Probability Distributions:
Foundation for Statistical Inference and Modeling

DATA SCIENCE INSTITUTE

What is frequency distribution?

X(Number of Children in a family)	Households (Frequency)
0	20
1	130
2	200
3	40
4	10
	400

This is the frequency distribution.

What is probability distribution?

X(Number of Children in a family)	Households (Frequency)	Relative Frequency
0	20	0.05
1	130	0.325
2	200	0.5
3	40	0.1
4	10	0.025
	400	1

. Seven coins are tossed and number of heads noted.
The experiment is repeated 128 times and the following distribution is obtained.

No. of Heads	0	1	2	3	4	5	6	7
Frequencies	7	6	19	35	30	23	7	1

Number of Heads	Frequencies
0	7
1	6
2	19
3	35
4	30
5	23
6	7
7	1
	128

Probability distribution for coin experiment

Number of Heads	Frequencies	Relative Frequency
0	7	0.0547
1	6	0.0469
2	19	0.1484
3	35	0.2734
4	30	0.2344
5	23	0.1797
6	7	0.0547
7	1	0.0078
	128	1

Standard Discrete Distributions

1. Bernoulli Distribution
2. Binomial Distribution
3. Poisson Distribution

Bernoulli Distribution

A trial in which there are two possible outcomes is called as a Bernoulli trial. Two outcomes are generally called as ‘success’ and ‘failure’.

A Bernoulli random variable takes values 1 and 0 with probabilities p and (1-p) respectively.

$Pr(X=1)=p$ and $Pr(X=0)= 1-p$

The Bernoulli distribution is named after Swiss scientist Jacob Bernoulli.

Binomial Distribution

Assume that Bernoulli trial is repeated ‘n’ times independently under identical conditions.

Probability of success is constant in all n trials.

Let X: Number of successes in ‘n’ trials

X is said to follow Binomial distribution with parameters ‘n’ and ‘P’

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

K=0,1,2…….n

Probability distribution for coin Experiment using "Binomial Distribution"

Number of Heads	Frequencies	Relative Frequency	Binomial Distribution
0	7	0.0547	0.0078
1	6	0.0469	0.0547
2	19	0.1484	0.1641
3	35	0.2734	0.2734
4	30	0.2344	0.2734
5	23	0.1797	0.1641
6	7	0.0547	0.0547
7	1	0.0078	0.0078
	128	1	1

DATA SCIENCE INSTITUTE

Binomial Distribution: Probabilities using R

The probability of getting a defective item is 2% (0.02)
What is the probability of getting 3 defective items in a pack of 10 items ? What is the probability of getting at most 3 defective items?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here n=10 an d p=0.02

Pr(X=3)

```
dbinom(3,10,0.02)

[1] 0.0008334005
```

Pr(X<=3)

```
pbinom(3,10,0.02)

[1] 0.9999695
```

Poisson Distribution

Poisson distribution can be considered as a limiting case of Binomial distribution where ‘n’ is large and ‘p’ is small. In other words, chance of a success is very small and trial is repeated large number of times.
The mean np is of intermediate magnitude.

The distribution is named after French mathematician Siméon Denis Poisson.

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$x = 0, 1, 2, 3, 4, \ldots\ldots$

$e = 2.71828$

$\lambda =$ long run average

Poisson Distribution has Mean=variance=λ

Poisson Distribution: Probabilities using R

Probability of receiving a ‘complaint’ call in the call center is 0.01.
Out of 1200 calls expected in a day, what is the probability of receiving more than 10 complaint calls in a day ?

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$x = 0,1,2,3,4,\ldots\ldots$

$e = 2.71828$

$\lambda = \text{long run average}$

λ= 1200 * 0.01= 12

Pr(X >10)= 1- Pr(X<=10)

```
1-ppois(10,12)
```

[1] 0.6527706

Standard Continuous distribution

1. Normal Distribution.

2. Chi-square Distribution.

3. Student’s t Distribution.

4. F Distribution.

Normal Distribution

Parameters: Mean = μ

Variance = σ^2

If mean = 0 and Variance = 1 then the normal distribution is called as standard normal distribution.

