STAT 135 LECTURE 1

SUMMARISING DATA

Objectives:

- **To showcase summaries (quantitative and categorical) generated using random samples provides an insight into the structure of data.**
- **To discuss central tendencies (example mean) as a realization of n independent random variables.**
- **To perceive and comprehend the variability in sample data.**
- **To juxtapose the strength and weaknesses of various measures of central tendencies and variability.**

Sections Covered: 10.1, 10.2.3, 10.3, 10.4.1, 10.4.2, 10.5,10.6,10.7

Video: Practical Usage of Statistics:

http://www.history.com/shows/mankind-the-story-of-all-of-us/videos/cholera-outbreak

Readings:

Tufte's Statistical Analysis regarding John Snow and Cholera epidemic as well as Challenger Tragedy.

https://www.sfu.ca/cmns/courses/2012/801/1-Readings/Tufte%20Visual%20and%20Statistical%20Thinking.pdf

R SOFTWARE

R Installation:

https://www.r-project.org/

Rtutorial:

https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=r+tutorial+beginners

http://www.r-tutor.com/elementary-statistics/numerical-measures

R: Document by Emmanuel Paradis on bCourses

CASE STUDY SCENARIO:

Heart Disease Detection: UCI Irvine Repository

http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart) This is a data set consisting of 270 observations for 13 parameters/attributes related to the heart .The last variable is the diagnostic variable determining whether a person suffers from heart disease or not.

Information regarding Heart Attacks:

http://www.nytimes.com/health/guides/disease/heart-attack/print.html

Attribute Information:

---

1. age
2. sex (0:Female 1:Male)
3. chest pain type (4 values)

Value 1: typical angina (Definite heart discomfort)
Value 2: atypical angina (Probable)
Value 3: non-anginal pain
Value 4: asymptomatic (no symptoms)

4. resting blood pressure
5. serum cholestoral in mg/dl HDL+LDL+Triglycerides Normal is below 200.Above 400 is high risk.
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2) Normal ,Having ST T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria. Higher levels higher risk of heart attack
8. maximum heart rate achieved (Used by doctors for heart risks)

9. exercise induced angina an indicator for heart disease related to obstruction of artery
10. oldpeak = ST depression induced by exercise relative to rest `(slope (up, flat, down)`
11. the slope of the peak exercise ST segment Heart rate slope calculated while excercising. The higher the slope the more chances of heart attack
12. number of major vessels (0-3) colored by flourosopy Detecting coronary calcification which is an indicator of heart disease
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable

QUANTITATIVE DATA

Measures of Central Tendency:

Mean : Is the average (Most typical value) of the given numerical values. All the observations will have to be added and divided by number of observations. MEAN is like a fulcrum balance point ie CENTER OF GRAVITY that balances the weight values.

$\bar{x} = \frac{x_1 + x_2 \ldots\ldots\ldots\ldots x_n}{n}$ **Sample mean is depicted by x bar or y bar**

$\bar{x} = \frac{\sum x}{n}$

$\mu = \frac{\sum X}{N}$ **Population mean is depicted by mhu**

```
getwd()
"C:/Users/seemasaharan/Documents"

setwd("C:/Users/seemasaharan/Documents/STAT135")

list.files(getwd())

library(readxl)
```

```
HeartAttack <- read_excel("HeartAttack.xlsx")

Resting_Blood_Pressure <- HeartAttack$Resting_Blood_Pressure
mean(Resting_Blood_Pressure)

[1] 131.3444
```

Median: Is the midmost value. It is called the 50th percentile. This is the observation X of the data set which has 50% of the values below it and 50% of the values above it.

```
median(Resting_Blood_Pressure)
[1] 130
```

Percentile: Is the percentage of values from the lower end of the distribution to the percentage required. p% percentile includes the values that are less than or equal to p% of the values.

