HW3_Videtti

#2. For the remaining exercises in this set, we will use one of R's built-in data sets, called the "ChickWeight" data set. According to the documentation for R, the ChickWeight data set contains information on the weight of chicks in grams up to 21 days after hatching. Use the summary(ChickWeight) command to reveal basic information about the ChickWeight data set. You will find that ChickWeight contains four different variables. Name the four variables. Use the dim(ChickWeight) command to show the dimensions of the ChickWeight data set. The second number in the output, 4, is the number of columns in the data set, in other words the number of variables. What is the first number? Report it and describe briefly what you think it signifies.

```
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

#The four variables in ChickWeight are weight, Time, Chick, and Diet.

```
dim(ChickWeight)
```

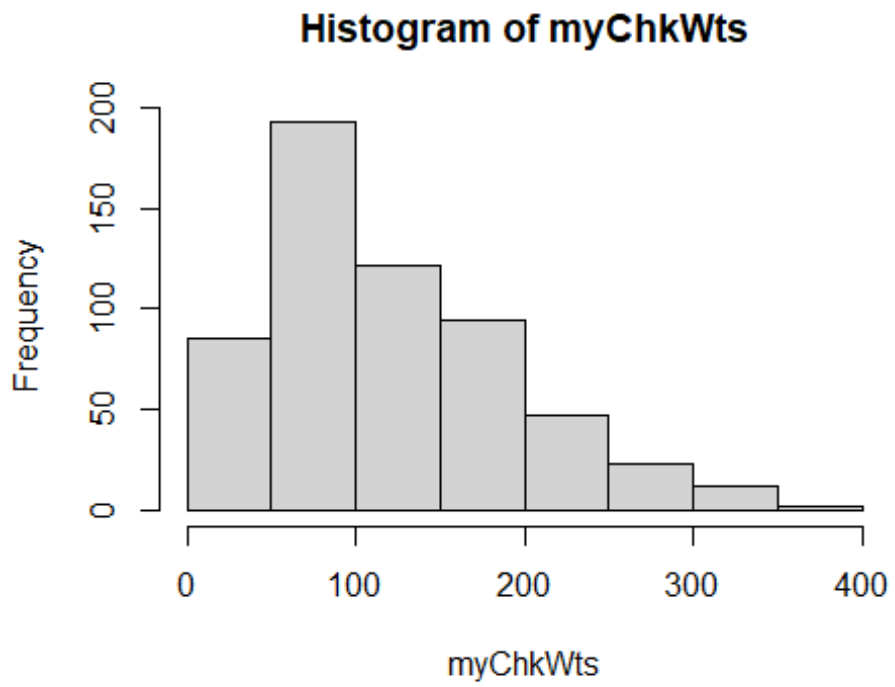
```
## [1] 578   4
```

#The first number is 578 and it indicates that there are 578 rows in the ChickWeight data set.

#3. When a data set contains more than one variable, R offers another subsetting operator, $, to access each variable individually. For the exercises below, we are interested only in the contents of one of the variables in the data set, called weight. We can access the weight variable by itself, using the $, with this expression: ChickWeight$weight. Run the following commands, say what the command does, report the output, and briefly

```
explain each piece of output:

summary(ChickWeight$weight)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    35.0    63.0   103.0   121.8   163.8   373.0

#This gives the summary statistics of the weight variable in the ChickWeight
data set. We see that the output shows the following:
#The minumum of the weight variable is 35
#The 1st quartile (25th percentile) of the weight variable is 63
#The median of the weight variable is 103
#The mean (average) of the weight variable is 121.8
#The 3rd quartile (75th percentile) of the weight variable is 163.8
#The maximum of the weight variable is 373

head(ChickWeight$weight)

## [1] 42 51 59 64 76 93

#This gives the first few rows of the weight variable in the ChickWeight data
set

mean(ChickWeight$weight)

## [1] 121.8183

#This gives the mean or average of the weight variable in the ChickWeight
data set. We see this is 121.8, which matches the mean we found using the
summary command.

