Homework 1

Maya Mileva

10/7/2019

Homework 1 by Maya Mileva: I consulted with the lecture slides and materials provided by my instructor and materials and knowledge from IST 638.

Exercises

1. Using the material from this chapter and possibly other information that you look up, write a brief definition of these terms in your own words: **mean, median, mode, variance, standard deviation, histogram, normal distribution,** and **Poisson distribution.**

mean - that is the average(arithmetic mean); central value of a discrete set of numbers; to find the mean of the values in a dataset, simply add up all the numbers and devide by how many numbers you have. Sensitive to outliers.

median- middle point of the data(balancing point); the median in the data set is the *middle data value* when data are put in ascending order. Half of the data values lie below the median, and half above. If the simple size n is odd, then the median is the middle value. If the sample size n is even, then the median is the mean of the two middle data values.Not sensitive to outliers.

mode - most frequent value; the mode in the data set is the value that occurs with the greatest frequency.

! Mean, median and mode are key measures of Central tendency(center of the data).!

variance - the average "squared deviation" from the mean; the sum of squared deviations from the mean divided by the number of observations.

standard deviation - may be interpreted as the typical difference between a data value and the sample mean for given data set; the square root of the variance

! Variance, sd and range are 3 ways to measure a Dispersion of the data.!

histogram - devides data on bins, awlays equal size; graphical display of data using bars of different size, display the shape and the spread of continuous sample data(representation of the distrubution of numerical data); visual representation of the distribution of the data set.

In his book Jeffrey Stanton describes it as: "The histogram comprises a series of bars along a number line (on the horizontal or X-axis), where the height of each bar indicates that there are a certain number of observations, as shown on the Y-axis (the vertical axis). This histogram is considered a univariate display, because in its typical form it shows the shape of the distribution for just a single variable."

normal distribution - "the bell curve"; probability function that describes how the values of a variable are distributed; symetric tails an gradual curves towards a peak in the middle.

Poisson distribution - different shapes of different levels of lambda, all observations are positive.

"Discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume." - Wikipedia

2. Use the data() function to get a list of the data sets that are included with the basic installation of R: just type "data()" at the command line and press enter. Choose a data set from the list that contains at least one numeric variable— for example, the Bio- chemical Oxygen Demand (BOD) data set. Use the summary() command to summarize the variables in the data set you selected— for example, summary(BOD). Write a brief description of the mean and median of each numeric variable in the data set. Make sure you define what a "mean" and a "median" are, that is, the technical definition and practical meaning of each of these quantities.

```
## data() # function to get a list of the data sets that are included with the basic installation of R
summary(BOD) # produce result summaries
```

```
##       Time           demand     
##  Min.   :1.000   Min.   : 8.30  
##  1st Qu.:2.250   1st Qu.:11.62  
##  Median :3.500   Median :15.80  
##  Mean   :3.667   Mean   :14.83  
##  3rd Qu.:4.750   3rd Qu.:18.25  
##  Max.   :7.000   Max.   :19.80  
```

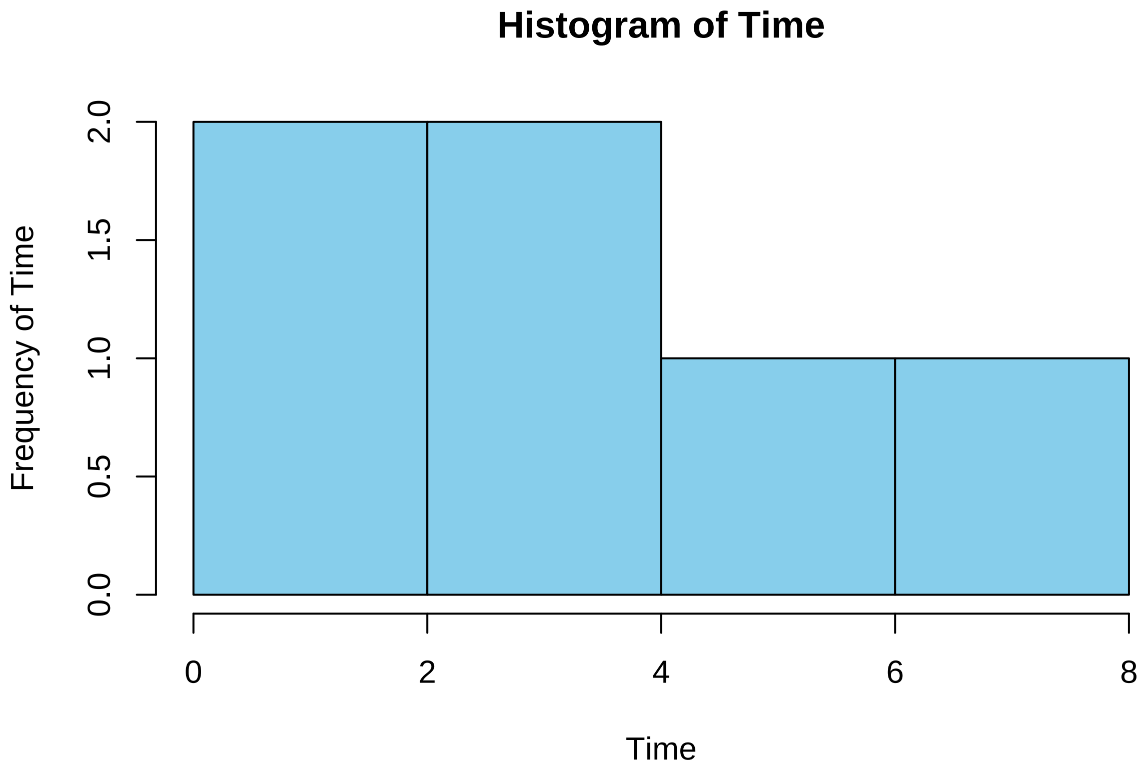
```
head(BOD, n=5) # checking the variables in the data set
```

```
##   Time demand
## 1    1    8.3
## 2    2   10.3
## 3    3   19.0
## 4    4   16.0
## 5    5   15.6
```

