STAT408_HW1

2022-09-13

Question 5 (10 points)

Dataset births.csv contains the information for 1992 newborns and their parents.

part 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("births.csv")
```

part 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?

```
weight <- NCbirths$weight
```

Since the weights are of the newborns and they are pretty large they are likely not in pounds but instead in a smaller unit such as ounces.

part 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 <- weight*0.0625
```

part d

Print the first 20 babies' weight in pounds.

```
weights_in_pounds[1:20]
```

```
##  [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
```

Part e

What is the mean weight of all babies in pounds?

```
mean(weights_in_pounds)
```

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

part 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.

```
smokers <- dim(NCbirths[NCbirths$Habit == "Smoker",])[1]
total <- dim(NCbirths)[1]
smokers/total
```

```
## [1] 0.0938755
```

9% of mothers in the sample smoke.

part 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?

The percentage found in part (f) is about 5 percentage points less than the CDC's report. Therefore, the percentage of mothers who smoke in our sample is about 5 percentage points less than the percentage of adult Americans who smoke.

Question 6 (10 points)

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

part 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("flint.csv")
```

part 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?

```
high_lead <- flint[flint$Pb >= 15,]
total <- dim(flint)[1]
(dim(high_lead)[1])/total
```

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

The proportion of all locations tested that were then found to have dangerous lead levels is approximately 0.044.

part c

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

```
north <- flint[flint$Region == "North",]
mean(north$Cu)
```

```
## [1] 44.6424
```

part d

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

```
mean(high_lead$Cu)
```

```
## [1] 305.8333
```

part e

Report the mean lead and copper levels for all locations.

```
mean(flint$Pb)
```

```
## [1] 3.383272
```

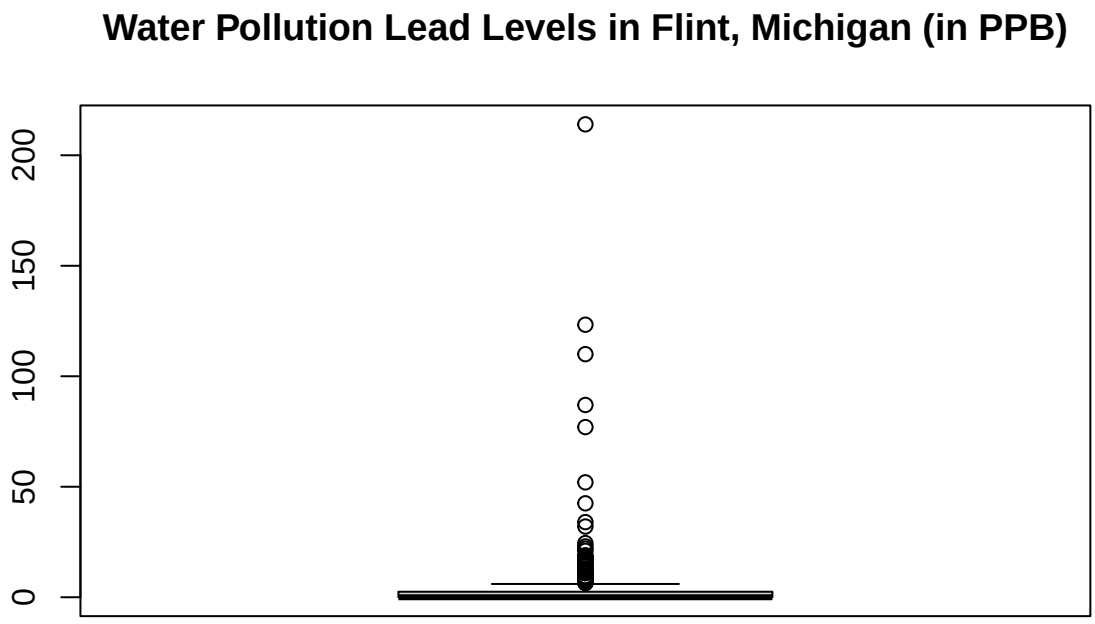
```
mean(flint$Cu)
```

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

part f

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