Normal distribution curve is symmetric bell shaped curve.

Assessment of ‘Normality’

Testing Normality Assumption

- An assessment of the normality of data is a prerequisite for many statistical methods.

- Normality can be assessed using two approaches: graphical and numerical.

- Normality assumption can be assessed using

- ❑ Box-Whisker Plot (actually for assessing symmetry)
- ❑ Quantile-Quantile Plot (QQ Plot)
- ❑ Shapiro-Wilks test
- ❑ Kolmogorov-Smirnov Test

Example
Assessing Normality Assumption

- The following data has two variables recorded on 80 guests in a large hotel.

❑ Customer Satisfaction Index (csi)

❑ Total Bill Amount in thousand rs. (billamt)

id	csi	billamt
1	38.35	34.85
2	47.02	10.99
3	36.96	24.73
4	43.07	7.9
5	38.77	9.38
6	63.04	9.49
7	43.17	19.58
8	35.14	6.15
9	38.33	13.29
10	38.7	9.62
11	31.44	8.51
12	34.87	14.49
13	24.49	13.59
14	36.84	5.3
15	58.05	15.55

Starting With Box Plot

Box Whisker Plot summarizes a variable using 5 points:

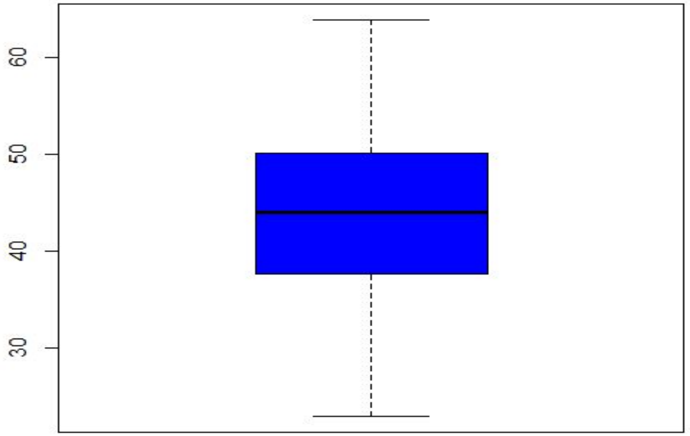
- Minimum
- First quartile (q1)
- Median(q2),
- Third quartile(q3)
- Maximum.

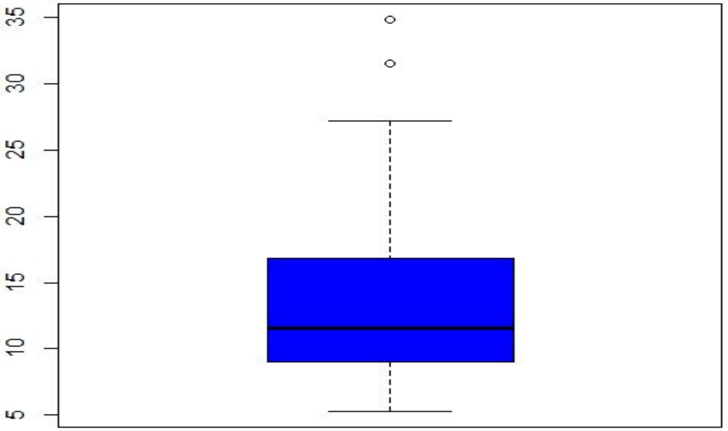
It is used to assess symmetry rather than Normality.

