IST-772 Quantitative Reasoning in Data science

Week1/HW-1: Getting Started and Statistical vocabulary

Reasoning with Data (Page-20: Exercise-1, 3, 4)

Sharat Sripada (vssripad@syr.edu)

Question-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 – Also commonly referred to as the average, is obtained by calculating the sum of values divided by number of observations. It can be represented as:

$$Mean/mu = sigma(xi)/N$$

Where xi are data points and N are the total number of observations

In R, the function mean() can be used to compute the mean of a vector.

NOTE: Mean is easily influenced by outliers in a dataset.

Median – Is also known as middle value. It is obtained by sorting a dataset or values of a vector in an ascending order and picking the middle value. While mean is pressurized by outliers, median tends to be resilient and mostly represents the middle value.

In R, the function median() can be used to compute the median of a vector.

Mode – The most frequently occurring value in a dataset is represented by the mode. Speaking more generally, mode can be the best way of finding the 'most typical' value in a vector of numbers because it picks out the value that occurs more often than any other.

In R, mode exists via a library called modeest and function, mfv() standing for most frequent value.

Mean, Median and Mode are methods of descriptive statistics commonly representing the measures of central tendency of data.

Variance – Originates from complexities in representing deviation from mean.

Typically, the best way to represent deviation was known to be by the sum of squares method. However, sum of squares tends to grow with more data points and it becomes hard to compare it as method of finding dispersion across datasets.

This problem is solved by dividing it the number of observations (N), hence known as variance. Variance can hence be represented mathematically as:

$$Var = sigma(xi – mean/mu)^2 / N$$

Where sigma(xi – mean/mu)^2 is called Sum of squares.

In R, the var() method can be used to compute the variance of a vector.

Standard Deviation – While variance is a good method to represent dispersion, its units can be confusing or hard to associate. For instance, when representing a dataset of temperatures in Celsius/Fahrenheit, variance would be square Celsius/square Fahrenheit.

Standard deviation originated there. To be able to represent dispersion in data using the same units as the data and hence, is defined as:

Std. deviation/sigma = (var) ^ ½

Variance and Standard Deviation is used to measure dispersions in data.

Histogram – Comprises of a series of bars along a number line (x-axis), where the height of each bar indicates that there are a certain number of observations, as show on the y-axis.

Histogram is also known as a univariate display since it typically represents shape of distribution for a single variable.

In R, the hist() function can be used to plot a histogram of vector.

Normal Distribution – When a distribution or plot of values takes the form of a 'bell curve' it is referred to as a normal distribution.

For example, heights of individual people are normally distributed because there are variety of different genes controlling bone shape and length in the neck, torso, hips, thighs, calves and so forth. Variations in these lengths tend to cancel out such that:

- a person of average height has data that falls in the middle of the distribution
- In few cases, all the bones are little shorter, making a person shorter, therefore in the lower part of the distribution
- And in other cases, all the bones are little longer, making a person taller, therefore in the higher part of the distribution

In a typical normal distribution, the values at the low-end of the x-axis are referred to *as left-hand tail* and the values at the high end of the axis as *right-hand tail*.

Using the rnom() alongwith a hist() function can result in plotting out a normal distribution.

Example:

```
> hist(rnorm(n=1000, mean=100, sd=10)
```

Poisson Distribution – Poisson distribution woks well when modeling arrival times. Examples of this include – arrival of buses or subway cars at a station, the arrival of customers at a cash register, or the occurrence of telephone calls at a particular exchange.

In R, the rpois() function can be used in the following manner to generate a Poisson distributon:

```
hist(rpois(n=1000, lambda=1))
```

where lambda is the Greek letter that statisticians like to represent the mean.

Question-2: Write a brief description of the mean and median of each numeric variable in the dataset Biochemical Oxygen Demand (BOD). Explain the technical and practical definition.

```
> BOD
  Time demand
1    1    8.3
2    2   10.3
3    3   19.0
4    4   16.0
5    5   15.6
6    7   19.8

> summary(BOD)
      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
```

NOTE: For numeric variable **Time**, mean and median are 3.5 and 15.8. For variable **demand** the mean and median are 3.667 and 14.83 respectively.

Using the mean and median R functions we get the same values:

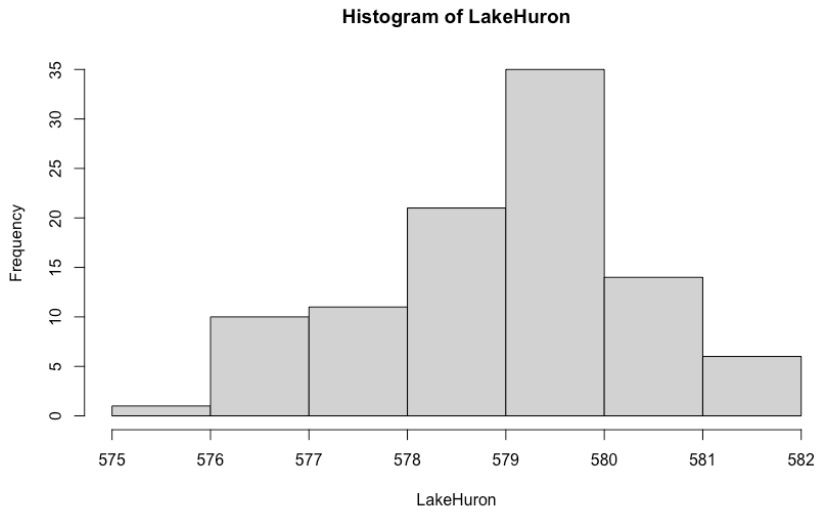
```
> df <- BOD
> colnames(df)
[1] "Time"   "demand"
> median(df$demand)
[1] 15.8
> mean(df$demand)
[1] 14.83333
> median(df$Time)
[1] 3.5
> mean(df$Time)
[1] 3.666667
```

Summary: The BOD dataset represents the biochemical oxygen demand versus time (in days) in an evaluation of water quality. The mean and median values are interpreted as follows:

- The mean or the average oxygen demand is 14.83units per 3.66days
- The median oxygen demand is 15.8units per 3.5days

Question-3: Use the hist function to create a histogram of a variable. Describe the shape of the histogram in words. Which of the distribution types do you think these data fit most closely?

Using the hist() function on the LakeHuron dataset we obtain the following chart:



This follows the pattern of a normal distribution (slightly skewed right). When drawing a curve around the edges of the bars it tends to follow a 'bell curve' with:

- Most values hovering around the mean or median
- Few values at the low-end of the x-axis are in left-hand tail – range 575-576
- And few others in the right-hand tail – range 581-582