myChkWts <- ChickWeight$weight
#This assigns the weight variable in the ChickWeight data set to a new
variable called myCkkWts

quantile(myChkWts,0.50)

## 50%
## 103

#This gives the 0.5 quantile of the weight variable in the ChickWeight data
set. This is equivalent to the 2nd quartile, the 50th percentile, and the
median. We see that this is 103, which matches the median we found using the
summary command.
```

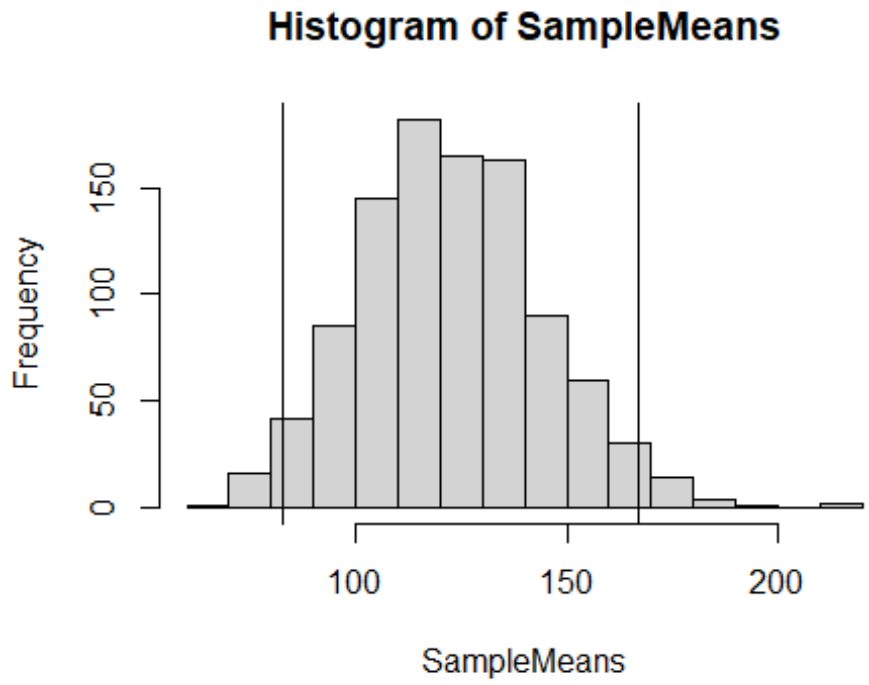
```
#4. In the second to last command of the previous exercise, you created a
copy of the weight data from the ChickWeight data set and put it in a new
vector called myChkWts. You can continue to use this myChkWts variable for
the rest of the exercises below. Create a histogram for that variable. Then
write code that will display the 2.5% and 97.5% quantiles of the distribution
for that variable. Write an interpretation of the variable, including
descriptions of the mean, median, shape of the distribution, and the 2.5% and
97.5% quantiles. Make sure to clearly describe what the 2.5% and 97.5%
quantiles signify.

hist(myChkWts)
```




```
#This variable is right skewed, meaning the mean is greater than the median.
We saw in Exercise 3 that the median is 103 and the mean is 121.8.

quantile(myChkWts,0.025)

## 2.5%
##   41

quantile(myChkWts,0.975)