MEAN

- Is the average of all the data values
- Data has one mean value
- Mean is sensitive to a few extreme values (outliers) and therefore is not a resistant measure.
- Mean value for 90,90,95,90,95,2,3 is 66.7
- Mean of 2 subsets can be combined to get a cumulative mean for the 2 groups
- Used for Quantitative data only
- Cannot be calculated for open intervals

MEDIAN

- Is the midmost value from amongst all values.
- Data has one median value
- Median is not sensitive to a few extreme values (outliers) and therefore is a resistant measure.
- Median value for 2,3,90,90,90,95,95 is 90
- Median of 2 subsets cannot be combined to get a cumulative median for the 2 groups
- Median is applicable to quantitative and qualitative data.
- Can be calculated for open interval

MEASURES OF DISPERSION

Standard Deviation: Is the square root of the variance.
Variance :is the average of square of deviations therefore the unit of variance is a squared unit of the original data wheras the data has an unsquared unit. To bring it back the variance to an unsquared unit we take a square root. This is measure called the standard deviation.

Σ Summation is the cumulative addition sign

$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$ **sigma square σ^2 is the variance of the population.**

$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$ **sigma σ is the standard deviation of the population**

$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ **s is the sample standard deviation and n-1 is the number of degree of freedom.**

```
> var(Resting_Blood_Pressure)
[1] 319.0371
> sd(Resting_Blood_Pressure)
[1] 17.86161
```

STANDARD DEVIATION

- Standard deviation is represented by an algebraic equation and can be manipulated algebraically (used in research)
- Most robust as it takes all data values into account
- Standard Deviation is sensitive to a few extreme values (outliers) and therefore is not a resistant measure.

INTERQUARTILE RANGE

- IQR is Simple to calculate.
- Not as robust as it does not takes all data values into account
- Interquartile range is not sensitive to a few extreme values (outliers) and therefore is a resistant measure.

First Quartile: 25 th Percentile. Splits the data into lower 25% of the observations and higher 75% of the data

Second Quartile: Is the median of the observations Splits the data into lower 50% of the observations and higher 50% of the data
Third Quartile: 75 th Percentile. Splits the data into lower 25% of the observations and higher 75% of the data

```
quantile(Resting_Blood_Pressure)
 0%  25%  50%  75% 100%
 94  120  130  140  200
```

Summary:

Minimum Q1 Q2 (Median) Mean Q3 Maximum

```
summary(Resting_Blood_Pressure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   94.0   120.0   130.0   131.3   140.0   200.0
```

CATEGORICAL DATA

Categorical Variables can be summarized by the use of Counts, Proportions and Relative Frequencies . Relative Frequencies provide a good measure if there are unequal values in the given categories .

The Counts Percentages generated by R : Variable used are Pain Type and Gender.

Following these measures the Crosstabulated Data (Contingency Table) is given for these two variables.

Pain Type

Sex 1:Male 0:Female

```
PainType <- table(HeartAttack$Chest_Pain_Type)
> PainType

 1   2   3   4
20  42  79 129

>prop.table(PainType)*100

    1        2        3        4
7.407407 15.555556 29.259259 47.777778


>addmargins(prop.table(PainType)*100)

     1        2        3        4      Sum
 7.407407 15.555556 29.259259 47.777778  100


>Gender <- table(HeartAttack$Sex)
> Gender

 0  1
87 183
> addmargins(prop.table(Gender)*100)

    0        1      Sum
32.22222 67.77778 100.00000
PainTypeGender <- table(HeartAttack$Test_Pain_Type,HeartAttack$Sex)

addmargins(PainTypeGender)

      0   1   Sum
1     4  16    20
2    16  26    42
3    32  47    79
4    35  94   129
Sum  87 183   270

addmargins(prop.table(PainTypeGender)*100)

           0         1       Sum
1   1.481481  5.925926  7.407407
2   5.925926  9.629630 15.555556
3  11.851852 17.407407 29.259259
4  12.962963 34.814815 47.777778
Sum 32.222222 67.777778 100.000000
```

