Lecture Notes -3

Descriptive Statistics:
It is used to provide summary statistics of the data.

History:
Here are a few people who contributed to the statistical party.
Francis Galton- eugenics and peas
Karl Pearson- correlation and regression
William Sealy Gosset- small sample statistical techniques and beer
Ronald Fisher- analysis of variance and farms

There is always uncertainty in data and we always consider sample of the data for which descriptive statistics and inferential statistics are used.

Key measures of data:
Measure of central tendency- mean, median and mode
Dispersion is how much data is distributed
Measure of Dispersion- range, variance and standard deviation

Why is understanding central tendency and measure of dispersion useful?
Central tendency and measure of dispersion are useful to get more information about the data. Central tendency is measured by mean, median and mode and dispersion is measured by range, variance and standard deviation.

Histogram:
Histogram is a picture that shows central tendency and dispersion. It is designed to show frequencies.

Code:
Hist (USstatePops$april10census, breaks=20) #creates a histogram with 20 bins
Bell or normal distribution is meant to represent normal shape of a bell which is the most typical distribution.
Hist (rnorm(51,605,682))
#hist is used to create a histogram and rnorm() function returns ‘n’ data points from a normal #distribution. rnorm(n,mean,standard deviation)

Key pieces of information to enable comparisons:
The distribution had a characteristic shape. The distribution had a center point, mean and a “spread” (variability) which was the standard deviation.

Use:
This can be used to understand for example a incase of gym- Minutes spent in the gym, how often, many people visit the gym

Function:
It is a bundle of code that can be used over again without retyping.
Example:
Mean ()- Any vector is an argument to the mean which returns the mean value of that vector.
function() - Creates a new function.
return() - Completes a function by returning a value.
tabulate() - Counts occurrences of integer-valued data in a vector.
unique() - Creates a list of unique values in a vector.
match() - Takes two lists and returns values that are in each.
mfv() - Most frequent value (from the modest package).

Code:

```
#MyMode-function name, myVector-Input argument

MyMode<-function(myVector) #create function MyMode
{
Return(myVector) #send back results to function
} #curly brackets send function code

tinyData<-c (1,2,1,2,3,3,3,4,5,4,5) #create vector tinyData
tinyData #display content of tinyData
MyMode(tinyData) #pass argument tinyData to function MyMode
MyMode #display contents of arguments passed

MyMode<-function(myVector)
{
uniqueValues<-unique(myVector) #add function to MyMode
uniqueCounts<-tabulate(myVector) #add tabulate function to MyMode
return(uniqueValues[which.max(uniqueCounts)] #return the unique value that has highest unique
count associated with it
}
MyMode(tinyData)
```

Provide some examples of "Data Science in the real world"
Data Science can be used by a supermarket chain which can be used to analyze revenue for a particular month/week. It can be used to find out analysis for a single product to manage inventory of the product, the sales of the product, effectiveness of marketing campaign on sales and many other uses. Other application is analyzing the correlation between products for example - people purchasing milk also buy bread

MIS vs Data Science
MIS is Management Information Systems which is used for decision making by managers for control and coordination. It can involve dashboards, reports, analysis of historic data for decision making. It is managing information and using information in the right way.

Data Science is field that uses scientific methods, process, modeling, algorithms to understand data to gain insights from it. It involves various domains such as statistics, data analysis, machine learning and various methodologies.