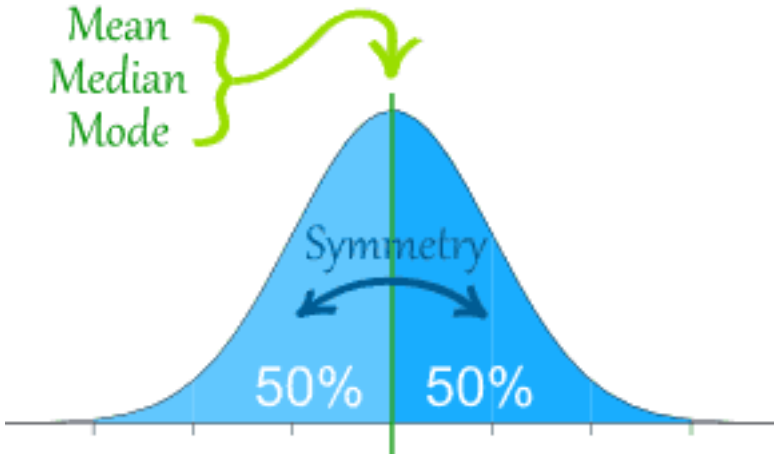
NORMAL DISTRIBUTION

- A function that represents the distribution of random variable as a symmetrical bell-shaped graph is a normal distribution, normal distribution has below key properties
  - The mean, mode and median are all equal
  - The curve is symmetric at the center (i.e. around the mean, μ).
  - The total area under the curve is 1 (or 100%)

