

Statistics?

Branch of mathematics dealing with

- **Data collection**
- **Organization**
- **Analysis**
- **Interpretation and presentation.**



Copyright 2015 Business Over Broadway

Statistically you can analyze the data using two methods,

Descriptive statistics: Summarizing data from a sample using indexes such as the mean or standard deviation

Inferential statistics : Drawing conclusions from data that are subjected to random variation (e.g., observational errors, sampling variation).

Descriptive Statistics

- **Measures of central tendency** – mean, median, mode
- **Measures of dispersion** – range, variance, standard deviation
- **Measures of shape** – skewness, kurtosis



Inferential Statistics

Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates.

Data Series & Dataset

- **Data Series:** A row or column of numbers that are plotted in a chart is called a data series.

19,4,33,2,51,32,2,41,18,2,4,1

- **Dataset :** A collection of all related sets of information that is composed of separate elements but can be manipulated as a unit by a computer

Air quality dataset

Serial no.	Solar Radiation	Wind	Temp	Month	Day
1	190	7.4	67	5	1
2	118	8	72	5	2
3	149	12.6	74	5	3
4	313	11.5	62	5	4
5	299	8.6	65	5	7
6	99	13.8	59	5	8
7	19	20.1	61	5	9
8	194	8.6	69	5	10
9	256	9.7	69	5	12
10	290	9.2	66	5	13

Mean

Central value of a discrete set of numbers:

- specifically, the sum of the values divided by the number of values.

- $$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

- It is the value separating the higher half from the lower half of a data sample (a population or a probability distribution).

$$\text{Median}(X) = \begin{cases} X_{(r+1)} & \text{If } m \text{ is odd, i.e., } r = (m-1)/2 \\ \frac{1}{2}(X_{(r)} + X_{(r+1)}) & \text{If } m \text{ is even, i.e., } r = m/2 \end{cases}$$

- Median is the middle number in a sorted list of numbers.

Mode

- The mode of a set of data values is the value that appears most often.
- The mode is found by collecting and organizing data in order to count the frequency of each result. The result with the highest number of occurrences is the mode of the set.
- For example, in the following list of numbers, 17 is the mode since it appears more times than any other number in the set:
- 4, 4, 6, 9, **17, 17, 17**, 27, 27, 37, 48,**17**

Data Distribution

- The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur.
- We can describe the below data series as: 18,5,32,2,51,31,2,41,18,2,4,1

“Minimum of 1, Maximum of 51, Average of 17.41.”

- Given this description of the data series, what picture do we form of the data? The easiest way to visualize data is to look at its “**distribution**”.
- In the next slides we will learn more about the distribution.

Measures of Dispersion

- Dispersion is the extent to which a distribution is stretched or squeezed.
- Summary statistics can also be used to understand variation or dispersion in the data.
- A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.

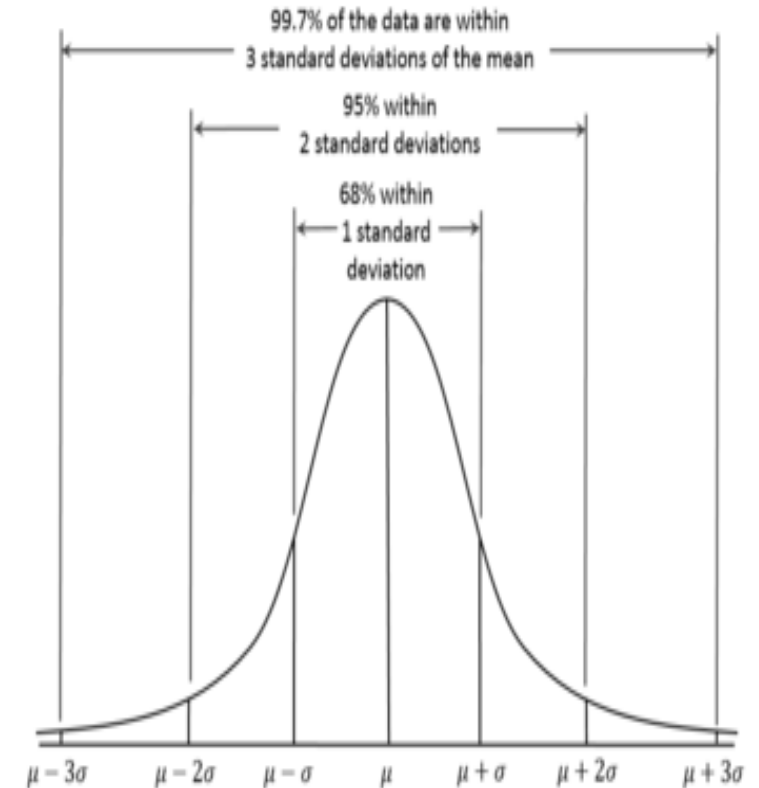
- ✓ Range
- ✓ Inter-Quartile Range
- ✓ Variance
- ✓ Standard Deviation



Normal Distribution/ Gaussian Distribution (Bell Shaped Curve)

- A normal distribution is the distribution in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.
- Features of normal distributions are listed below.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.



Range

Range is the difference between a highest and a lowest observation

Range = Highest observation - lowest observation

Example: In {4, 8, 12, 15, 19, 23, 27, 36, 41 } the lowest value is 4, and the highest is 41

Range: $41 - 4 = 37$

The range can sometimes be misleading when there are extremely high or low values.

Example: In {4, 8, 12, 15, 19, 23, 27, 36, 4100 }

the lowest value is 4,

and the highest is 4100,

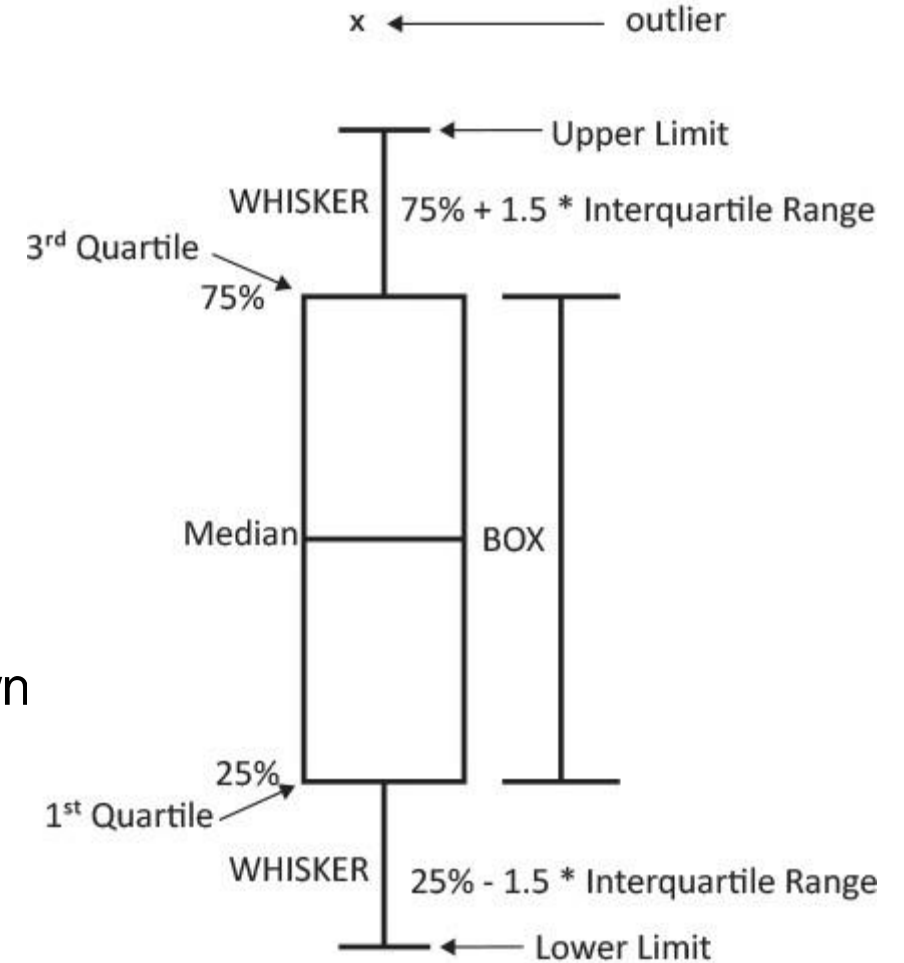
So the range is $4100 - 4 = 4096$

The single value of 4001 makes the range large, but most values are below 100.