```
boxplot(flint$Pb, main = 'Water Pollution Lead Levels in Flint, Michigan (in PPB)')
```



part 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.

No, the boxplot appears very skewed to the right with many outliers ranging from around 50 to 200 PPB. Therefore, the mean would be greatly affected by the outliers and would not be a good measure for the location of the center of the data. Instead, the median would be a better measure for the center of this data because it is not as heavily affected by outliers.

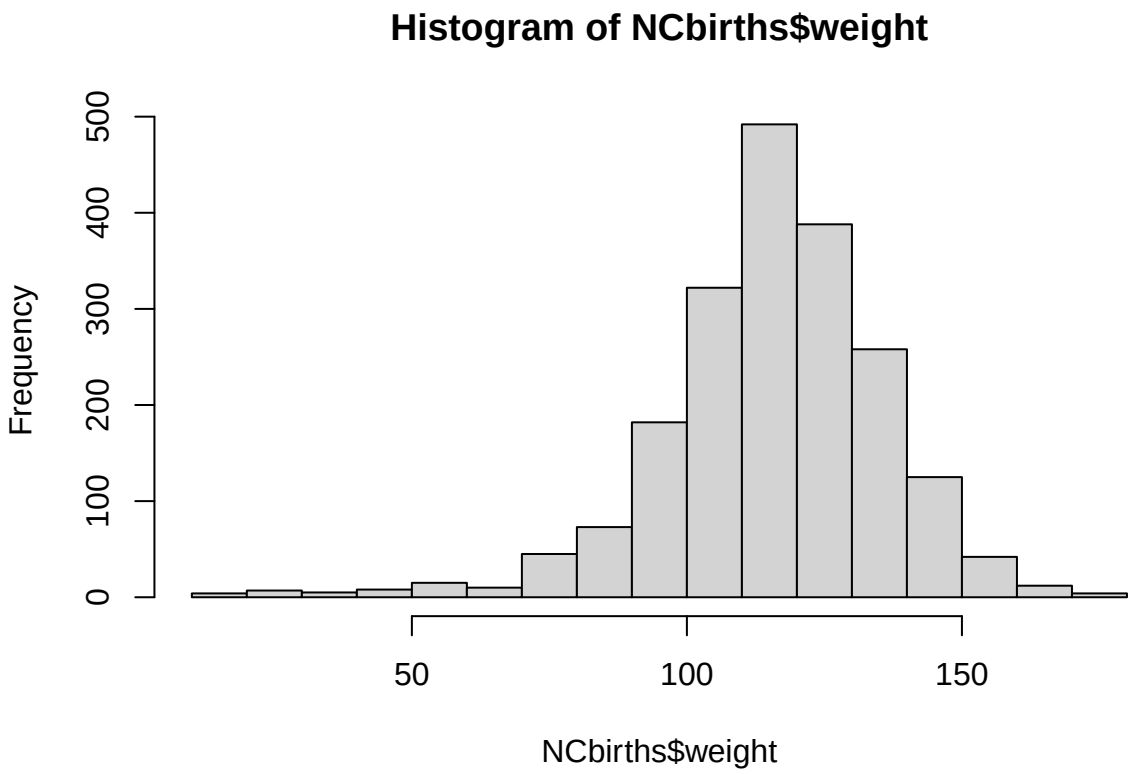
Question 7 (10 points)

We will use a simulation study to show central limit theorem.

part 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?

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



Based on the histogram above, the weight does not appear to follow a normal distribution because instead of showing a symmetrical bell curve the histogram appears skewed to the left.

part b

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