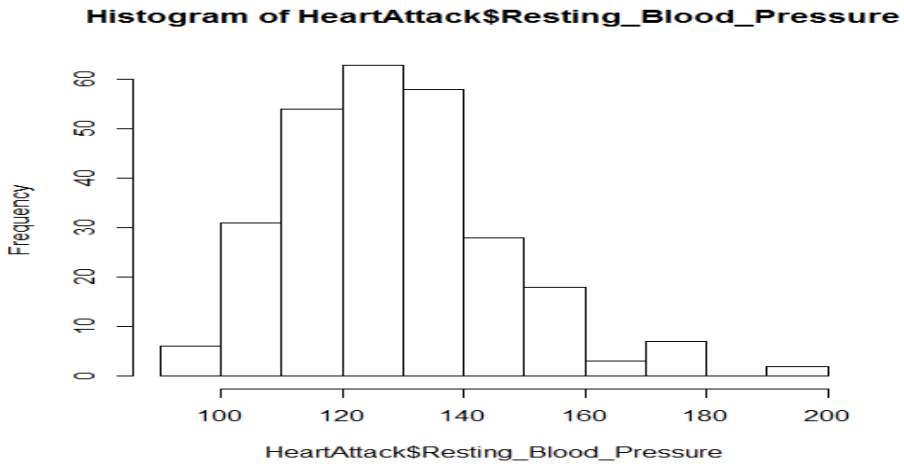
GRAPHICAL SUMMARIES **Graphical summary Enables us to visually ascertain the mean, median, mode (measures of central tendencies), standard deviation (spread/dispersion/deviation) and the shape (skewed or bell shaped) of the distribution.**

QUANTITATIVE DATA

Histogram

Consists of parallel vertical bars that depict the frequency distribution of the quantitative data set. The height of the bar is the frequency of that data interval.

```
> hist(HeartAttack$Resting_Blood_Pressure)
```



Slightly Right Skewed

Stem and leaf

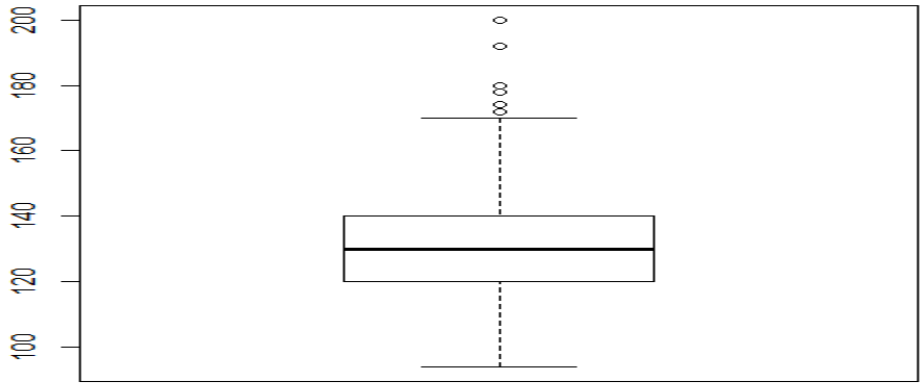
Is a graphical display that displays quantitative data according to the most significant digit in the data set. Each row is placed adjacent to the next according to the ascending order.

```
> stem(HeartAttack$Resting_Blood_Pressure)

 The decimal point is 1 digit(s) to the right of the |

  9 | 44
 10 | 000012245556888888
 11 | 00000000000000002222222255578888888
 12 | 000000000000000000000000000000002223444445555555556668888888889
 13 | 00000000000000000000000000000022222244445555556668888888888
 14 | 0000000000000000000000000000022245555568
 15 | 000000000000000002222568
 16 | 000000000005
 17 | 002488
 18 | 000
 19 | 2
```

```
20 | 0
```

Boxplot This is a visual display of quantitative data and is shows the division in terms of Min Q1 Q2 Q3 Max and outliers

```
boxplot(HeartAttack$Resting_Blood_Pressure)
```



Slightly right skewed. The lower 25% of the data set is less dispersed ie the BP of the first 25% of the data has less variation as compared to the highest 25% of the dataset.

