# Attribution Statement:

1. Homework 3 by Parin Patel: I did this homework by myself, with help from the book and the professor. In addition I used the following websites to help with the r code for the histogram questions: <https://www.stat.berkeley.edu/~stark/Java/Html/SampleDist.htm>

# Exercises:

**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.**

A picture containing wall

Description automatically generated

The four variables for the ChickWeight dataset are the following: weight, Time, Chick, and Diet.

A close up of a logo

Description automatically generated

The output of the dim(Chickweight) commands contains two numbers. The first reflects the number of rows, and the second number reflects the number of columns in the dataset. It is important to remember that the number of rows in a dataset signifies the number of observations, while the columns signify variables. Therefore, it can be said that the ChickWeight dataset contains 578 observations, or rows.

**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 ,withthisexpression:ChickWeightweight. Run the following commands, say what the command does, report the output, and briefly explain each piece of output:**

**summary(ChickWeight$weight)**

**head(ChickWeight$weight)**

**mean(ChickWeight$weight)**

**myChkWts <- ChickWeight$weight**

**quantile(myChkWts,0.50)**

A screenshot of a cell phone

Description automatically generated

The first, summary function of the variable weight returns an output about the quantile statistics of the variable. Since the weight variable is a numeric variable, it returns numeric statistics of all its data. The first number is the minimum, the second number is the 1st quartile, the third is the median, the fourth is the mean, the fifth is the 3rd quartile, and finally the maximum.

The head function returns the weight values of the first 6 observations.

The mean function returns the average weight of all the chicks. The formula tables the sum of all the weights of all the observations and divides it by the number of observations. In this case, since the total number of rows is 578, the sum of all the weights is divided by 576.

Finally, the variable “myChkWts” returns a vector of all the observations. This variable was created by subsetting the weight variable using a logical operator (<-) .

The created variable “myChkWts” is then used in a quantile function. This function divides the data into an arbitrary number of divisions. In this specific case, the 50% quantile returns a weight of 103. This is also the median of the variable and can be seen under “median” in the summary function.

**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**.

A screenshot of a cell phone

Description automatically generated

The histogram above shows a distribution that is skewed to the right, and therefore has a tail that falls on the right side, and larger values on the right side than the left. This distribution shows that a bulk fo the observations are less than the mean, while a few are larger. The mean of this histogram is 121.8 and the median is 103. Therefore, both are to the right of the peak of the histogram.

The first quantile lines displayed for 2.5% is 41. This shows how just 2.5% of all the sample means are at or below the value of 41. The second quantile line of 97.5% is at 294.574. This shows that just 97.5% of all the sample means are at or below 294.575. Likewise, this also means that just 2.5% of all sample means are above or at the value of 294.575. Finally, the first line also shows that 97.5% of all the sample means are above or at 41. G

**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**.

A screenshot of a cell phone

Description automatically generated

Code:

A screenshot of a cell phone

Description automatically generated

**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 sampling distribution of the sample means allows for a sample to be taken from the population data n times, with n being a fixed size randomly taken from the population repeatedly. For every sample, the mean is also computed and recorded to create a distribution of the sample mean. Therefore, in summary, we will have n means of n samples. This is different than the distribution of raw data that is the distribution of observations. This this case, it is the distribution of observations for the variable weight. This distribution mean falls within the the range of the original dataset; however, it will not be the mean from the original sample. The sample mean’s distribution however, generates a mean that is the same as the original sample mean. The reason why 2.5% and 97.5% quantiles vary between the two raw data distribution and the sample distribution of means is because the distribution in the sample mean is normalized through the specified 1000 trials. This causes decreases the chance of an outlier, seen in the raw data distribution. It is important to note that the mean and the sample mean do not have to loo kalike. The sample mean is, however, a better estimate of the mean of the population. And therefore, the larger the mean the better the estimate.

**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.”**

A screenshot of a cell phone

Description automatically generated

The histogram above has the sample size adjusted from 11 to 100. The results of the 2.5% and the 97.5% quantiles are different than the results for exercise 5 because we are utilizing a larger number of observations to create the sample mean. Therefore, we are more likely to have a better representation of the observations when n =100 than when n=11. As stated in number 6, the sample mean is a better estimate of the population. This estimate increases in accuracy as the mean increases.

# Appendix A: Final Script;

datasets::ChickWeight

#2. Use summary(chickweight) to reveal basic info about dataset

summary(ChickWeight)

#Use the dim(Chickweight) cmmd to show the dimensions of the ChickWeight dataset.

dim(ChickWeight)

#3.Run following commands to access the weight variable by itself.

summary(ChickWeight$weight)

head(ChickWeight$weight)

mean(ChickWeight$weight)

myChkWts <- ChickWeight$weight

quantile(myChkWts,0.50)

##4. Histogram and quantiles

myChkWts\_Hist<-hist(myChkWts, main="4. Histogram of myChkWts" )

myChkWts\_Hist<-abline(v=quantile(myChkWts,prob=0.025))

myChkWts\_Hist<-abline(v=quantile(myChkWts,prob=0.975))

myChkWts\_Hist

quantile(myChkWts,0.025)

quantile(myChkWts,0.975)

###5 sampling distribution of means from weight data

#at least 1,000 means.