```
weights_sample <- sample(NCbirths$weight, size = 10)
weights_sample
```

```
##  [1] 136 113 155 113 106  95 124 120 115 113
```

```
mean(weights_sample)
```

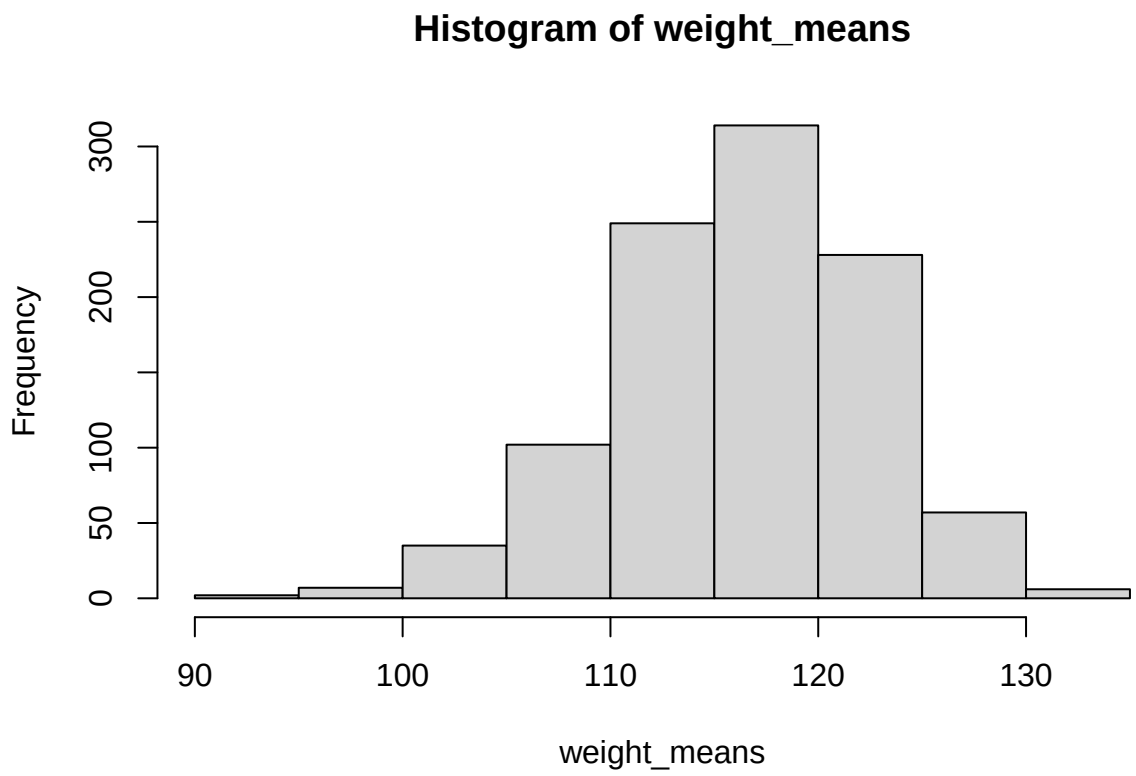
```
## [1] 119
```

part 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?

```r
weight_means <- c()
for (x in 1:1000) {
  weights_sample <- sample(NCbirths$weight, size = 10)
  weight_means <- c(weight_means, mean(weights_sample))