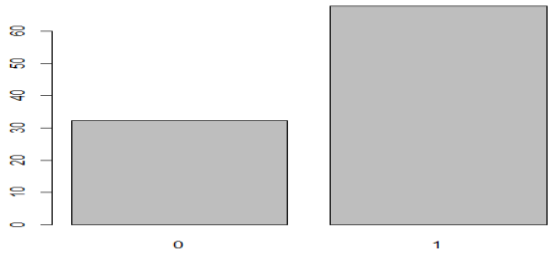
Lowest 25% of the data lie between 94---120 (Difference of 18 points)

Highest 25% of the data lie between 140---170 (Difference 30 points)

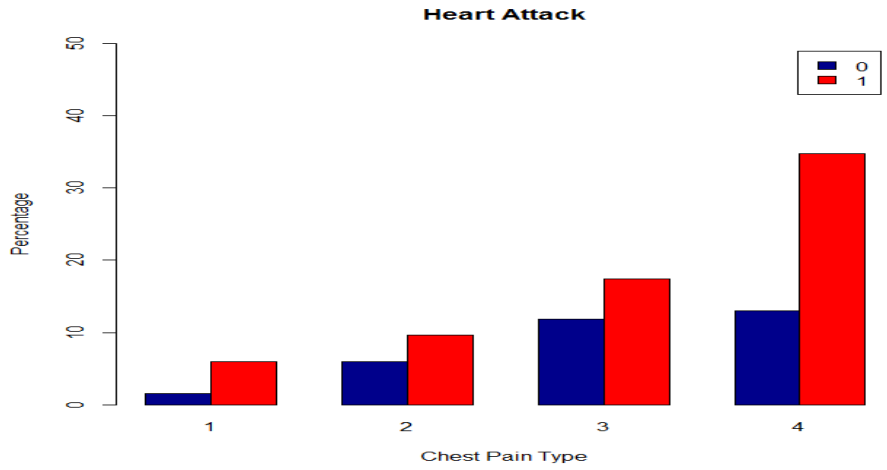
CATEGORICAL

Bar (Used for qualitative/categorical data) A bar diagram represents the frequency of occurrences of different values of X .It is represented by the height of a bar. The bar diagram helps differentiate between the independent variables and the dependent variables. Side by side bar graphs can be used for comparative analysis. Pareto graph is when instead of frequency, relative frequency is used. Horizontal bars are also used for bar diagrams. Start at zero rules are used for bar diagrams in order to ensure that insignificant differences are not exaggerated.

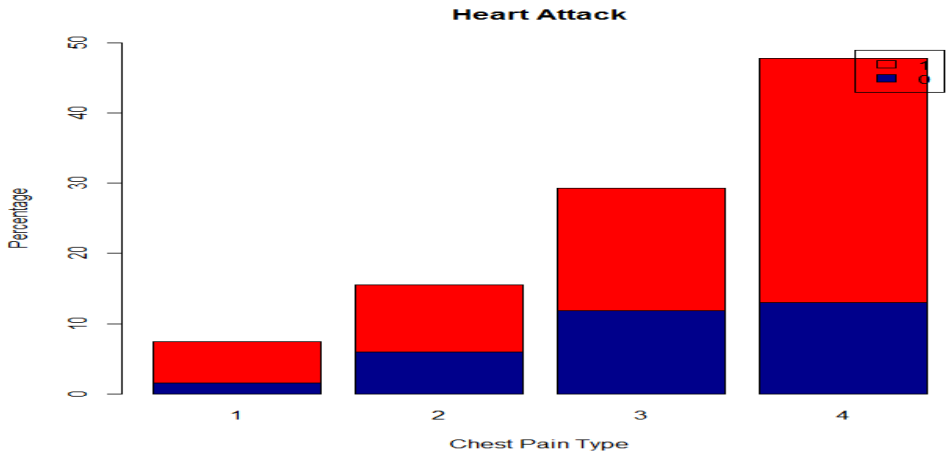
```
prop.table(Gender)*100

     0      1
32.22222 67.77778
Gender <- table(HeartAttack$Sex)
> barplot(prop.table(Gender)*100)
```




```
prop.table(Gender_Chest_Pain_Type )

          1          2          3          4
 0 0.01481481 0.05925926 0.11851852 0.12962963
 1 0.05925926 0.09629630 0.17407407 0.34814815
Gender_Chest_Pain_Type <- table(HeartAttack$Sex,HeartAttack$Chest_Pain_Type)
barplot(prop.table(Gender_Chest_Pain_Type)*100, main="Heart Attack",
xlab="Chest Pain Type",ylab="Percentage",col=c("darkblue","red"),legend =
rownames(Gender_Chest_Pain_Type),ylim=c(0,50),beside=TRUE)
```