So we may be better using Box Plot and Standard Deviation

Box and Whisker Plot

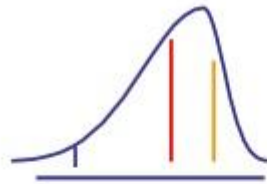
- **Box-and-whisker plots** are a handy way to display data broken into summary statistics by dividing data into four quartiles, each with an equal number of data values. It shows where the middle of the data, maximum and minimum lies.
- The first quartile represents the point where 25% of the data is below it.
- The median is the middle value of the data where half of the points are above and half are below this value.
- The third quartile represents the point where 75% of the data is below it.
- The whisker extends up to the highest value of upper limit and down to the lowest value of the lower limit.
- The lowest point of the lower whisker is called the lower limit. It equals $Q1 - 1.5 * (Q3 - Q1)$ or interquartile range).
- The highest point of the upper whisker is called the upper limit. It equals $Q3 + 1.5 * (Q3 - Q1)$.
- Outliers are points that fall outside the limits of the whiskers.
- The interquartile is represented by the distance between Q1 and Q3.





Distribution Shape & Box-and-Whisker

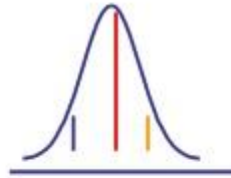
Left-Skewed



Q_1 Q_2 Q_3



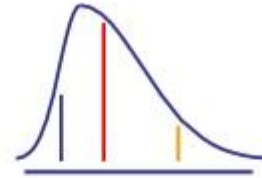
Symmetric



Q_1 Q_2 Q_3



Right-Skewed

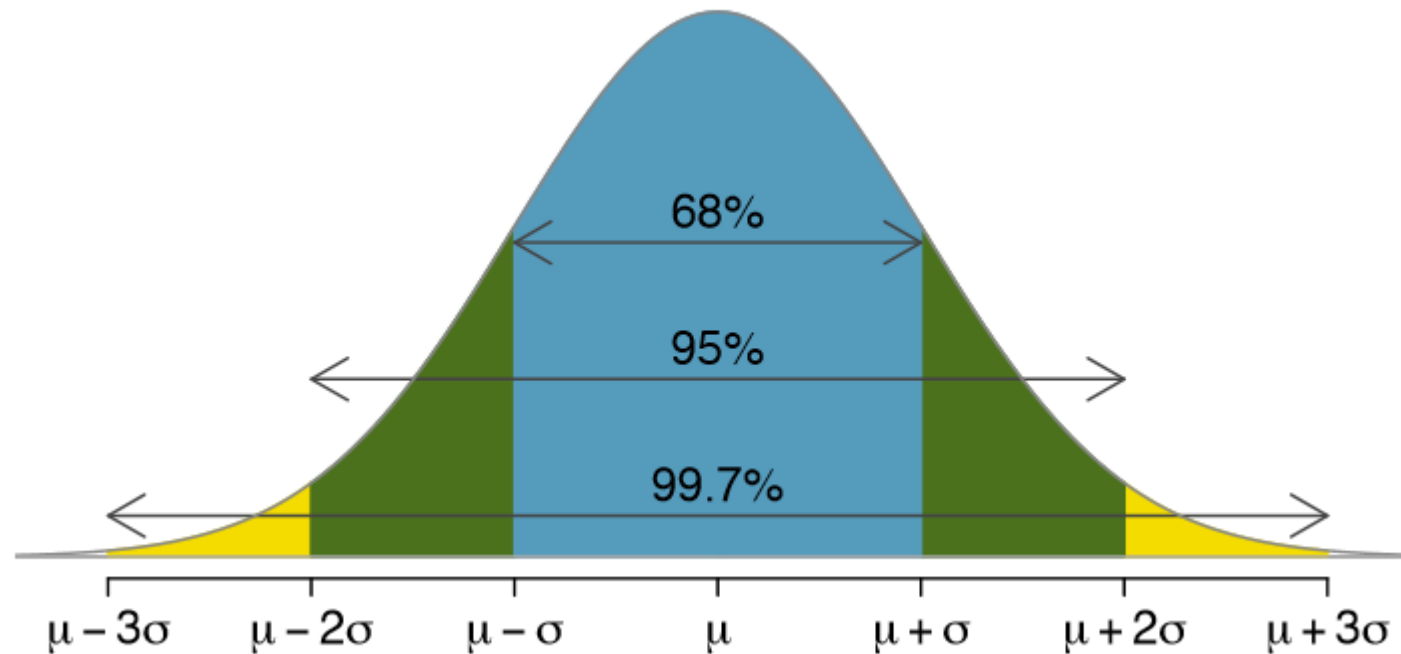


Q_1 Q_2 Q_3



Standard Deviation

- **Standard deviation** is a measure of the dispersion of a set of data from its mean
- **Standard deviation** s (or σ) is just the square root of variance s^2 (or σ^2)
- When we calculate the standard deviation of normal distribution we find that (generally):



KDnuggets

Empirical Rule

EXAMPLE: CAR SALES

Suppose you know that the prices paid for cars are normally distributed with a mean of \$17,000 and a standard deviation of \$500. Use the 68–95–99.7 Rule to find the percentage of buyers who paid

- | | |
|-----------------------------------|-----------------------------------|
| (a) between \$16,500 and \$17,500 | (b) between \$17,500 and \$18,000 |
| (c) between \$16,000 and \$17,000 | (d) between \$16,500 and \$18,000 |
| (e) below \$16,000 | (f) above \$18,500 |

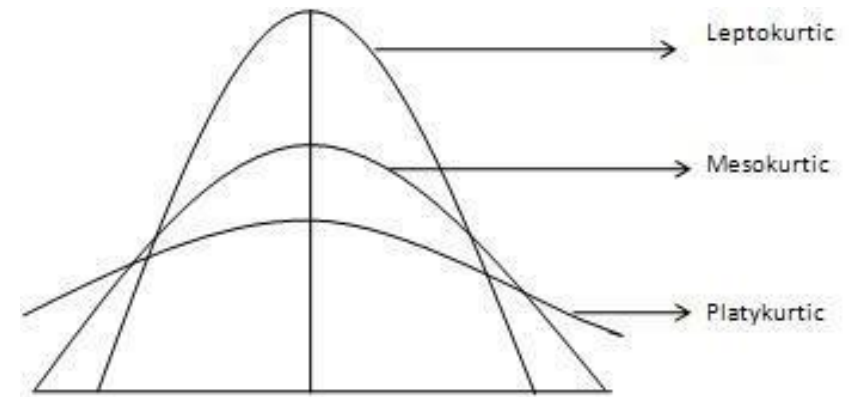
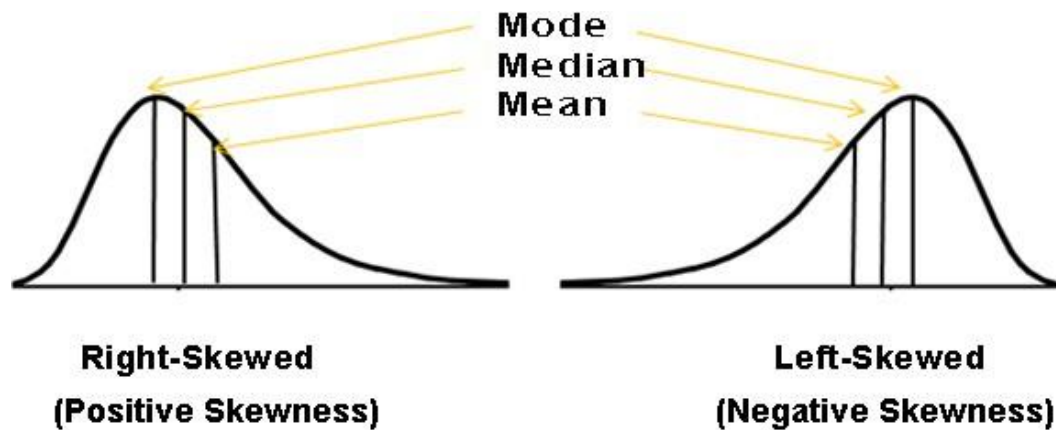


Measures of Shape

- Measures of shape describes the distribution or pattern of the data in a set
- The distribution shape of the quantitative data can be described as there is a logical order to the values and the low and high end values on the horizontal axis of the histogram
- The distribution shape of the qualitative data cannot be described.

