IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: August 27, 2020
Chapter Number: #1
Title of Chapter: **Getting started and Statistical Vocabulary**

Chapter 1 - DESCRIPTIVE STATISTICS

Descriptive statistics: summarizes a vector of data by calculating a scalar value that summarizes those data. There are normally two types of them:

1. **Measures of central tendency** (summarize a vector of data by figuring out the **location of the typical, the middle, or the most common point in the data)**
    - Mean: arithmetic mean
    - Median: the halfway value – resistant to outliers
    - Mode: data that occurs the most frequently – mfv() function in R
2. **Measures of dispersion**
    - Range
    - Deviation from the Mean: aggregation of the absolute values of the deviations
    - Sum of Squares: the sum of the squared deviations from the mean – used to describe how far points are from the mean
    - Variance: the sum of squared deviations from the mean divided by the number of observations – var() function in R
    - Standard Deviation: square root of the variance – sd() function in R
3. **Distribution of Data**: (shape of the data)
    a. Normal Distribution: the shape of normal distribution looks like an upside-down bell. Most of the data follow this distribution since that "the underlying phenomenon has many small influences that add up to a combined result that has **many cases that fall near the mean**". The shape of data that follows normal distribution is univariate and symmetric. – rnorm() function in R
    - Poisson Distribution: despite of having the normally distribution, Poisson distribution also fits the natural phenomena. One difference between poison distribution and normal distribution is that **poison takes discrete variable while normal distribution takes continuous variable.** – rpois() in R function

Insert Box - Mean and Standard Deviation Formula

Mean:

$$\mu = \frac{\sum x}{N}$$

The formula is used to calculate the mean value of a vector of data. Specifically, the μ equals to the sum of all the data divided by the number of data points. The symbol of Σ represents aggregation while the symbol of μ normally represents the population mean.

Standard Distribution:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

This formula equals to the square root of the aggregation of squared deviations divided by the number of data points. Similarly, σ represents population standard deviation.
Attention: small letter sigma (σ) indicates the standard deviation and the capital letter sigma (Σ) is the summation symbol.

Exercise Review

1. Mean: value obtained by the sum of all data points in the population and divide by the number of data points.
   Median: the value that occurs in the middle of a list of data points. If the number of data points is even, use the mean of the two numbers in the middle of the list.
   Mode: the number occurs the most frequently
   Variance: the sum of squared deviations from the mean divided by the number of observations.
   Standard Deviation: the square root of deviation.
   Histogram: a graph that can be used to demonstrate the distribution of a vector of data.
   Normal Distribution: A bell shape data distribution type. Specifically, majority of the data points fall around the mean area. Fewer data points fall inside the "extreme side".
   Poisson Distribution: the shape of Poisson distribution changes all the time. When the mean approaches the medians close enough, the shape of Poisson distribution looks similar with normal distribution.

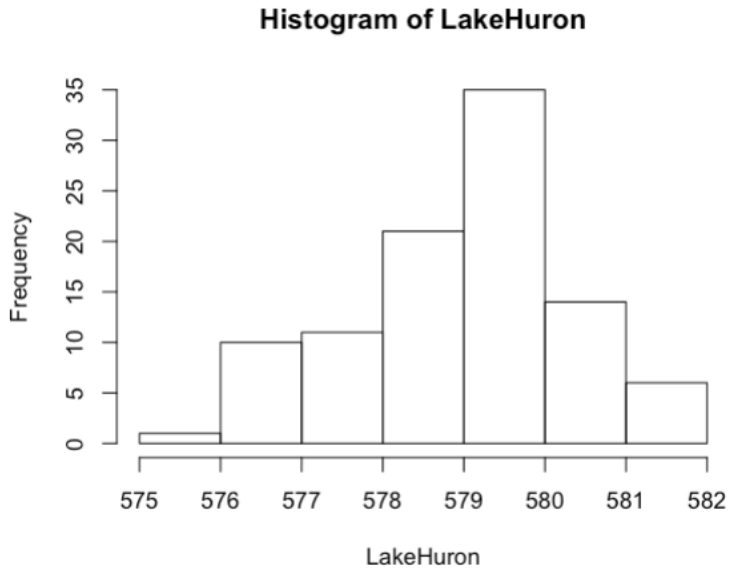
```
12  data("BOD")
13  summary(BOD)
13:13  (Top Level) ⇕

Console ~/

> data("BOD")
> 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
```

2. The mean of Time variable is 3.667 while the median is 3.500. Similarly, the mean of demand variable is 14.83 while the median is 15.80. Here, since I am not sure about the specific meaning of Time, I could only conclude that the mean of the seven number is 3.667 and the median is 3.500. For the mean of demand variable, it could be used to represent the average demand among those seven observation periods. The mean of demand variable is slightly lower than the median, so the graph of demand variable is slightly left-skewed.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.



3. The shape of this graph is slightly left-skewed. The mean (which is 579.0041) is slightly lower than the median (which is 579.12). However, the difference is pretty tiny, and the number of observations is large enough (which is 98). Hence, I reckoned that this graph fits the shape of normal distribution.

R Code Fragment and Explanation

hist(x, breaks, …) - computes a histogram of the given data values
Typical Parameters:

x: A vector of data for which histogram is desired

freq: optional parameter which indicates whether the histogram graphics is a representation of frequencies.

Breaks: optional parameter which indicates the breakpoints between histogram cells.

In the exercise, we can directly use hist() function to depict the distribution of data points.

Question for Class

1. Wish to know more detailed difference between Normal Distribution and Poisson Distribution.