}
hist(weight_means)
```



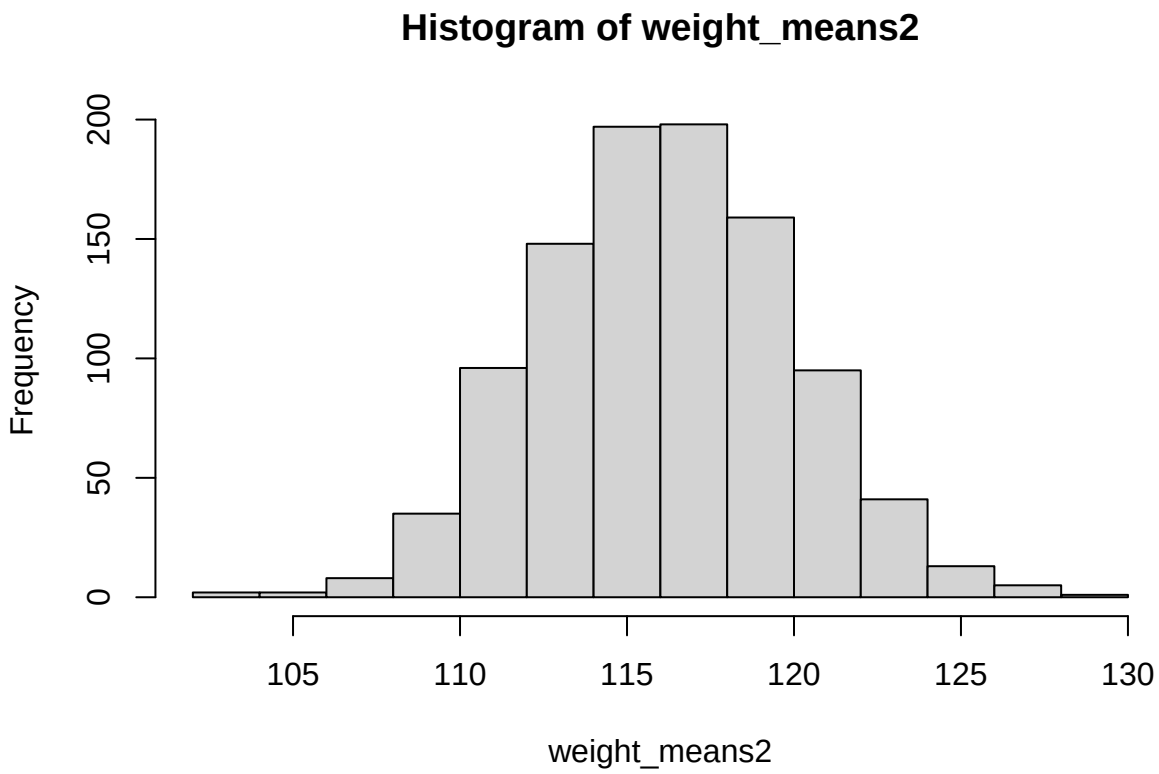
Yes, the histogram for the 1000 means appears approximately normal as the bell curve looks somewhat symmetrical.

part 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.

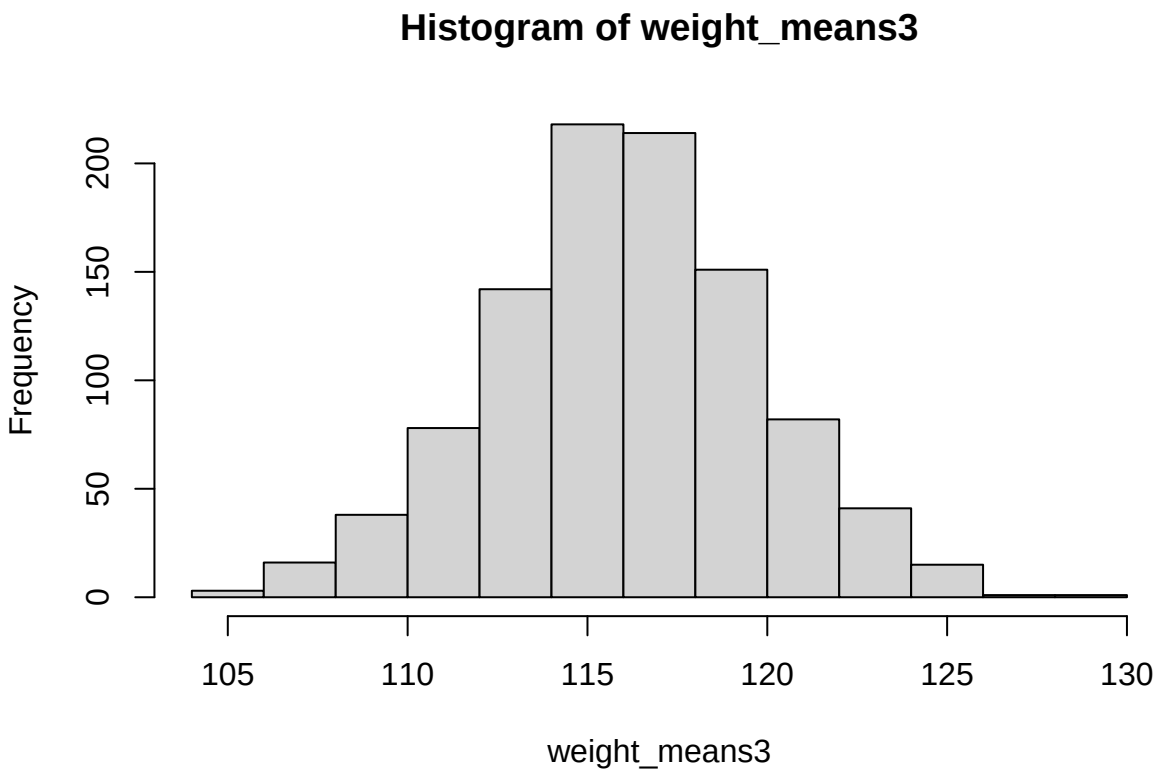
```r
# sample size 30
weight_means2 <- c()
for (x in 1:1000) {
  weights_sample2 <- sample(NCbirths$weight, size = 30)
  weight_means2 <- c(weight_means2, mean(weights_sample2))

}
hist(weight_means2)
```




```
# sample size 100
weight_means3 <- c()
for (x in 1:1000) {
  weights_sample3 <- sample(NCbirths$weight, size = 30)
  weight_means3 <- c(weight_means3, mean(weights_sample3))

}
hist(weight_means3)
```



Yes, the histogram for the 1000 means for both sample sizes 30 and 100 appear approximately normal as the bell curves both look somewhat symmetrical.Since these histograms look so similar and more normal than a sample size of 10, this suggests that increasing the sample size helps to achieve a more approximately normal distribution.