Box Plots in R

#import csv data “Normality Assessment Data”

```
testdata<-read.csv(file.choose(),header=T)
boxplot(testdata$csi,col="blue")
boxplot(testdata$billamt,col="blue")
```



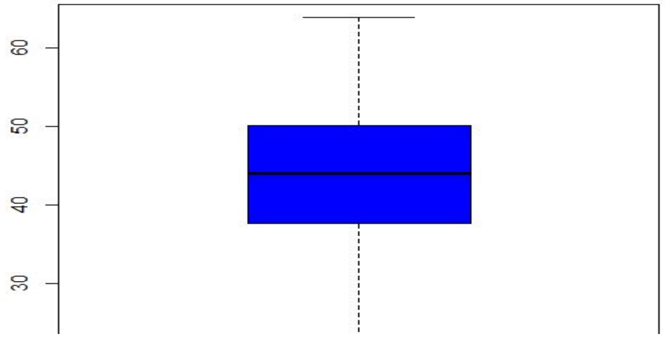


Checking Skewness Values

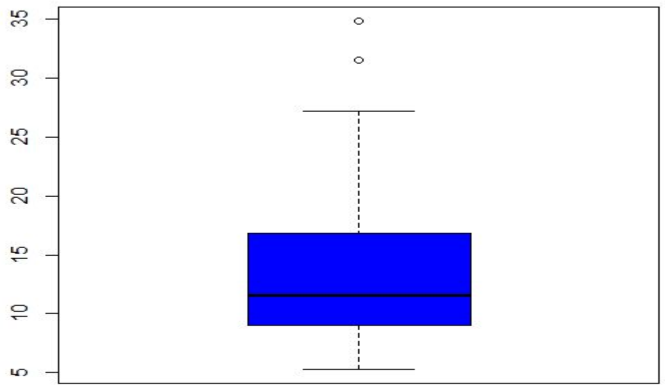
#Measure skewness for both variables

```
library(e1071)
skewness(testdata$csi)
skewness(testdata$billamt)
```

0.0379 for csi



1.3032 for billamt



#The distribution of ‘csi’ appears Normal whereas the distribution of billamt is non-normal(+vely skewed)

DATA SCIENCE INSTITUTE

Quantile-Quantile Plot

- Very powerful graphical method of assessing Normality.

- Quantiles are calculated using sample data and plotted against expected quantiles under Normal distribution.

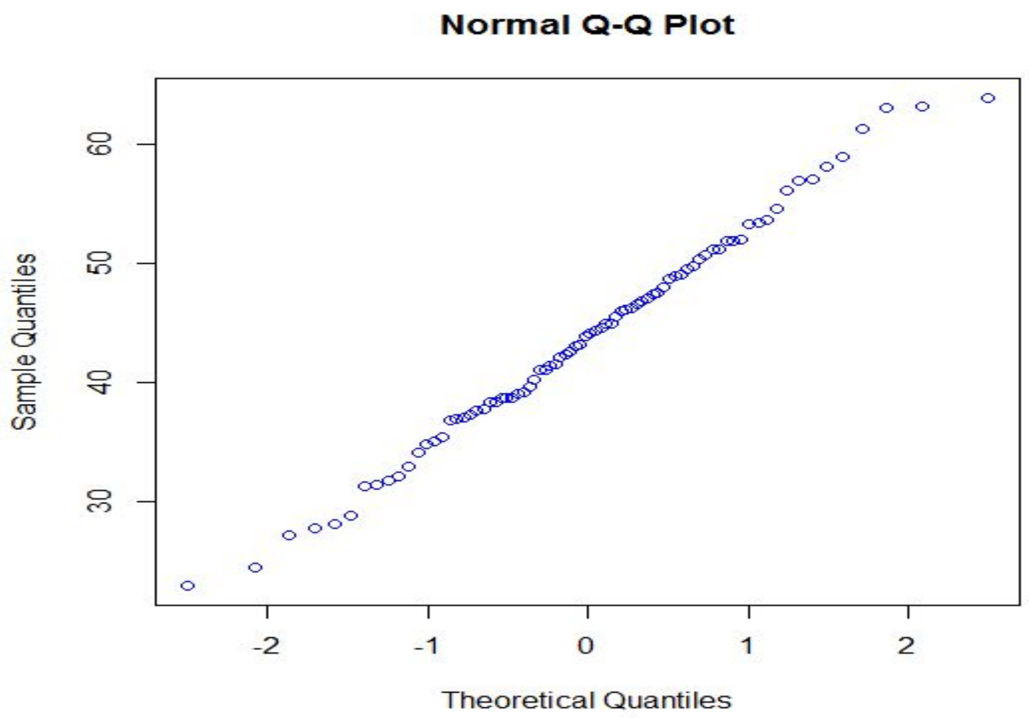
- If Normality assumption is valid then high correlation is expected between sample quantiles and expected(theoretical) quantiles.