##   97.5%
## 294.575
```

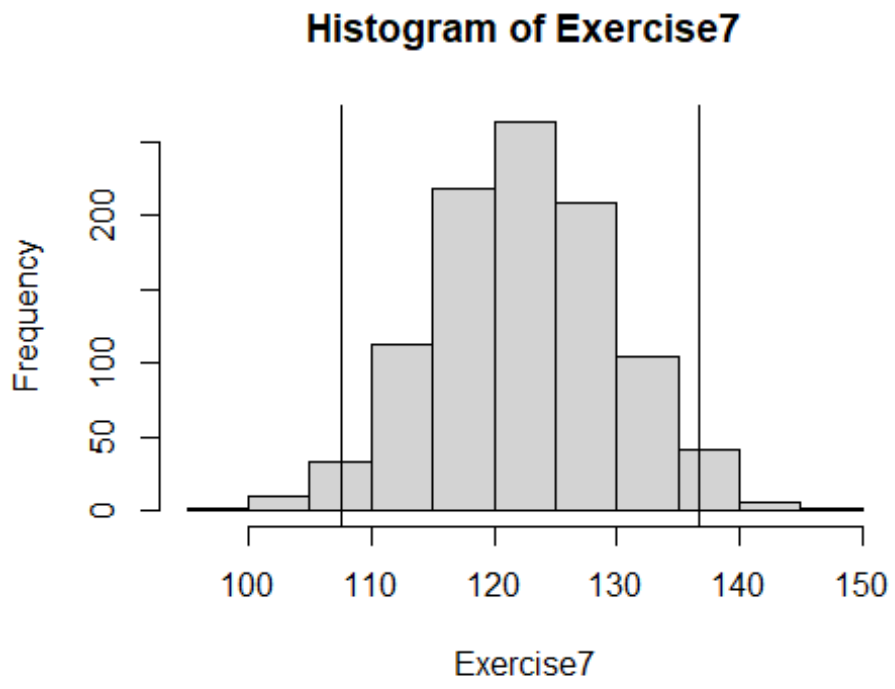
```
#The 2.5% quantile is 41 and the 97.5% quantile is 294.575. This means that
95% (97.5 - 2.5 = 95) of the values in the myChkWts variable are between 41
and 294.575.
```

```
#5. Write R code that will construct a sampling distribution of means from
the weight data (as noted above, if you did exercise 3 you can use myChkWts
instead of ChickWeight$weight to save yourself some typing). Make sure that
the sampling distribution contains at least 1,000 means. Store the sampling
distribution in a new variable that you can keep using. Use a sample size of
n = 11 (sampling with replacement). Show a histogram of this distribution of
sample means. Then, write and run R commands that will display the 2.5% and
97.5% quantiles of the sampling distribution on the histogram with a vertical
line.

SampleMeans <- replicate(1000,mean(sample(myChkWts,11,replace = TRUE)))
hist(SampleMeans)
abline(v=quantile(SampleMeans,0.025))
abline(v=quantile(SampleMeans,0.975))
```




```
#6. If you did Exercise 4, you calculated some quantiles for a distribution
of raw data. If you did Exercise 5, you calculated some quantiles for a
sampling distribution of means. Briefly describe, from a conceptual
perspective and in your own words, what the difference is between a
distribution of raw data and a distribution of sampling means. Finally,
comment on why the 2.5% and 97.5% quantiles are so different between the raw
data distribution and the sampling distribution of means.

#The distribution of raw data shows how the raw data are distributed and
includes all values that the variable could possibly be. The distribution of
sampling means shows the distribution of means of random samples of the
variable and may not (and is actually highly unlikely to) include all values
that the variable could possibly be. Since you take the mean of each sample,
this removes extreme values and the mean of the sample means tends to
converge toward the population mean after a larger and larger number of
samples are taken. The spread of the distribution of sample means will be
less than that of the distribution of the raw data, and thus the difference
between the 2.5% and 97.5% quantiles will be smaller among sample means than
the raw data (2.5% quantile will be larger, 97.5% quantile will be smaller).




#7. Redo Exercise 5, but this time use a sample size of n = 100 (instead of
```

```
the original sample size of n = 11 used in Exercise 5). Explain why the 2.5%
and 97.5% quantiles are different from the results you got for Exercise 5. As
a hint, think about what makes a sample "better."

Exercise7 <- replicate(1000,mean(sample(myChkWts,100,replace = TRUE)))
hist(Exercise7)
abline(v=quantile(Exercise7,0.025))
abline(v=quantile(Exercise7,0.975))
```




```
#We see the 2.5% quantile continues to get larger, the 97.5% quantile
continues to get smaller, and thus the difference between the two continues
to get smaller. This is due to a larger sample size being taken and each
sample mean thus being more reflective of the actual mean of the variable.
There are even less extreme values in this sample mean distribution than
there were in Exercise 5, also due to each sample being larger and more
reflective of the variable itself.
```