IST 777

Joe Hernandez

7/25/2019

Homework #3

2. Use data set ChickWeight

   Reveal basic information about the 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
```

   Name four variables in dataset

```
> colnames(ChickWeight)
[1] "weight" "Time"   "Chick"  "Diet"
```

   The four variables in the dataset are weight, Time, Chick, and Diet.

   Show dimensions of dataset

```
> dim(ChickWeight)
[1] 578   4
```

   The first number, 578, is the number of records in the dataset.

3. Run commands, describe function, report output, and 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
```

   Summary function reveals basic information about ChickWeight dataset specifically the weight column

   Output reveals range and quartile data within the ChickWeight data for weight

   Minimum weight for chick is 35.0 grams

   First quartile, represents the 25th percentile of the chicks' weight, is 63.0 grams

   Median, represents the 50th percentile or midpoint of the chicks' weight, is 103.0 grams

   Mean, expresses the average chick weight, is 121.8

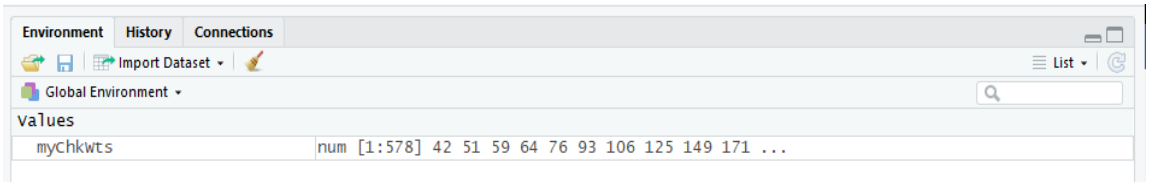
   Third quartile, represents the 75th percentile of the chicks' weight, is 163.8 grams

   Maximum weight for chick is 373.0 grams

```
> head(ChickWeight$weight)
[1] 42 51 59 64 76 93
```
Head function shows the first six entries of the ChickWeight dataset for the weight column
Output shows six entries that make up the chicks' weight
Each piece of output shows the first six chick weights are 42 grams, 51 grams, 59 grams, 64 grams, 76 grams, and 93 grams respectively

```
> mean(ChickWeight$weight)
[1] 121.8183
```
Mean function takes the average of the ChickWeight dataset for the weight column
Output shows a more precise average up to four decimal places
Result is a more precise average of chicks' weight at 121.8183 compared to the summary function 121.8

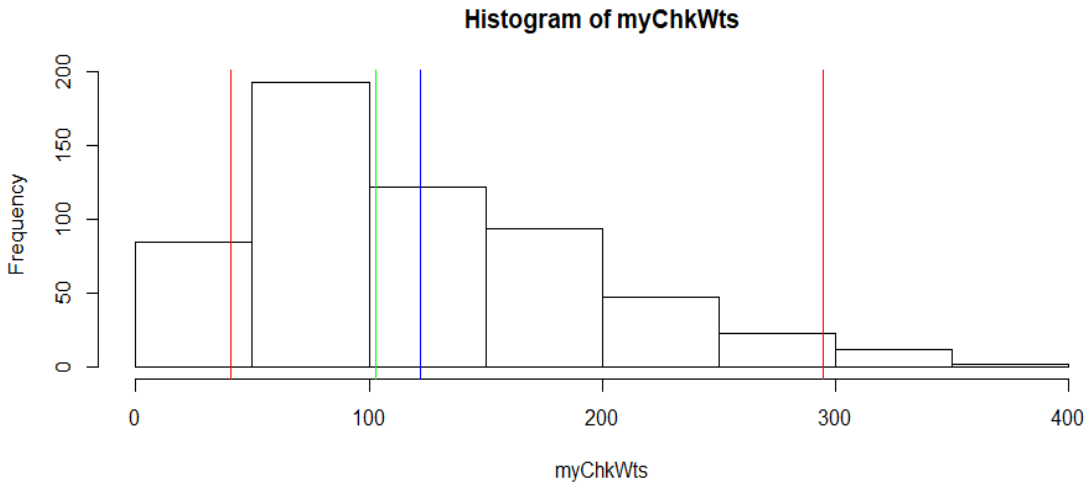
```
myChkWts <- ChickWeight$weight
```
This function creates a variable called myChkWts where it will store only the weights in the ChickWeight dataset
No output will be shown in the console
In the environment of RStudio myChkWts will populate as a variable with a description of the type, length, and a set of values




```
> quantile(myChkWts, 0.50)
50%
103
```
This function calls for the quantile of the variable myChkWts which is set at the 50th percentile
Output will show the 50th percentile quantile, also known as the median
Result from running this function shows the 50% quantile setting followed by the value 103 grams for the median chick weight

4. Create a histogram for the myChkWts variable

```
> hist(myChkWts)
> abline(v = quantile(myChkWts, 0.975), col = "red")
> abline(v = quantile(myChkWts, 0.025), col = "red")
> abline(v = quantile(myChkWts, 0.5), col = "green")
> abline(v = mean(myChkWts), col = "blue")
```




Display 2.5% and 97.5% quantiles of distribution

```
> quantile(myChkWts, 0.025)
2.5%
  41
```

> quantile(myChkWts, 0.975)

```
  97.5%
294.575
```

Display median of distribution

> quantile(myChkWts, 0.5)

```
50%
103
```

Display mean of distribution

```
> mean(myChkWts)
[1] 121.8183
```

Summarize overall data for quick reference

```
> summary(myChkWts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0    63.0   103.0   121.8   163.8   373.0
```

The variable collects all of the ChickWeight entries specifically for weight where the shape of the histogram shows a Poisson distribution. This is a valid observation as it will be rare for chicks to weigh higher than the mean or median values.