- The Y axis plots the actual values. The X axis plots theoretical values.

- If the data are truly sampled from a Gaussian(Normal) distribution, the QQ plot will be linear.

Quantile-Quantile Plot
QQ Plot of CSI

```
qqnorm(testdata$csi,col="blue")
```

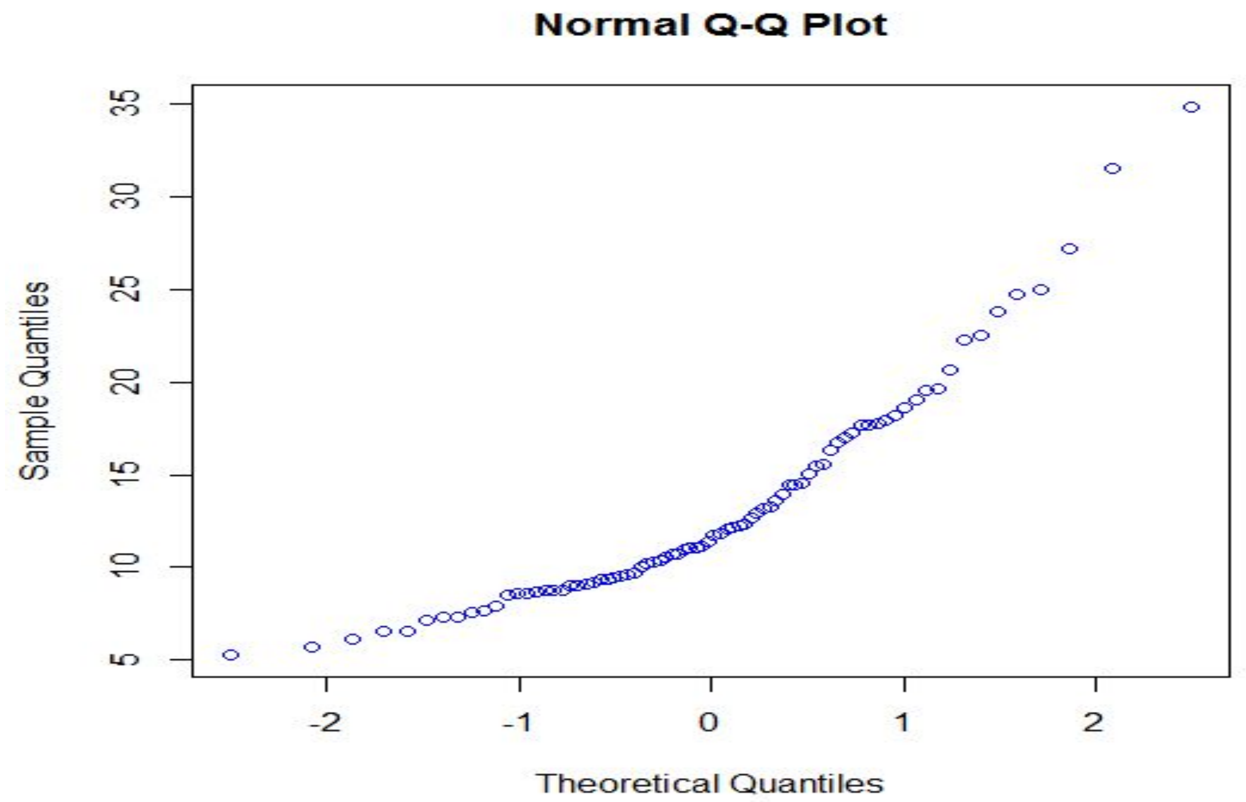


#Check what qqline(csi,col="red") gives

The distribution can be assumed 'Normal'

Quantile-Quantile Plot
QQ Plot of BillAmt

```
qqnorm(testdata$billamt,col="blue")
```



The distribution appears to be non-normal