The data set Bio- chemical Oxygen Demand (BOD) has 2 variables - "Time" and "demand". From the summary() we can notice the mean and median calculated for each of those variables. The mean is the average value(model of our data set) in the data and the median is the middle point of the data. The median provides a helpful measure of the centre of a dataset. It separates the upper and lower half and median is a commonly used measure of the properties of a data set in statistics and probability theory. By comparing the median to the mean, we can get an idea of the distribution of a dataset. When the mean and the median are the same, the dataset is more or less evenly distributed from the lowest to highest values.

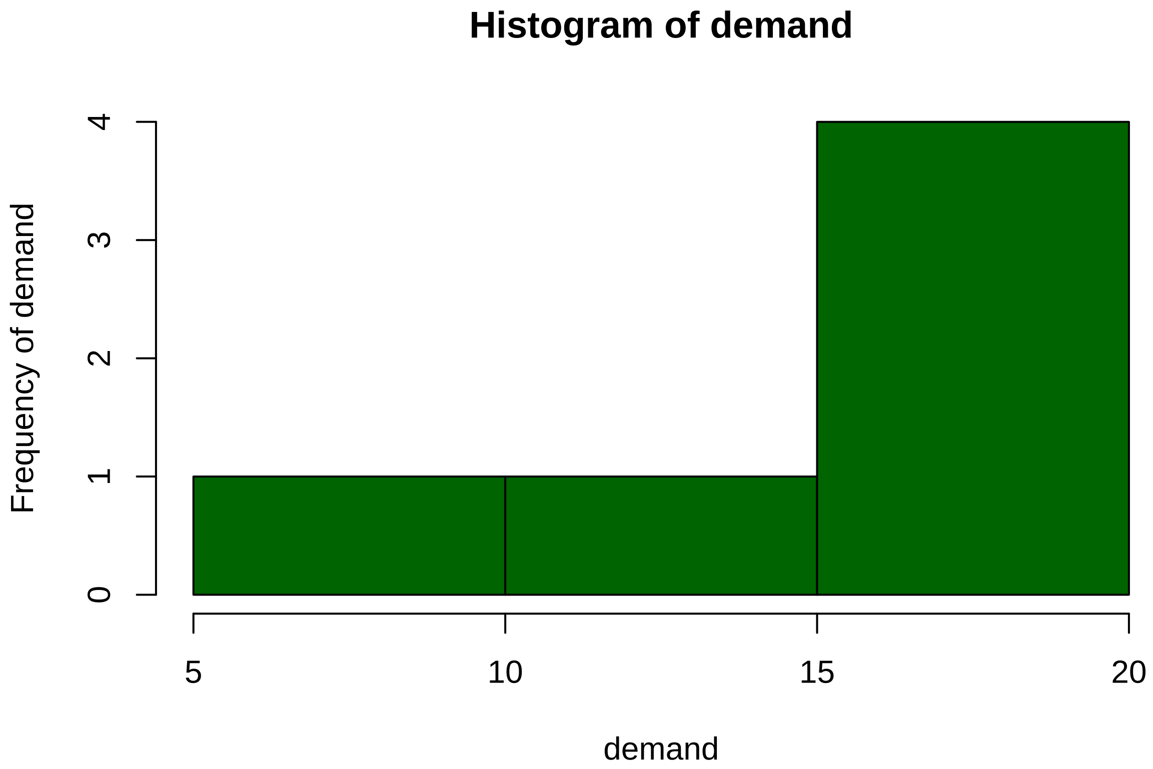
For "Time" variable we can see that the middle point is 3.5 and the average is 3.667. The range is from 1 to 7. The two values are really close.

```
hist(BOD$Time
     ,main = "Histogram of Time"
     , xlab="Time"
     , ylab="Frequency of Time"
     , col = "skyblue"
     , bty = "n"
     )
```