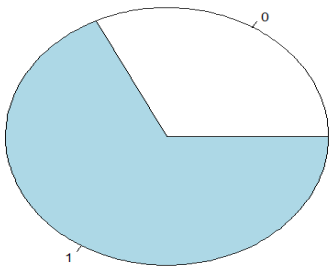
```
prop.table(Gender_Chest_Pain_Type )

        1           2           3           4
0 0.01481481 0.05925926 0.11851852 0.12962963
1 0.05925926 0.09629630 0.17407407 0.34814815
Gender_Chest_Pain_Type <- table(HeartAttack$Sex,HeartAttack$Chest_Pain_Type)
barplot(prop.table(Gender_Chest_Pain_Type)*100, main="Heart Attack",
xlab="Chest Pain Type",ylab="Percentage",col=c("darkblue","red"),legend =
rownames(Gender_Chest_Pain_Type),ylim=c(0,50))
```



It can be clearly seen that approximately 6% of the individuals experienced Type 1 pain. Out of these 1% were women and 5% were men.

PIE Chart

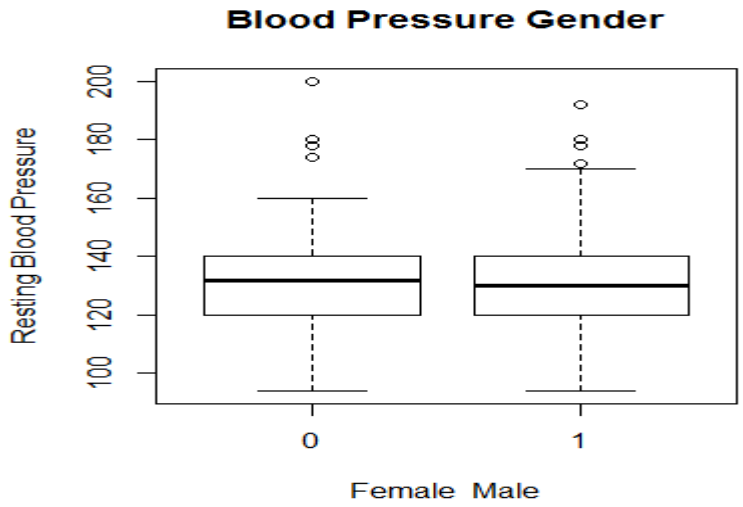
```
pie(prop.table (Gender)*100)
```



QUANTITATIVE GROUPED BY CATEGORICAL

BOXPLOTS Provides the graphical display of Resting Blood Pressure grouped by Gender

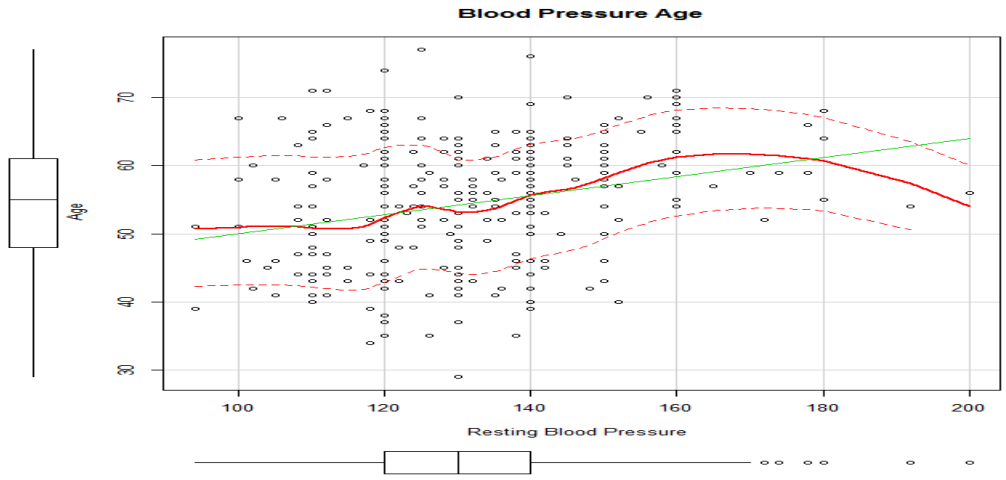
```
boxplot(Resting_Blood_Pressure~Sex,data=HeartAttack, main="Blood Pressure Gender",
                    xlab="Female(0)    Male(1)", ylab="Resting Blood Pressure")
```