#store sampling distribution as a new variable.

#use n=11 as sample size with replacement.

samplingDistribution <- replicate(1000,mean(sample(myChkWts,size=11,replace=TRUE)))

myChickWts\_SamplHist<-hist(samplingDistribution, main="5. Histogram of Sampling Distribution of myChkWts") #histogram

#add quantiles abline at 2.5% and 97.5%

myChickWts\_SamplHist<-abline(v=quantile(samplingDistribution,prob=0.025))

myChickWts\_SamplHist<-abline(v=quantile(samplingDistribution,prob=0.975))

myChickWts\_SamplHist #call histogram

#check of quantiles

quantile(samplingDistribution,0.025) #84.725

quantile(samplingDistribution,0.975) #166.2841

###7. Redo #5 but use n=100

samplingDistribution100 <- replicate(1000,mean(sample(myChkWts,size=100,replace=TRUE)))

myChickWts\_SamplHist100<-hist(samplingDistribution100, main="7. Histogram of Sampling Distribution of myChkWts, n=100") #histogram

#add quantiles abline at 2.5% and 97.5%

myChickWts\_SamplHist100<-abline(v=quantile(samplingDistribution100,prob=0.025))

myChickWts\_SamplHist100<-abline(v=quantile(samplingDistribution100,prob=0.975))

myChickWts\_SamplHist100 #call histogram

# Appendix B: Final Output;

|  |
| --- |
| >  > #2. Use summary(chickweight) to reveal basic info about dataset  > 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  >  >  >  > #Use the dim(Chickweight) cmmd to show the dimensions of the ChickWeight dataset.  > dim(ChickWeight)  [1] 578 4  >  >  >  >  > #3.Run following commands to access the weight variable by itself.  > summary(ChickWeight$weight)  Min. 1st Qu. Median Mean 3rd Qu. Max.  35.0 63.0 103.0 121.8 163.8 373.0  >  > head(ChickWeight$weight)  [1] 42 51 59 64 76 93  >  > mean(ChickWeight$weight)  [1] 121.8183  >  > myChkWts <- ChickWeight$weight  >  > quantile(myChkWts,0.50)  50%  103  >  >  > ##4. Histogram and quantiles  > myChkWts\_Hist<-hist(myChkWts, main="4. Histogram of myChkWts" )  > myChkWts\_Hist<-abline(v=quantile(myChkWts,prob=0.025))  > myChkWts\_Hist<-abline(v=quantile(myChkWts,prob=0.975))  >  > myChkWts\_Hist  NULL  >  > quantile(myChkWts,0.025)  2.5%  41  > quantile(myChkWts,0.975)  97.5%  294.575  >  >  > ###5 sampling distribution of means from weight data  > #at least 1,000 means.  > #store sampling distribution as a new variable.  > #use n=11 as sample size with replacement.  >  >  > samplingDistribution <- replicate(1000,mean(sample(myChkWts,size=11,replace=TRUE)))  > myChickWts\_SamplHist<-hist(samplingDistribution, main="5. Histogram of Sampling Distribution of myChkWts") #histogram  > #add quantiles abline at 2.5% and 97.5%  > myChickWts\_SamplHist<-abline(v=quantile(samplingDistribution,prob=0.025))  > myChickWts\_SamplHist<-abline(v=quantile(samplingDistribution,prob=0.975))  > myChickWts\_SamplHist #call histogram  NULL  >  > #check of quantiles  > quantile(samplingDistribution,0.025) #84.725  2.5%  80.27045  > quantile(samplingDistribution,0.975) #166.2841  97.5%  165.6568  >  >  >  >  > ###7. Redo #5 but use n=100  > samplingDistribution100 <- replicate(1000,mean(sample(myChkWts,size=100,replace=TRUE)))  > myChickWts\_SamplHist100<-hist(samplingDistribution100, main="7. Histogram of Sampling Distribution of myChkWts, n=100") #histogram  > #add quantiles abline at 2.5% and 97.5%  > myChickWts\_SamplHist100<-abline(v=quantile(samplingDistribution100,prob=0.025))  > myChickWts\_SamplHist100<-abline(v=quantile(samplingDistribution100,prob=0.975))  > myChickWts\_SamplHist100 #call histogram  NULL |
|  |
| |  | | --- | | > | |