For “demand” variable we can see that the middle point is 15.8 and the average is 14.83. The range is from 8.30 to 19.80.(more spread out)

```
hist(BOD$demand
     ,main = "Histogram of demand"
     , xlab="demand"
     , ylab="Frequency of demand"
     , col = "darkgreen"
     , bty = "n")
```



3 As in the previous exercise, use the data() function to get a list of the data sets that are included with the basic installation of R. Choose a data set that includes just one variable, for example, the LakeHuron data set (levels of Lake Huron in the years 1875 through 1972). Use the hist() command to create a histogram of the variable— for exam- ple, hist(LakeHuron). Describe the shape of the histogram in words. Which of the distri- bution types do you think these data fit most closely (e.g., normal, Poisson). Speculate on why your selected data may fit that distribution.

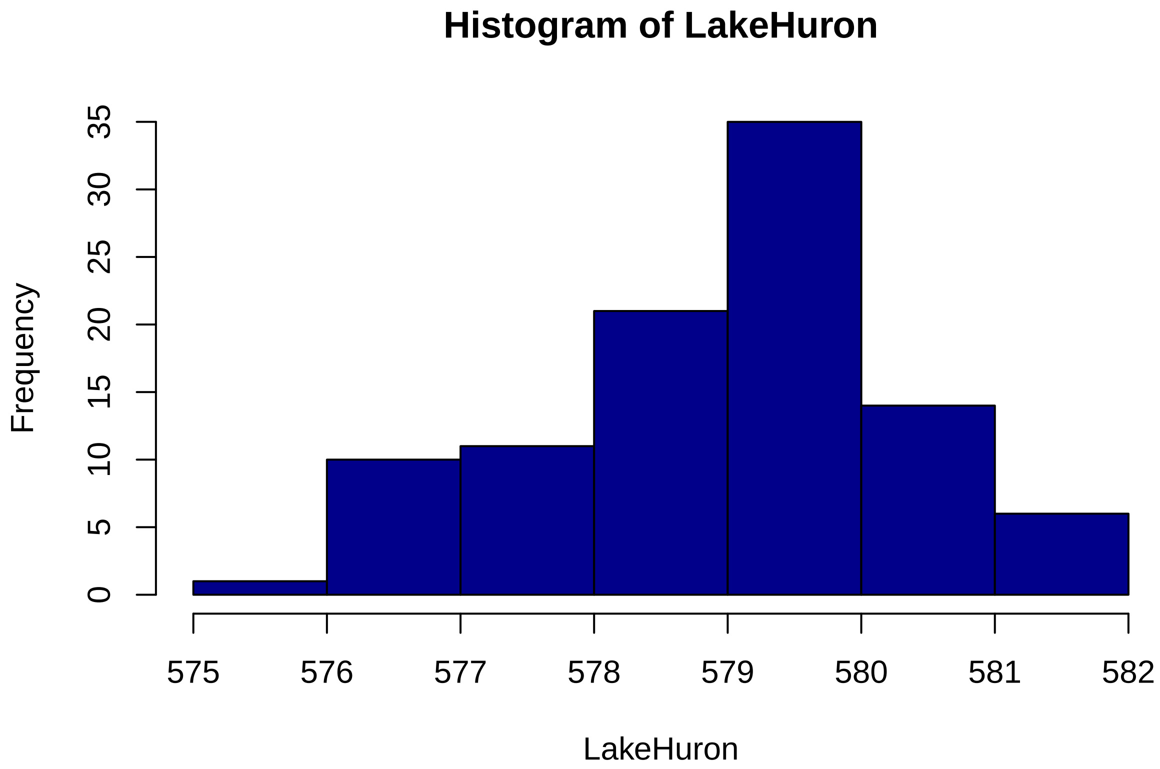
```
## data() # function to get a list of the data sets that are included with the basic installation of R
summary(LakeHuron)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   576.0   578.1   579.1   579.0   579.9   581.9
```

```
head(LakeHuron)
```

```
## [1] 580.38 581.86 580.97 580.80 579.79 580.39
```

```
hist(LakeHuron,
     col = "darkblue")
```