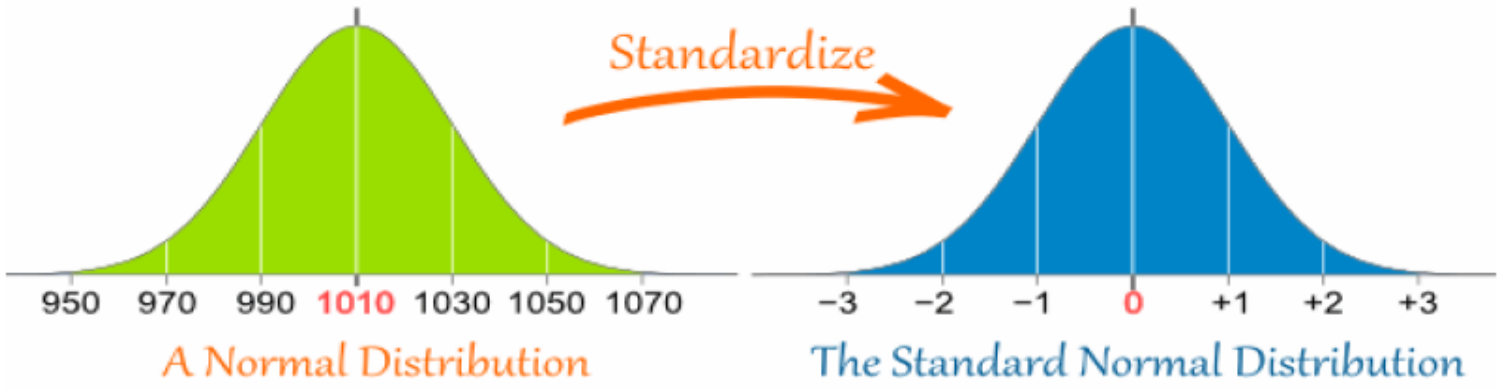


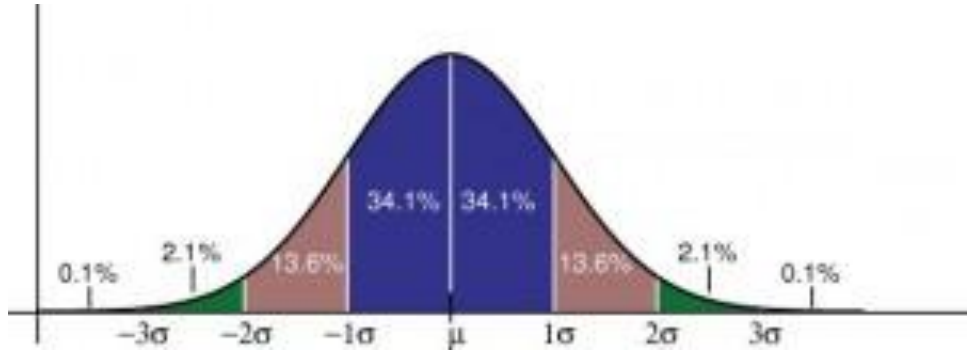
- Many groups (psychological, educational variables) follow this type of pattern. That's why it's widely used in business, statistics and by government bodies
- It is a function of mean and standard deviation of the variable
  - $X \sim N(\mu, \sigma^2)$

STANDARD NORMAL DISTRIBUTION

When we have mean=0 and standard deviation as 1, then normal distribution is termed as standard normal distribution.

$X \sim N(\mu,\sigma^2) \rightarrow Z \sim N(0,1)$



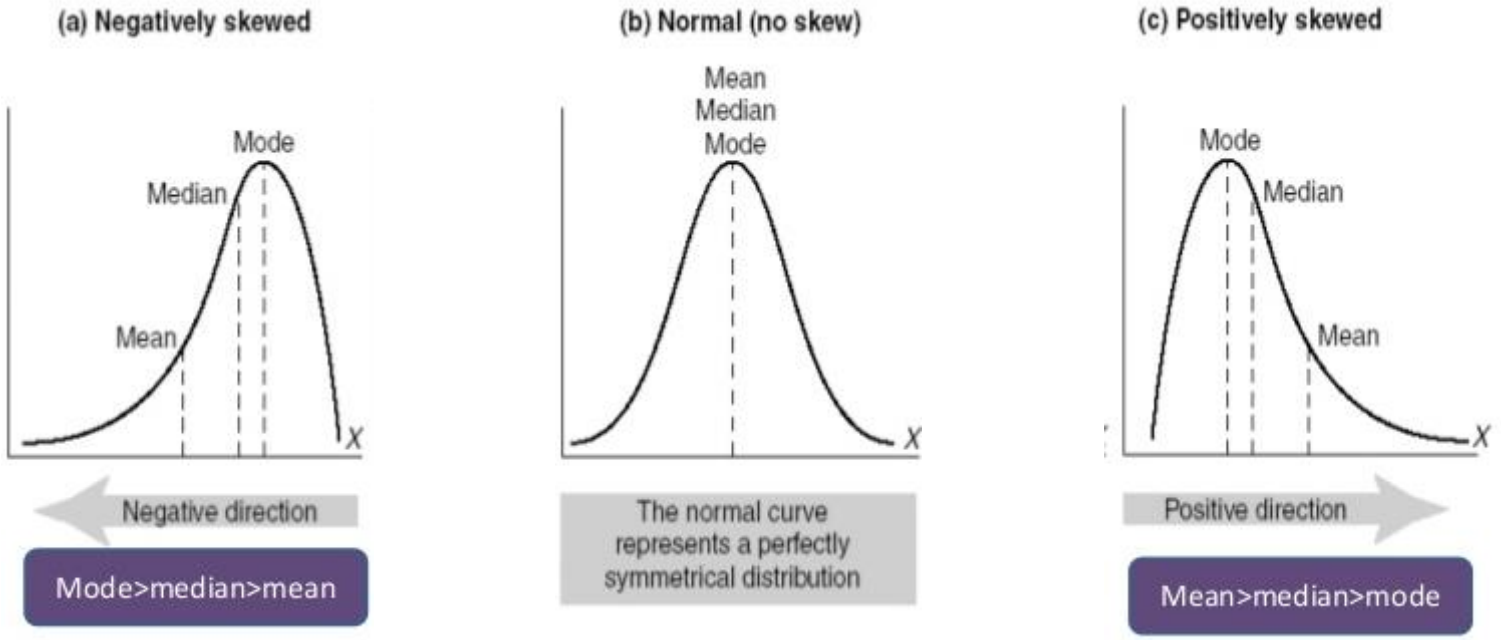


- 68% of the data falls within one standard deviation of the mean
- 95% of the data falls within two standard deviations of the mean
- 99.7% of the data falls within three standard deviations of the mean

SKEWNESS

Skewness is the measure of asymmetry of the distribution.