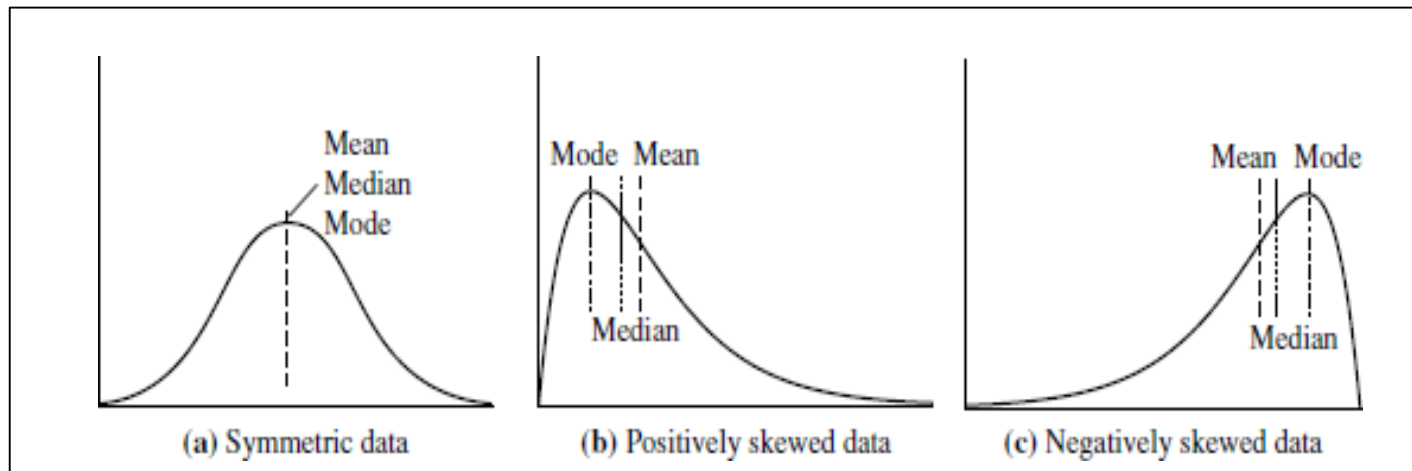
Measures of shape are as follows:

- ✓ Degree of Skewness
- ✓ Kurtosis



Degree of Skewness

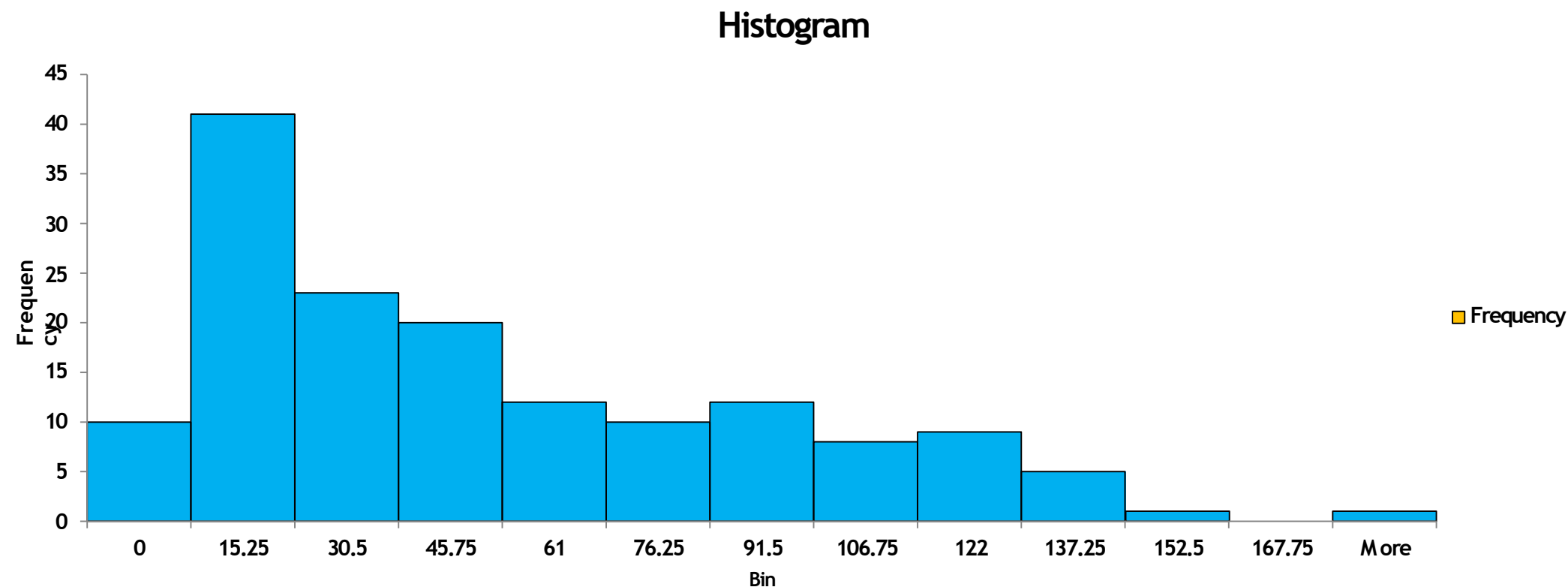
- Skewness is the tendency for the values to be more frequent around the high or low ends of the x axis
- Skewness is a measure of symmetry
- **Symmetric data** - The data is symmetrically distributed on both side of medium
 - ✓ $mean = median = mode$
- **Positively skewed** -
 - ✓ Tail on the right side is longer than the left side.
 - ✓ $mode < median < mean$
- **Negatively skewed** -
 - ✓ Tail on the left side is longer than the right side.
 - ✓ $mode > median > mean$



Skewness Interpretation

- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric, close to Normal Distribution.

Skewness Interpretation



The Skewness is 0.91, Mean is 43.17 and Median is 31 which indicates that the data is Positively skewed.

Kurtosis

- Kurtosis is the sharpness of the peak of a frequency-distribution curve.
- It describes the shape of the distribution of the tail's in relation to its shape

Types of Kurtosis

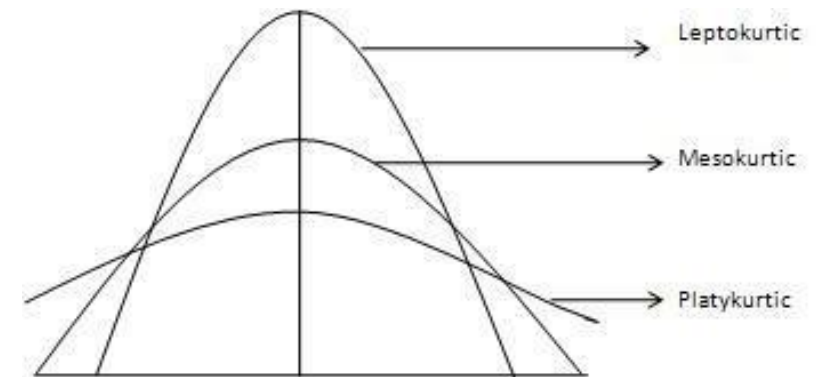
- ✓ Mesokurtic – It has flatter tail than standard normal distribution and slightly lower peak
- ✓ Leptokurtic – It has extremely thick tail and a very thin and tall peak
- ✓ Platykurtic – It has slender tail and a peak that's smaller than Mesokurtic distribution

Kurtosis - Measure of the relative peak of a distribution.

$K=3$ indicates a normal “bell-shaped” distribution (mesokurtic).

$K < 3$ indicates a platykurtic distribution (flatter than a normal distribution with shorter tails).

$K > 3$ indicates a leptokurtic distribution (more peaked than a normal distribution with longer tails).



Population and Sample

Population:

• A **population** is any large collection of objects or individuals, such as Working professionals, students, or house makers about which information is desired.