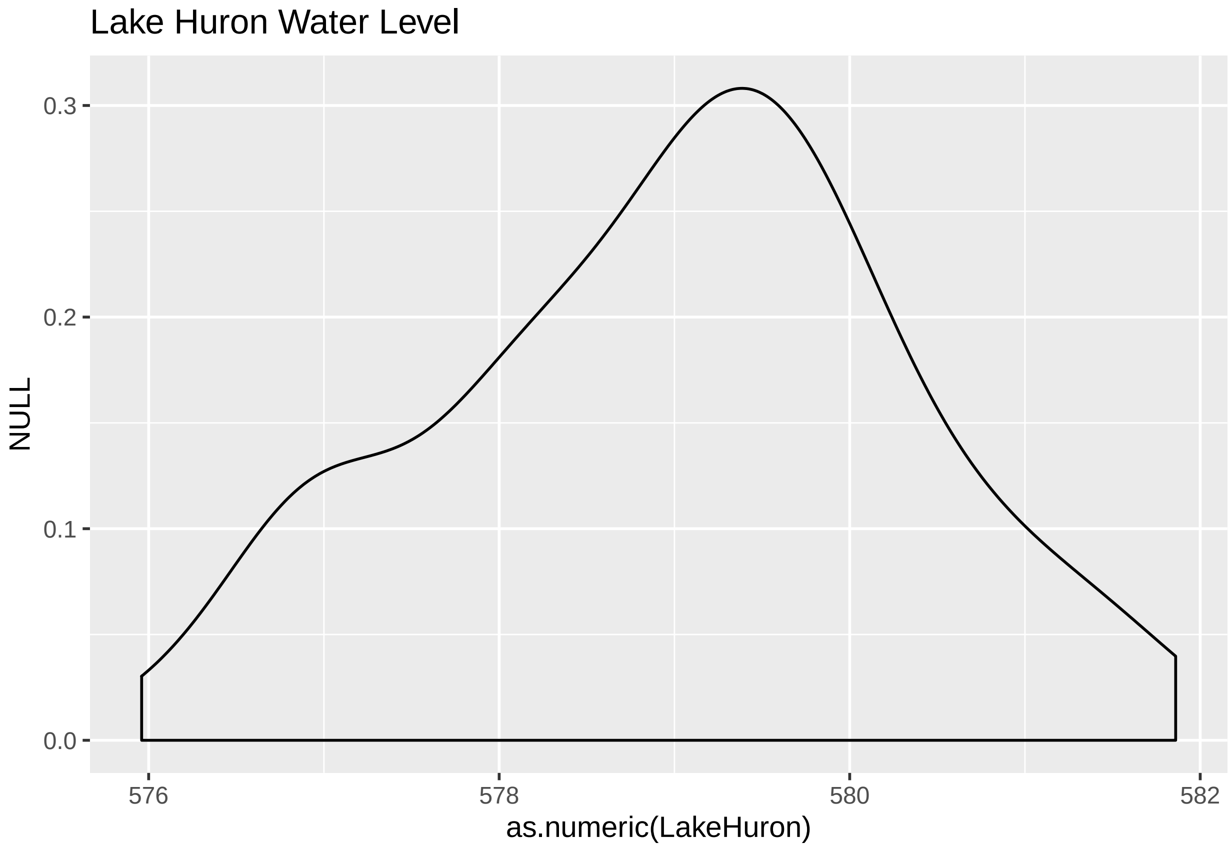


```
# pie(LakeHuron) # can see all parts are enly distributed
sd(LakeHuron)
```

```
## [1] 1.318299
```

The mean of the LakeHuron data set is 579 and the median is 579.1. By comparing the median to the mean, we can get an idea of the distribution of a dataset. When the mean and the median are the same, the dataset is more or less evenly distributed from the lowest to highest values. The histogram shape look like bell shape so it fits closely to a normal distribution. The histogram shows the distribution, if we draw a line from the edges oh the histogram we will draw a kind of bell curve. It lookes like is right skewed.

```
# how closely the data is to a normal distribution
qplot(as.numeric(LakeHuron)
      , geom="density"
      , main ="Lake Huron Water Level")
```



Normal distributions are symmetric, unimodal, and asymptotic, and the mean, median, and mode are all equal. A normal distribution is perfectly symmetrical around its center. The distribution of the LakeHuron is not symmetrical around its center so we can think of Binomal distribution, but its more likely to be Normal distribution.