Matthew L. Pergolski
IST 772
Dr. Block
10/11/2021

Homework 1

Beginning Statement

"I produced the material below with no assistance [direct quote from IST 772 class syllabus]."
Note: Homework questions have been copied/pasted into the document for both the student and viewer's convenience.

Homework Questions

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.
    a. Mean
        i. The mean is the average of a vector of numbers.
    b. Median
        i. The median is the middle most value in a vector of numbers – after the vector was sorted from smallest to largest (or vice versa).
    c. Mode
        i. The mode is the value that appears the most in a vector of numbers.
    d. Variance
        i. The variance measures the spread of a vector of numbers.
        ii. It can be described as the 'sum of squares,' as many refer to it, divided by the number of values in a vector; that is, the differences in each data value from the mean, squared, summed, and divided by however many values appear in a vector of numbers.
        iii. Another way to think of this term is being the average value away from the mean value (but inflated due to the values being squared).
    e. Standard deviation
        i. Like variance, the standard deviation also measures the spread of a vector of numbers.
        ii. The difference between standard deviation and variance is that standard deviation is the square root of the variance; the average value away from the mean (except this time, the square root function is applied to the variance).
    f. Histogram
        i. A histogram illustrates data in a 'bar chart' format and measures frequency.
    g. Normal distribution
        i. A normal distribution is a 'bell curve' shape that theoretically shows the median, mode, and mean holding the same value.

    h. Poisson distribution
        i. This distribution shows how likely something is to occur over time (i.e., arrival times).

3. 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 Biochemical 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.
    a. The dataset "cars" prints out the following when using the summary() function. There are two variables showing under this dataset: Speed and Distance. For the first variable, the average (or mean) speed was 15.4 units. As a reminder, the mean value was created by taking all numbers in the 'speed' vector, adding them together, and dividing by the total number of observations seen. The mean (or average) distance for these cars was 42.98 units.
    b. The median, or middle-most number (when re-arranging the dataset from smallest to largest or vice versa), for speed is 15 while the median for distance is 36. Outliers typically do not affect the median since it's the middle-most number shown in the dataset.
    c. Also in the summary function, we see other information such as the min and max values as well as the 1st and 3rd quartiles.

```
> summary(cars)
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

4. 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 example, hist(LakeHuron). Describe the shape of the histogram in words. Which of the distribution types do you think these data fit most closely (e.g., normal, Poisson). Speculate on why your selected data may fit that distribution.
    a. The dataset "Nile," which measures the flow of the famous Nile river, appears to show something similar to a normal distribution since we get an approximate bell-shaped curve.
    b. We do see, however, it is a not a perfect bell-shape since we appear to have a mean value of 919.4 while the median value is 893.5. We can see that some

outliers may be affecting the value of the mean, which is why it's a larger value than the median (i.e., the median is resistant to outliers).

c. I suspect this appears to be similar to a normal distribution because, over time, central tendency will correspond to a typical 'average' flow number value for the river. More often than not, the flow number value of the river will be near the mean.

```
> summary(Nile)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  456.0   798.5   893.5   919.4  1032.5  1370.0
```