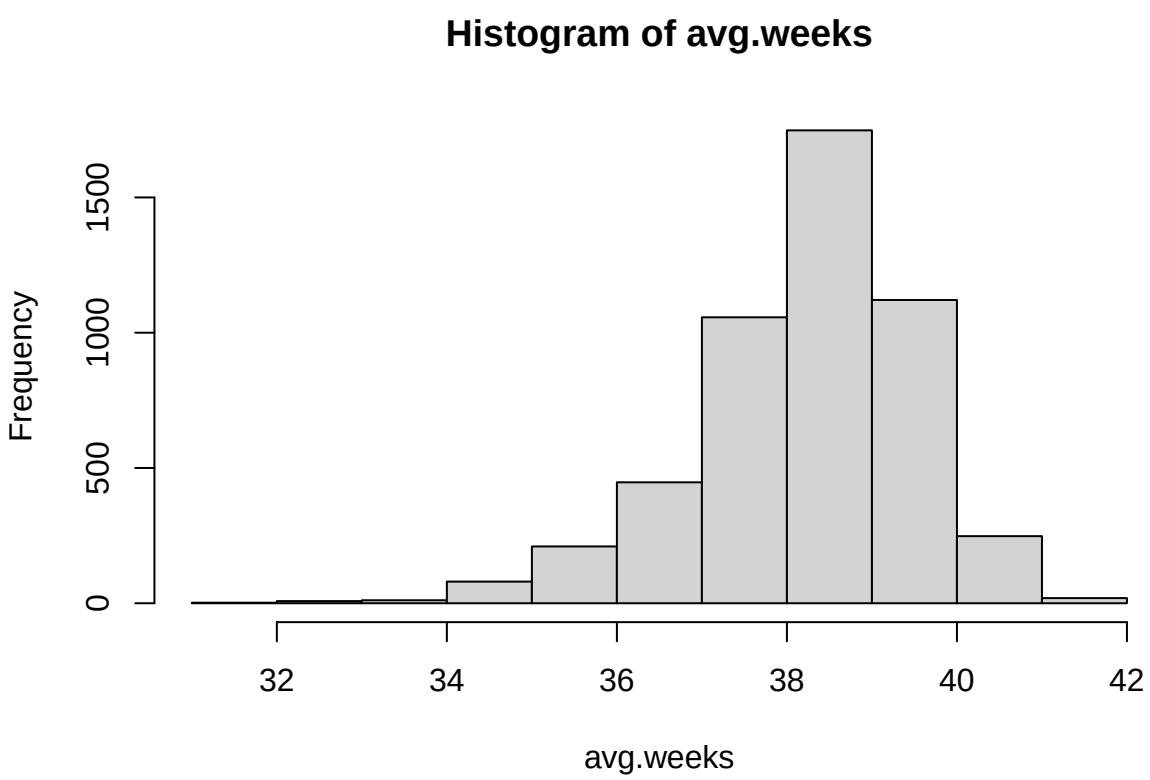

Histogram of avg.premie
Frequency
0
500
1500
2500
1.0
1.2
1.4
1.6
1.8
2.0
avg.premie


Histogram of avg.premie
Frequency
0
500
1000
1500
1.0
1.1
1.2
1.3
1.4
avg.premie



REFERENCES: Rpubs https://rstudio-pubs-static.s3.amazonaws.com Accesssed on November 8, 2021 ### Question 4

Plot the histogram of the variable `weeks` and comment on the shape, skewness of the plot. Take repeated sample (5000 samples) of size 5 and find the average `weight` $\bar{x}_{sk}$ for each sample. Thus write a for-loop to sample $\bar{x}_{sk}$ for n = 5.



REFERENCES: Rpubs https://rstudio-pubs-static.s3.amazonaws.com Accesssed on November 8, 2021 ### Question 5

Plot the histograms for Questions 2-4.