Example:

- ✓ Collection of items
- ✓ A group of people suffering from a particular disease,
- ✓ Collection of books,

Sample:

• Sample is the representative unit of the target population, which is worked upon by the researchers



Image credits : Towards DataScience

Advantages and Disadvantages of sampling

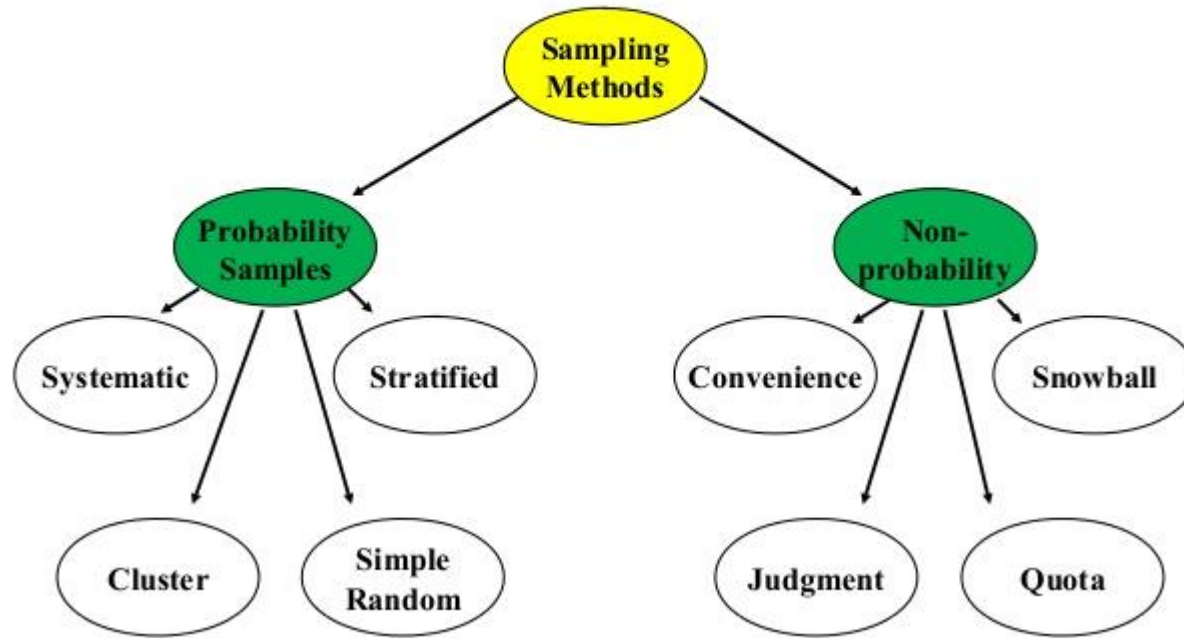
Advantages of Sampling

- ✓ Low cost
- ✓ Less time consuming
- ✓ Suitable in limited resources

Disadvantages of Sampling

- ✓ Difficult to select a truly representative sample
- ✓ It is important to have subject specific knowledge
- ✓ Chances of bias
- ✓ Sampling is impossible when population is too small and heterogeneous

Classification of Sampling Methods



Importance of Probability

- Probability is the measure of the likelihood that an event will occur
- Probability trains you to make decisions in situations which there are observable patterns, but a degree of uncertainty. Uncertainty and randomness occur in just about every field of application and in daily life for example probably the price of X share will go up, probably 'X' team will win the match , so it is extremely useful and interesting to understand probability.

Conditional Probability

- In probability theory, **conditional probability** is a measure of the probability of an event (some particular situation occurring) given that (by assumption, presumption, assertion or evidence) another event has occurred

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where:

- $P(A|B)$ = Conditional probability that event A will occur given that event B has occurred already
- $P(A \cap B)$ = Unconditional probability that event A and event B both occur
- $P(B)$ = Probability that event B occurs

Conditional Probability

Example 1: A Company evaluates the performance of an employee with two tests. 25% of the employees passed both tests and 42% of the employees passed the first test. What percent of those who passed the first test also passed the second test?

Solution:

$$P(\text{Second}|\text{First}) = \frac{P(\text{First and Second})}{P(\text{First})} = \frac{0.25}{0.42} = 0.60 = 60\%$$

Bayes theorem

In probability theory and statistics, **Bayes' theorem** (alternatively **Bayes' law** or **Bayes' rule**) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Binomial Distribution

- The number of successes x in n repeated trials of a binomial experiment is called binomial random variable
 - ✓ Toss a coin it has only two outcome i.e. Head or Tail
 - ✓ Result of an exam has two outcomes, pass or fail
- The **binomial distribution** is a common discrete **distribution** used in statistics, as opposed to a continuous **distribution** such as the normal **distribution**. This is because the **binomial distribution** only counts two states, typically represented as 1 (for a success) or 0 (for a failure) given a number of trials in the data
 - The experiment consists of n repeated trials.
 - Each trial can result in just two possible outcomes – Success or Failure
 - The probability of success, denoted by P , is the same on every trial.
 - Independent trials i.e. the outcome on one trial does not affect the outcome on other trials

Formula for Binomial Distribution

- The mathematical formula to calculate these probabilities is called **probability distribution function**.
- **For Binomial Distribution:**

$$\text{PDF} = P(x) = \frac{n!}{x! (n-x)!} p^x q^{n-x}$$

Where,

x = Outcomes

n = Trials

p = probability of success on each trials

Binomial Distribution Example

- If you Toss a coin 5 times, find the probability of getting exactly 2 heads

Solution:

Number_s	2
trials	5
Probability_s	0.5
Cumulative	FALSE

The probability is 31%

In Excel: =BINOM.DIST(2,5,0.50,FALSE)

QUARTILE		X		✓	f _x	=BINOMDIST(2,5,0.5,FALSE)		
	A	B	C	D	E	F	G	H
1								
2								
3			=BINOMDIST(2,5,0.5,FALSE)					
4			BINOMDIST(number_s, trials, probability_s, cumulative)					
5								

Binomial Distribution Example

Number_s	2
trials	5
Probability_s	0.5
Cumulative	TRUE

- If you toss the same coin 5 times, find the probability of getting upto 2 heads.
- When you have to find the values upto two heads which means: probability of getting 1 head & 2 head, you will write “True” in cumulative in excel instead of “false”
- Solution:

The probability is 50%

In Excel: =BINOM.DIST(2,5,0.50,True)

QUARTILE							
				=BINOMDIST(2,5,0.5,TRUE)			
	A	B	C	D	E	F	G
1							
2							
3							
4							
5							

=BINOMDIST(2,5,0.5,TRUE)
BINOMDIST(number_s, trials, probability_s, cumulative)

Binomial Distribution Example

- If you toss the coin 5 times, find the probability of getting more than 2 heads.

Solution:

Number_s	2
trails	5
Probability_s	0.5
Cumulative	TRUE

The probability is 50%

In Excel: =1-BINOM.DIST(2,5,0.50,true)

[illegible]

Binomial Distribution Example

- An online company delivers 3.4% defective goods to its company . A sample of 10 deliveries is taken , What is the probability that the sample contains exactly 2 defective parts?

Solution:

Number_s	2
trails	10
Probability_s	0.034
Cumulative	FALSE

The probability that it examines 10 samples and finding 2 defects is 4%

In Excel: =BINOM.DIST(2,10,0.034,FALSE)

[illegible]

Binomial Distribution Cumulative Value

- An online company delivers 3.4% defective goods to its company . A sample of 30 deliveries is taken , What is the probability that the sample contains upto 2 defective parts?

Solution:

Number_s	2
trials	30
Probability_s	0.034
Cumulative	TRUE

The probability that it examines 30 samples and finding upto 2 defects is 91.92%
In Excel: =BINOM.DIST(2,30,0.034,True)

NORM.DIST ▾		✕ ✓ <i>fx</i>		=BINOM.DIST(2,30,0.034,TRUE				
	A	B	C	D	E	F	G	H
1	=BINOM.DIST(2,30,0.034,TRUE							
2	BINOM.DIST(number_s, trials, probability_s, cumulative)							
3								

Negative Binomial Distribution Example

- This type of distribution concerns the number of trials that must occur in order to have a predetermined number of successes or in other words it is concerns with the number of trials X that must occur until we have r successes.
- For example: "What is the probability that we get three heads in the first X coin flips?"
- An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. What is the probability that the first strike comes on the third well drilled?