The mean of the variable is 121.8183 grams

The median of the variable is 103 grams

The area between the 2.5% and 97.5% quantiles shows where 95% of the myChkWt data will fall which is up to 2 standard deviations from the mean. Chick weights between 41 - 294.575 grams range holds 95% of the data.

In the histogram, the 2.5% and 97.5% quantiles are shown in red, median weight is green, and mean is blue.

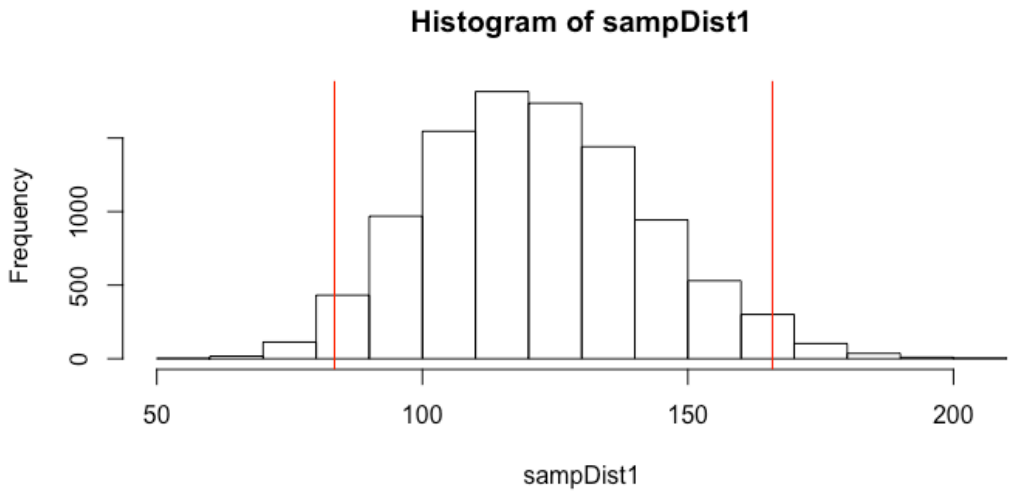
5. Construct sampling distribution of means for myChkWts dataset

Set seed for replication

```
> set.seed(1234)
> sampDist1 <- replicate(10000, mean(sample(myChkWts, size = 11,
replace = TRUE)),
                       simplify = TRUE)
```

Show histogram of distribution of sample means

```
> hist(sampDist1)
```




Display 2.5% and 97.5% quantiles of sampling distribution on histogram

```
> abline(v = quantile(sampDist1, 0.025), col = "red")
> quantile(sampDist1, 0.025)
    2.5%
83.27273
> abline(v = quantile(sampDist1, 0.975), col = "red")
> quantile(sampDist1, 0.975)
   97.5%
166.7295
```

6. Describe difference between distribution of raw data and sampling means

   The raw data collected is from actual chick weights recorded whereas the sampling means takes a sample set of chick weights from the raw data and takes the average for each set.

   Run summary of both raw data and sampling means

```
> summary(myChkWts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0    63.0   103.0   121.8   163.8   373.0
> summary(sampDist1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  57.36  106.73  120.64  121.73  135.55  209.18
```

   Although the mean for the data is fairly close the inter quantile ranges between 2.5% and 97.5%

for the raw data and sampling means shows a difference due to the sampling size which randomly selects chick weights. The average will skew the results closer to the mean than the raw data resulting in higher weight result of 83.45227 grams for 2.5% quantile range and 165.9114 grams for 97.5% quantile range.

```
> quantile(myChkWts, 0.025)
2.5%
  41
> quantile(myChkWts, 0.975)
  97.5%
294.575
> quantile(sampDist1, 0.025)
    2.5%
83.45227
> quantile(sampDist1, 0.975)
   97.5%
165.9114
```

7. Repeat Exercise 5 but change n = 100

   Set seed for replication

```
> set.seed(1234)
> sampDist2 <- replicate(10000, mean(sample(myChkWts, size = 100,
replace = TRUE)),
+                        simplify = TRUE)
> summary(myChkWts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0    63.0   103.0   121.8   163.8   373.0
> summary(sampDist2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  94.07  116.78  121.55  121.69  126.41  149.36
> quantile(myChkWts, 0.025)
2.5%
  41
> quantile(myChkWts, 0.975)
  97.5%
294.575
> quantile(sampDist2, 0.025)
  2.5%
108.04
> quantile(sampDist2, 0.975)
 97.5%
136.13
```

   What makes this sample better is that is makes a narrower set and better representation of the chick weight distribution. The larger sample size shows a larger set of chick weights to collect the mean. This creates a tighter quantile range which can represent most of the data in from the raw data.