It can be clearly seen that overall Female have a slightly higher median Blood pressure. The lower 25% of both males and females have the same variability/range of blood pressures but for the upper most 25% of the data males have a higher variability as compared to females. They both have around 4 outliers and female have the most extreme outlier. Both distributions are slightly right skewed.

Scatter plots Provides the correlation estimate of a bivariate data set.

```
scatterplot(Age ~Resting_Blood_Pressure,data=HeartAttack, main="Blood Pressure Age",
                    xlab=" Resting Blood Pressure  ", ylab=" Age")
```



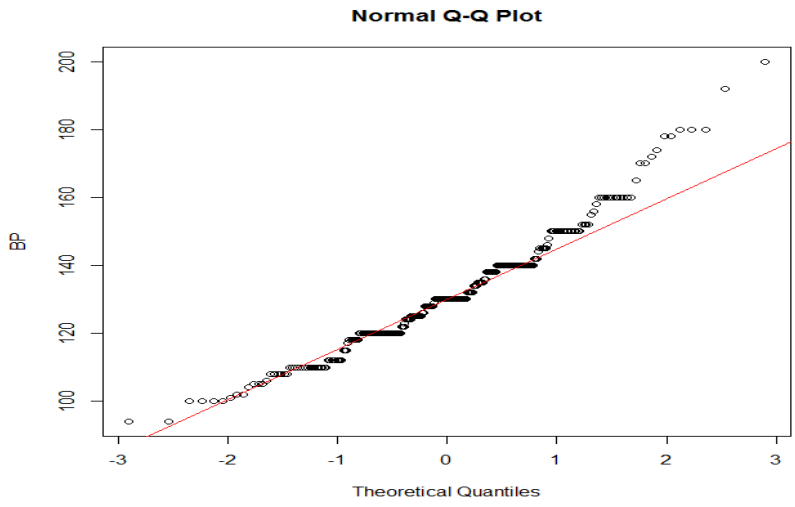
Quantile Quantile plots Are used to check the validity of the distributional shape of a data set. The quantile quantile plots can be constructed for two data sets or a dataset and a theoretical distribution. If the dataset follows a theoretical distribution then it will follow the straight diagonal line y=x. If the data set does not follow the line then the dataset does not exactly follow the theoretical distribution and might be skewed.

If the data is curved like a C facing upwards close to and cutting the line the line y=x then it is right skewed (maybe will fit lognormal distribution) wheras if it is C facing downwards close to and cutting the line it is left skewed. If it follows the line it is normally distributed.

If the data is s shaped about the line y=x and has short tails ie few points deviate away from the tails and heads then it is not normal.