In our Example:

number_f(no. of failure)	2
number_s(no. of successes)	1
probability_s	0.2
Cumulative	FALSE

In Excel: `NEGBINOM.DIST(2,1,0.2,false)` = 0.128

<code>=NEGBINOM.DIST(</code>					
<code>NEGBINOM.DIST(number_f, number_s, probability_s, cumulative)</code>					

<code>=NEGBINOM.DIST(2,1,0.2, FALSE</code>					
<code>NEGBINOM.DIST(number_f, number_s, probability_s, cumulative)</code>					

Geometric Distribution

- Geometric distribution is a special case of negative binomial distribution where number of successes(r) is equal to 1
- The experiment consists of a sequence of trials with the following conditions:
 - ✓ The trials are independent.
 - ✓ Each trial can result in one of two possible outcomes, success and failure.
 - ✓ The probability of success is the same for all trials.

Geometric Distribution

- Example – In a country 10% of the people evade legitimate taxes. What is the probability that a tax official will need to raid at most 20 people, before finding a tax evader (first tax evader)

Solution:

number_f	19
number_s	1
probability_s	0.1
Cumulative	FALSE

The probability that the tax official will need to raid at most 20 people is 1%.

	A	B	C	D	E	F
1	=NEGBINOM.DIST(
2	NEGBINOM.DIST(number_f, number_s, probability_s, cumulative)					
3						

Poisson Distribution

- A statistical distribution showing the frequency probability of specific events when the average probability of a single occurrence is known. The Poisson distribution is a discrete function.
- Example: On an average, 12 people visit a restaurant in one hour, what is the probability that 15 people may visit in next one hour.
- **Properties of Poisson Experiments :**
 - ✓ The experiment results classified as successes or failures.
 - ✓ Average number of successes that occurs in a specified region is known.
 - ✓ Probability that a success will occur is proportional to the size of the region.
 - ✓ Probability that a success will occur in an extremely small region is virtually zero.
 - ✓ Events have to be counted as a whole number

Poisson Distribution

- Poisson Probability can be calculated as :

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

Where,

Lamda is a mean number of occurrences in a given interval of time

Applications of Poisson Distribution

- Manufacturing
- Operations and Supply chain
- Insurance

Poisson Distribution Example

- On an average, 12 people visit a restaurant in one hour, what is the probability that 15 people may visit in next one hour.

Solution:

x	15
mean	12
Cumulative	FALSE

- The probability that 15 people may visit in next one hour is 7%.

=POISSON.DIST(15,12,FALSE			
D	E	F	G
=POISSON.DIST(15,12,FALSE			
POISSON.DIST(x, mean, cumulative)			

Poisson Distribution Example

- A burger shop has a staff of 25 workers, which deliver 175 burgers a day. A long weekend is coming up and 5 of the workers have asked for a holiday. You estimate remaining 20 workers can manage 15% greater delivery but want to plan for the chance of greater than 25% increase of delivery.

Solution: $175/25 = 7$ delivery a day

If 15% greater delivery is received with 5 less resources = $(175 \times 1.15)/20 = 10.6 = 11$

- We need a probability that if there is a requirement of 11 or more burgers delivered in a day when the average is 7.

QUARTILE							
	A	B	C	D	E	F	G
1	=1-POISSON(11,7,TRUE)						
2	POISSON(x, mean, cumulative)						
3							
4							
5							

Poisson or Binomial Distribution?

Poisson Distribution

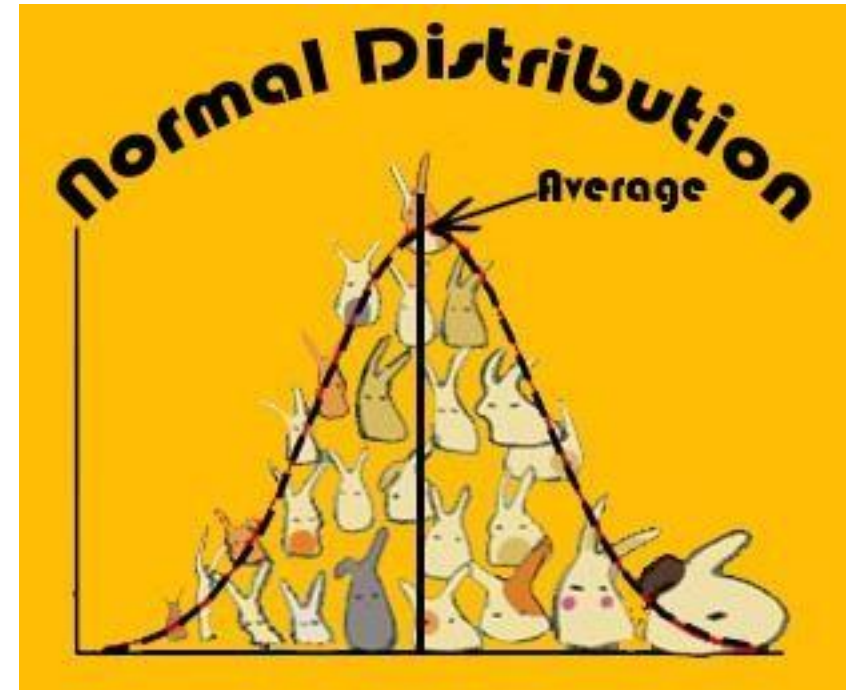
- A Poisson Distribution is used, If a mean / average probability of an event happening per unit time/ per page/per mile cycled etc., is given, and you are asked to calculate a probability of n events happening in a given time / number of pages / number of miles cycled.
- It describes the distribution of binary data from a infinite sample. Thus, it gives the probability of getting r events in a population

Binomial Distribution

- The Binomial Distribution is used when an exact probability of an event happening is given or implied, in the question and you are asked to calculate the probability of this event happening k times out of n .
- It describes the distribution of binary data from a finite sample. Thus, it gives the probability of getting r events out of n trials

Continuous Probability Distribution

- The probabilities of the possible values of a continuous random variable is a continuous distribution.
- A continuous random variable is a random variable with a set of possible values i.e. infinite and uncountable. For example: Height of women in Pune : 60 inch, 60.5 inch, 70.1 inch and soon.
- Normal Distribution is the most common kind of a continuous probability distribution due to its applications in statistics.
- Types of Continuous Probability Distribution
 - ✓ Normal Probability Distribution
 - ✓ Standard Normal Probability Distribution

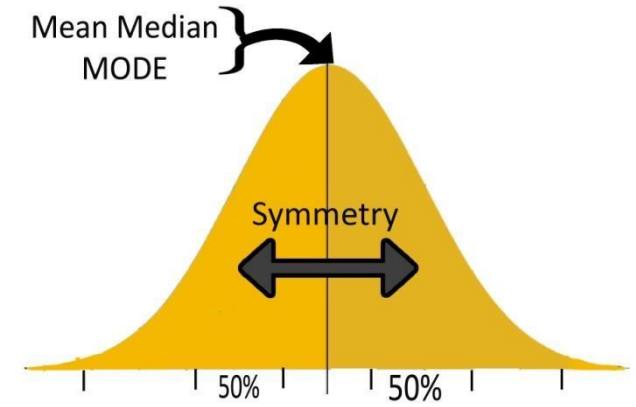


Normal Probability Distribution

- There are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" this is also known as Bell Curve.
- A normal distribution is a very important statistical data distribution pattern occurring in many natural phenomena.
- For example, the bell curve is seen in Exam Results. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an A or F. This creates a distribution that resembles a bell (hence the nickname).
- In Corporate also, many of the times, HR uses Bell curve to evaluate the performance of the candidates.
- **Examples of Normal Distribution**
 - ✓ Heights of people
 - ✓ Size of things produced by machines
 - ✓ Errors in measurements
 - ✓ Blood pressure
 - ✓ Marks on a test
- Note: You can refer Normal Distribution, Standard Deviation and Empirical rule from chapter.....

Normal Probability Distribution

- This is important to understand if a distribution is normal, there are certain qualities that are consistent and help in quickly understanding the scores within the distribution.
- The Normal Distribution has:
 - ✓ Mean = Median = Mode
 - ✓ Symmetric about the center
 - ✓ 50% of values less than mean and 50% greater than mean



EXAMPLE

A principal at a school claims that the students in his school are above average intelligence.