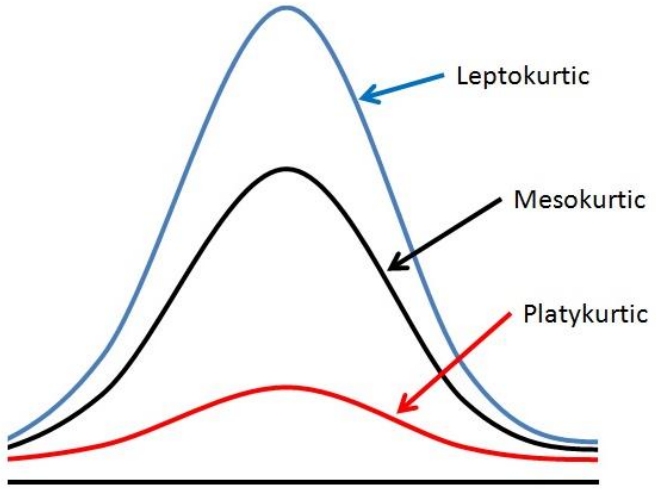
- Negatively skewed indicates a long left tail.
- Positively skewed indicates a long right tail.
- No skew indicates a symmetry around the mean.



KURTOSIS

Kurtosis is the measure of flatness or peakedness of the probability distribution.

- The normal curve is called **Mesokurtic** curve.
- If the curve of a distribution is more peaked than a normal or mesokurtic curve then it is referred to as a **Leptokurtic** curve.
- If a curve is less peaked than a normal curve, it is called as a **Platykurtic** curve.



STANDARD DEVIATION & VARIANCE (MEASURE OF DISPERSION)

- The variance (σ2) of a data set is calculated by taking the arithmetic mean of the squared differences between each value and the mean value or the weighted average of the squared deviations from the mean
- σ = population standard deviation
- σ2 = population variance
- **s** = estimate of population standard deviation based on sampled data
- **s2** = estimate of population variance based on sampled data

The population variance is defined as:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

COVARIANCE

- Covariance is a measure of how much two random variables change together.
- If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, the covariance is positive.
- In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, the covariance is negative.
- The sign of the covariance therefore shows the tendency in the linear relationship between the variables.
- The covariance between two jointly distributed real-valued random variables X and Y is defined as

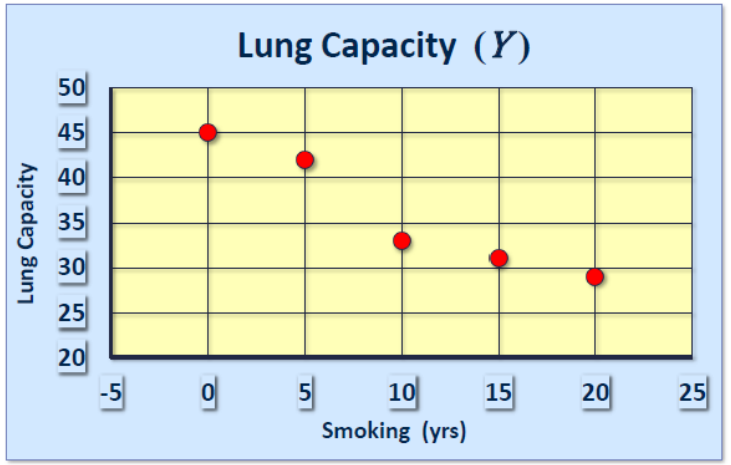
$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Example: investigate relationship between cigarette smoking and lung capacity

Data: sample group response data on smoking habits, and measured lung capacities, respectively

COVARIANCE

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29



Cigs (X)	Lung Cap (Y)
0	45
5	42
10	33
15	31
20	29
10	36

Cigs (X)				Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		$\sum$= -215		

Evaluation yields,

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

CORRELATION COEFFICIENT

- The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.
- The most common correlation coefficient, called the **Pearson product-moment correlation coefficient**, measures the strength of the linear association between variables.
- The population correlation coefficient ρ The population correlation coefficient ρX,Y between two random variables X and Y with expected values μX and μY and standard deviations σX and σY is defined as:

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

HYPOTHESIS TESTING

Hypothesis testing is a statistical way of evaluation of claim/current assumption about characteristic of data.