All the histograms are plotted in the code chunks above

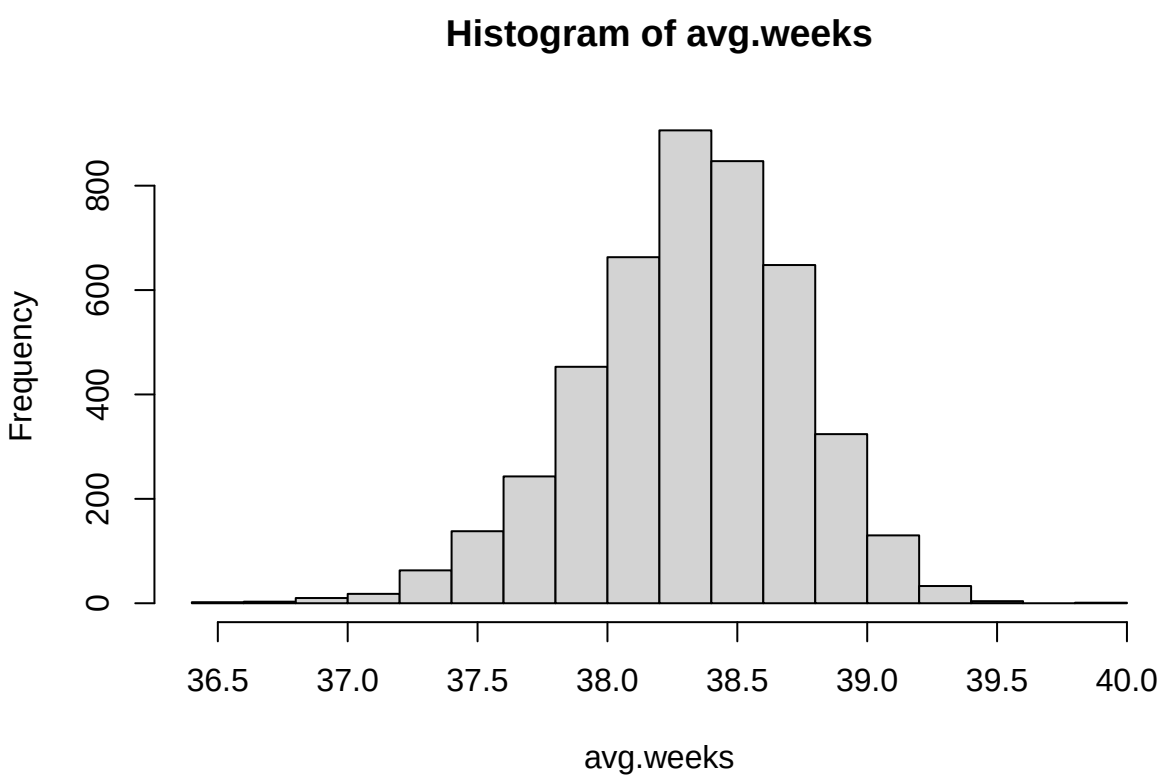
Question 6

Comment on the similarities or disimilarities of the histograms. (1-3 sentences)

As the sample size increases in the histograms of both variables, the mean of the sample distribution becomes closer and close to the center. However, the histograms based on weight shift from right to left and the histograms based on premie shift left to right.

Question 7

`If you alter the for-loop in Question 4 to sample n= 50 rather than n = 5, does the updated histogram b`



Yes, the histogram much better matches 2 and 3.

Question 8

From the results in the previous section what can you conclude about the distribution of the sample mean, $\bar{x}$ and sample proportion, $\hat{p}$?

Each graph was skewed in some way based on the sample size that was given. In the weights which focused on sample mean, the graph was skewed to the right, while the proportion was skewed more to the left. As the sample size increased, however, both the sample mean and proportion shifted towards the center.

Question 9

Comment on the center and spread of the distributions of the sample statistics based on the histograms in Question 5?

The larger the sample size, the less spread will occur and the sampling distribution mean will be more shifterted to the center. **Sample mean:**

Sample proportion:

REFERENCES: Rpubs https://rstudio-pubs-static.s3.amazonaws.com Accesssed on November 8, 2021 ### Conclusion

Thus, we saw that the Central Limit Theorem guarantees normality under one of two assumptions: normality or approx. normality with sufficient sample size. The distribution of the sample statistic is called the sampling distribution.

These concept are the key underlying concepts in inference when we will be testing a hypothesized value of the population mean or the population proportion based on the random sample or when we will be developing confidence interval for the population mean or the population proportion