A random sample of thirty students' IQ scores have a mean score of 112.5. Is there sufficient evidence to support the principal's claim?

The mean population IQ is 100 with a standard deviation of 15.

EXAMPLE

- **Step 1:** State the Null hypothesis.

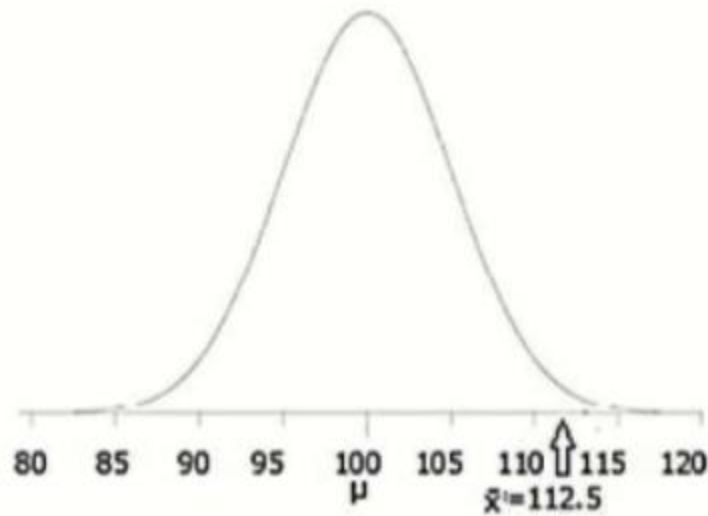
The accepted fact is that the population mean is 100, so: $H_0: \mu=100$.

- **Step 2:** State the Alternate Hypothesis.

The claim is that the students have above average IQ scores, so:
 $H_1: \mu > 100$.

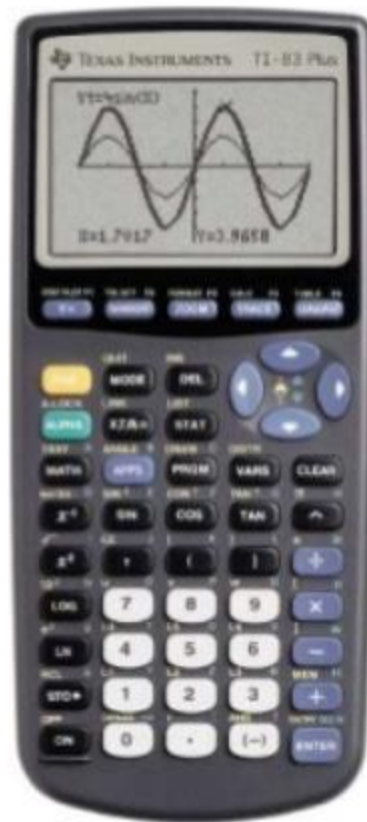
EXAMPLE

- **Step 3:** Draw a picture to help you visualize the problem.



- **Step 4:** State the alpha level. If you aren't given an alpha level, use 0.05, An alpha level of 0.05 is equal to a z-score of 1.645.

To calculate z-score use TI-83 calculator.



- **Step 5:** Find the **Z** using this formula:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

For this set of data:

$$Z = (112.5 - 100) / (15 / \sqrt{30}) = 4.56$$

- **Step 6:** If Step 5 (4.56) is greater than Step 4 (1.645), reject the null hypothesis. If it's less than Step 4, you cannot reject the null hypothesis. In this case, it is greater, so you can reject the null and principal's claim is right.

Two tailed z test

Example:

Suppose it is up to you to determine if a certain state (Michigan) receives a significantly different amount of public school funding (per student) than the USA average. You know that the USA mean public school yearly funding is \$6800 per student per year, with a standard deviation of \$400.

Next, suppose you collect a sample ($n = 1000$) from Michigan and determine that the sample mean for Michigan (per student per year) is \$6873. Use the z-test and the correct H_0 and H_a to run a hypothesis test to determine if Michigan receives a significantly different amount of funding for public school education (per student per year).

Step 3: Calculate the z-test statistic

Now, calculate the test statistic.

$z = (\text{sample mean} - \text{population mean}) / [\text{population standard deviation}/\sqrt{n}]$

$z = (6873 - 6800) / [400/\sqrt{100}]$

$z = 73 / [400/10]$

$z = 73 / [40]$

$z = 1.825$

Example

Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive effect on blood glucose levels. A sample of 36 patients who have tried the raw cornstarch diet have a mean glucose level of 108. Test the hypothesis that the raw cornstarch had an effect or not.

Solution:- Follow the above discussed steps to test this hypothesis:

Step-1: State the hypotheses. The population mean is 100.

H0: $\mu = 100$

H1: $\mu > 100$

Step-2: Set up the significance level. It is not given in the problem so let's assume it as 5% (0.05).

Step-3: Compute the random chance probability using z score and z-table.

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

- For this set of data: $z = (108 - 100) / (15 / \sqrt{36}) = 3.20$
- You can look at the probability by looking at z- table and p-value associated with 3.20 is 0.9993 i.e. probability of having value less than 108 is 0.9993 and more than or equals to 108 is $(1 - 0.9993) = 0.0007$.
- Step-4: It is less than 0.05 so we will reject the Null hypothesis i.e. there is raw cornstarch effect.
- **Note:** Setting significance level can also be done using z-value known as critical value. Find out the z- value of 5% probability and it is 1.65 (positive or negative, in any direction). Now we can compare calculated z-value with critical value to make a decision.

Example

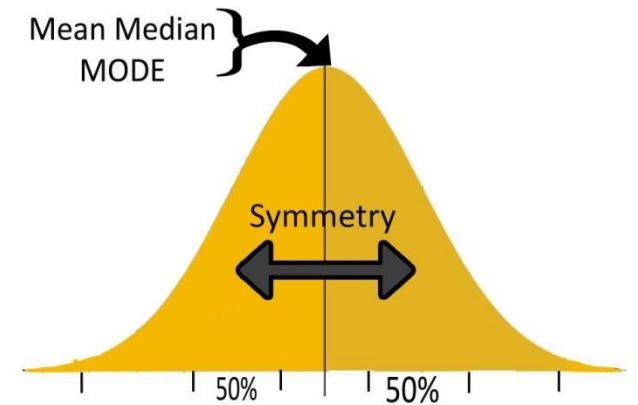
- A company has 500 employees, salary of whom is normally distributed, with an average of Rs.40,000 and Standard deviation of Rs.6000. Suppose you pick a random employee from the 500 employees, what are chances he/she earns less than Rs.30,000
- The following information is available:

Distribution: Normally distributed

Mean: 40,000

SD: Rs.6000

Make a bell
curve
showing
mean and
.....



Standard Scores or Z score

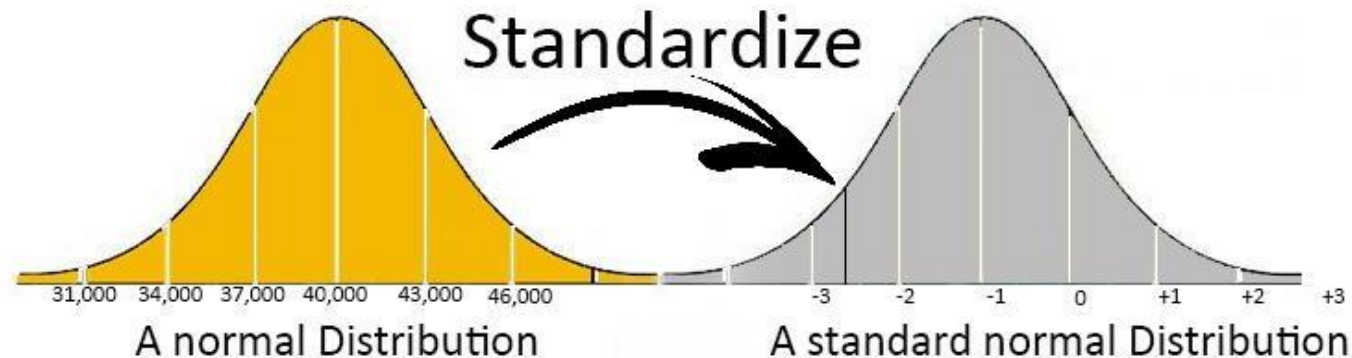
To find out the answer of previous questions, first of all we need to understand the standard scores or Z score