Null Hypothesis (H0) : Established fact / Current assumption about data; this is what is always tested for validity

** Takes signs : '=', '>=','<='

Examples:

1. Life of an electric fan is greater than or equal to 3000hrs
2. Average diameter of balls manufactured in bearing company is 10mm
3. Average marks obtained by engineering students are less than or equal to 60%

Alternate Hypothesis (H1) : Research hypothesis, challenge to the null hypothesis

** Takes signs : '<>'(not equal to), '>','<'

Types of hypothesis tests based on hypothesis definition:

- If H1 has sign '<' : Lower tailed test
- If H1 has sign '>' : Upper tailed test
- If H1 has sign '<>': Two tailed test

STATISTICAL SIGNIFICANCE LEVELS

Level of confidence (LoC) : How sure/certain are you about the results. Typical values 99%, **95%**, 90%.

- Helps in identification of critical values/range of values in hypothesis testing
- It means that 'we can be LoC% sure that range of values contain true mean of the population'.

Level of Significance (LoS): Error that we are ready to accept while performing a hypothesis test.

LoS = 1-LoC

TEST STATISTIC

Measures difference between observed statistic for sample with its hypothesized population parameter and is calculated from sample data

- z - value: Used when population standard deviation is known and sample size is greater than 30, data is normally distributed

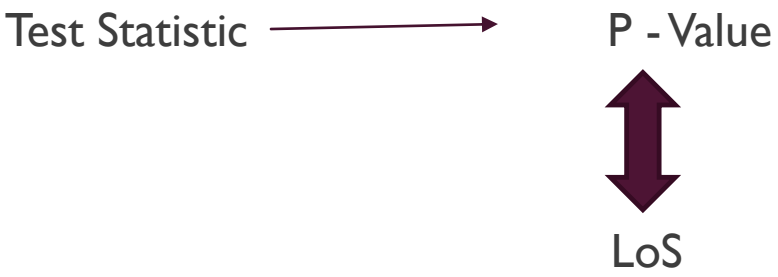
$$z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$$

- t - value: Used when population standard deviation is not known or sample size is less than 30

$$t = (\bar{x} - \mu_0) / (s / \sqrt{n})$$

P - VALUE

P - Value refers to probability associated with the result or extreme values (than actual result).



IF P is Low, then Null will go

- R function for calculating p value for normal distribution : pnorm(test statistic, lower.tail=T/F).. multiply output with 'lower tail=T' by 2 for two tailed test
- R function for calculating p value for t distribution : pt(test statistic, df=n-1)

HYPOTHESIS TESTING – EXAMPLE I

- Suppose the seller says that the mean life of a fan is more than 15,000 hours. In a sample of 40 fans, it was found that they only last 14,900 hours on average. Assume the population standard deviation is 110 hours. At .05 significance level, can we reject the claim by the seller?

The null hypothesis is $\mu \geq 15000$. Begin with computing the test statistic.

```
> xbar = 14900          # sample mean
> mu0 = 15000           # hypothesized value
> sigma = 110          # population standard deviation
> n = 40              # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                  # test statistic
[1] -5.749091
```

```
> pnorm(-5.749091,lower.tail = T)
[1] 4.486226e-09
```

HYPOTHESIS TESTING – EXAMPLE 2

- Suppose the chocolate wrapper states that there is at most 4 grams of saturated fat in a single chocolate. In a sample of 70 chocolates , it is found that the mean amount of saturated fat per chocolate is 4.2 grams. Assume that the population standard deviation is 0.50 grams. At .05 significance level, can we reject the claim on wrapper?