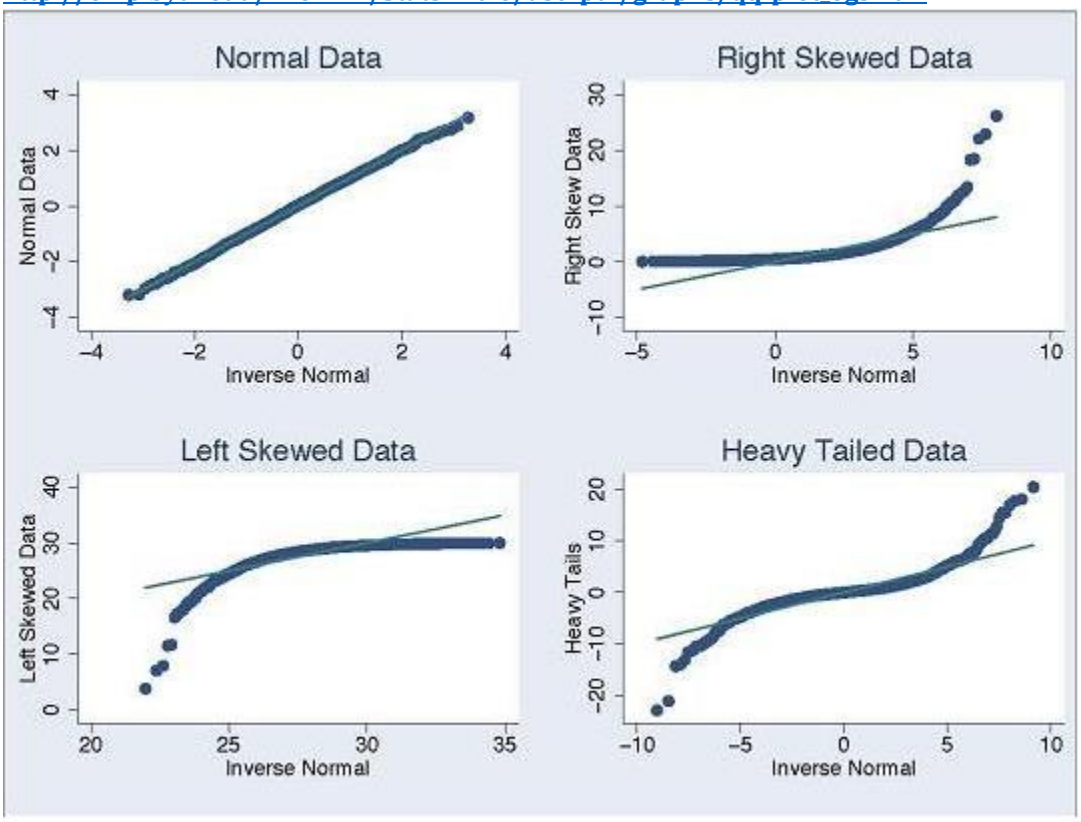
If the data is s shaped about the line y=x and has long tails ie a large number of points deviate away from the tails and heads then it is not normal.

```
qqnorm(HeartAttack$Resting_Blood_Pressure, ylab =
"BP");qqline(HeartAttack$Resting_Blood_Pressure,col=2)
```



The distribution is clearly right skewed as the right tail deviates from the line on the right.

http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot_egs.htm



A nice visual analysis of qq plots

http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot_egs.htm

Histogram

- This does not show all the values of the data.Practical for a moderate/large data set but not small data set.
- It used to compare two data sets
- Interval flexibility in terms of number and width. Generally 6-15 intervals shows the shape of the distributions.

Stem and Leaf

- This shows all the values.Practical for a small data set
- It used to compare two data sets.
- **6-15 stems intervals gives a good idea about the shape of the distribution.**
- **As compared to histogram the interval is constrained by rules of leaf allocation.**

Box and Plot

- This does not show all the values of the data.Practical for a large data set.
- It gives the 5 number summary of the data.outliers are identified.
- It used to compare two data sets
- **Bimodal data is not identified.**

Lab/Discussion

Q42,Q44,Q46 a) and c),47

Home Work:

Analyse the Heart Attack Indicator data set using R:

1) **Expore, analyse and record the descriptive summaries and graphical summaries for the variables Chest_Pain_Type, Serum_Cholesterol, Resting_Electrocardiographic_ Reading, Maximum_Heart_Rate, Number_Blood_Vessels_Calcified, thal, Heart_Attack_Diagnosis.**
   Make sure to use the correct measures depending on the attributes being either quantitative or categorical. Please explain your findings.
2) **Create a box plot of Serum_Cholesterol grouped by Heart_Attack_Diagnosis. Give appropriate observations like which category has a higher median, which group has more variability and in which quartile region, are there any outliers?**
3) **Create a scatter plot and analyse the correlation between Serum_Cholesterol and Age.**
4) **Using Quantile-Quantile plots analyze and report the distribution characteristics of Serum_Cholesterol.**