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score". Get used to those words!

$$z = \frac{x - \mu}{\sigma}$$

So to convert a value to a Standard Score ("z-score"):

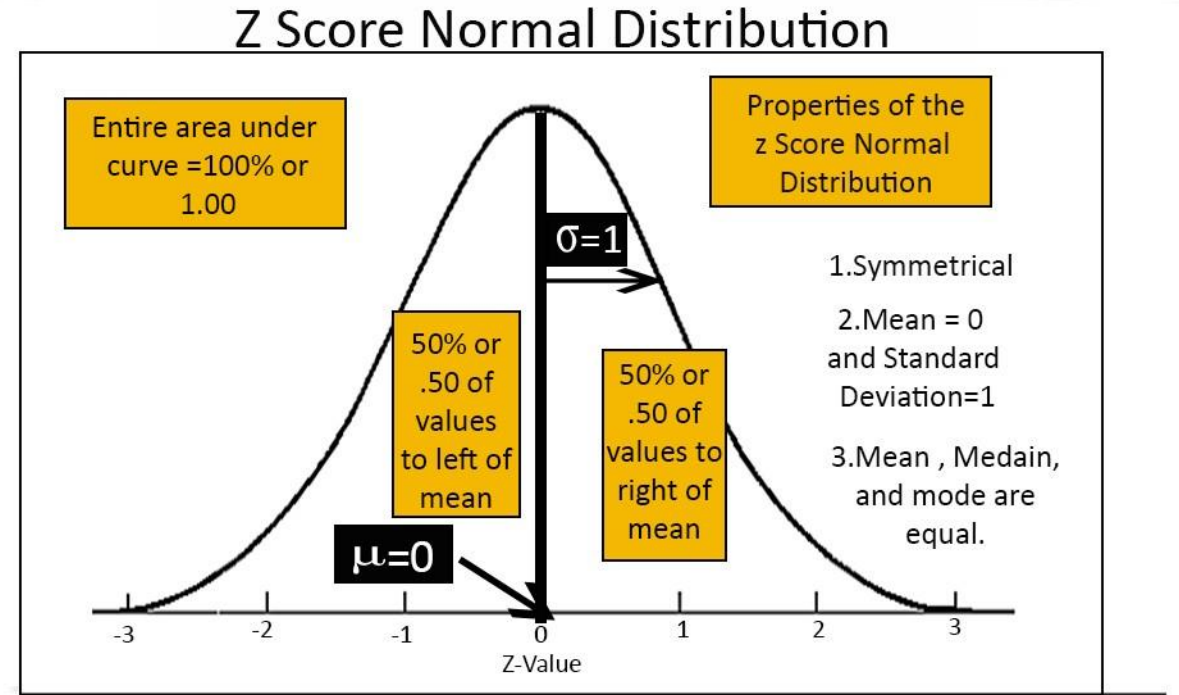
- first subtract the observation from the mean: $30,000 - 40,000 = -10,000$
- then divide by the Standard Deviation: $-10,000/6000 = -1.66$
- And doing that is called "Standardizing":
- We can take any Normal Distribution and convert it to The Standard Normal Distribution.



Z test

The area under the whole of a normal distribution curve is 1, or 100 percent. The z-table helps by telling us what percentage is under the curve at any particular point.

Now we need to use Ztable to find the what percentage of is under the curve at any particular time



Number in the table represents $P(Z \leq z)$

Number in the table represents $P(Z \leq z)$

[illegible]

Z test

Since the z score was negative (-1.66), we need to use the negative Z- Scores

From the table, we can find out the probabilities of z score at -1.66

= 0.0485 Or 4.85%

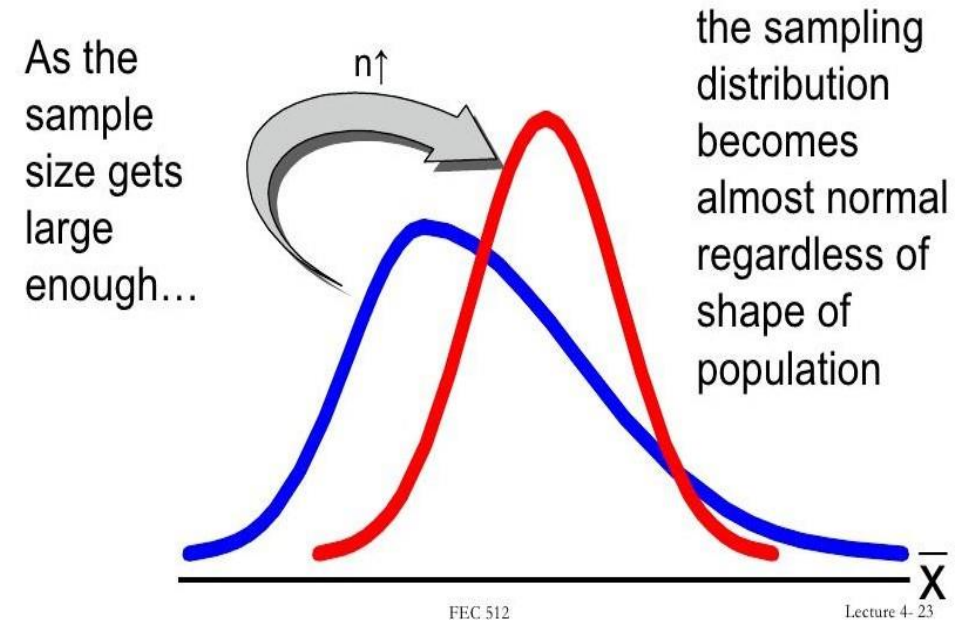
It means that when we pick a random employee from the 500 employees, the chances he/she earns less than Rs.30,000 is 4.85%

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867

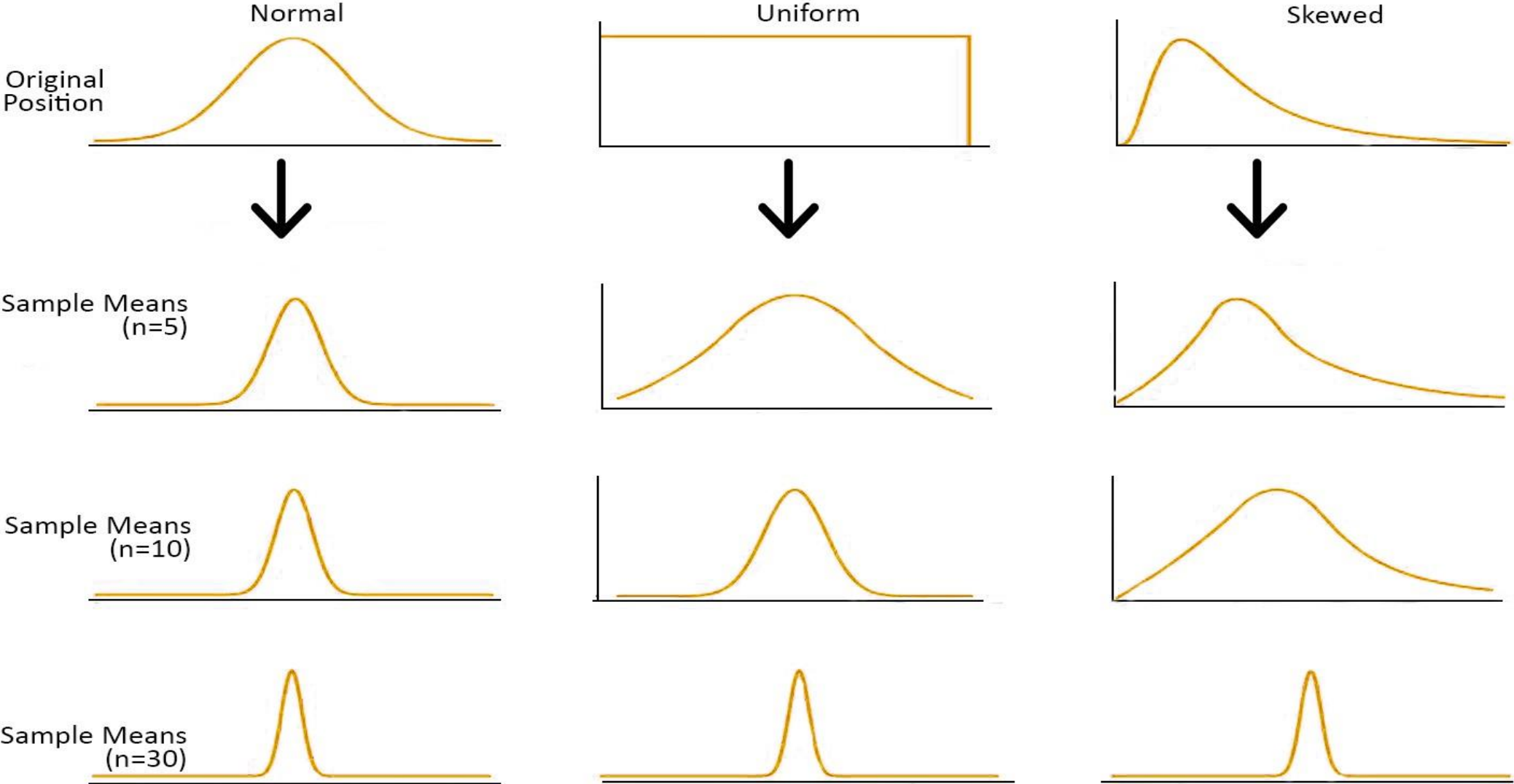
Central Limit Theorem

- The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger - no matter what the shape of the population distribution.
- This fact holds especially true for sample sizes over 30.
- All this is saying is that as you take more samples, especially large ones, your graph of the sample means will look more like a normal distribution.
- The standard deviation of the sample means will be smaller than the population standard deviation and will be equal to the standard deviation of the population divided by the square root of the sample size.

Central Limit Theorem



Shapes of Distributions as Sample Size Increases



Introduction to Hypothesis

Example 1: You are an Analytics advisor of an e-commerce company and find from the population that the average daily sales is 20,000 USD (approx.) and Standard Deviation is USD4000.

The Sales team picks up a random sample of 25 days and claims that the average daily sales is 19,800 USD (Provided that the data is normally distributed) and recommends that company needs to rethink its strategies to increase the sale.

What are the possible observations?

- Reductions in daily sales
- There is no difference between the population and sample, whatever you are seeing is simply “Random by chance?”

Probability of seeing a sample mean of 19,800 or lower if true population mean was 20,000

Function Arguments

NORMDIST

X	19800	= 19800
Mean	20000	= 20000
Standard_dev	$(4000/(25)^{0.5})$	= 800
Cumulative	TRUE	= TRUE

= 0.401293674

Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.401293674

[Help on this function](#)

OK Cancel

The probability is 0.4012 i.e. 40.12%

There is 40% probability of seeing sample mean of 19,800 when the population mean was 20,000

What does it mean?

There is a 40% chance that when you pick a random sample from a population with a mean of 20,000 and you get a sample mean of 19,800 or lower.

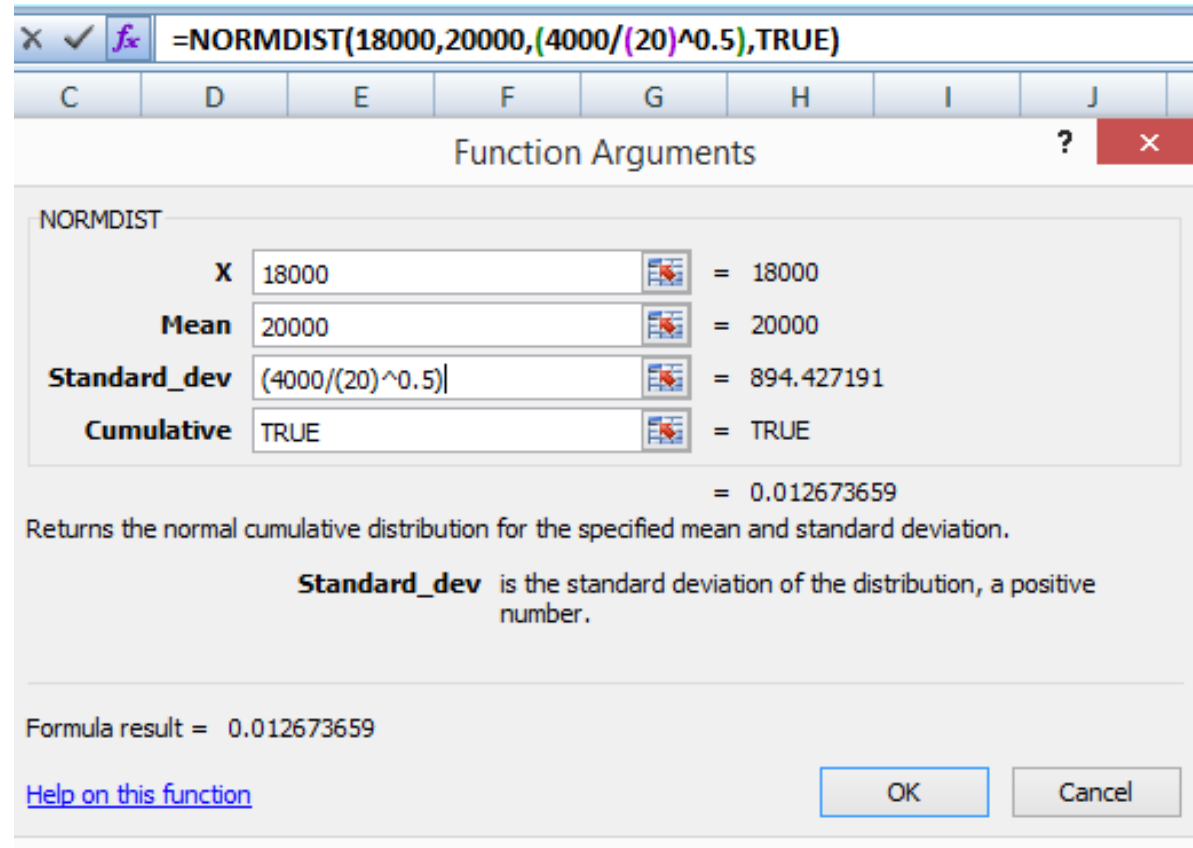
In other words, seeing a 19,800 sale or lower is very likely (40%) because of random by chance and you can still say that population average or sale is Rs.20,000. **Therefore, here we can conclude that there is no change in average daily sales, and we accept the null hypothesis.**

Why we are getting probability of 40% in seeing a sale of 19,800 or lower?

Now assume that population mean of Rs.20,000 we got from below 100 sales figure and we have taken the sample of 20 data points from the below shaded region. The average of 20 data points of shaded region is 19,800.

	1	2	3	4	5	6	7	8	9	10
1	12519	13786	16002	18074	19075	20010	21179	22808	24131	25559
2	12547	13951	16233	18357	19427	20069	21415	22876	24156	25737
3	12721	14133	16266	18542	19500	20170	21535	22934	24312	25971
4	13048	14189	16790	18587	19641	20319	21652	22949	24402	25982
5	13409	15466	17000	18732	19727	20365	21799	22954	24496	26024
6	13453	15600	17070	18752	19794	20392	21909	23431	24793	26298
7	13515	15690	17454	18791	19838	20968	21956	23537	24872	26338
8	13604	15807	17672	18962	19881	21111	21974	23948	25071	26388
9	13753	15820	18300	18992	19924	21116	22545	24006	25198	26490
10	13776	15846	17945	19017	20000	21160	22729	24092	25474	26839

Example 2: Now assume that the average of sample mean from 20 observations is 18,000. Now what is the probability of seeing the average of 18,000 or lower , if the true population mean was 20,000



The image shows the 'Function Arguments' dialog box for the NORMDIST function in Microsoft Excel. The formula bar at the top displays the formula: `=NORMDIST(18000,20000,(4000/(20)^0.5),TRUE)`. The dialog box has a title bar with a question mark and a close button. The main area is titled 'NORMDIST' and contains four input fields with their corresponding values and calculated results:

Argument	Value	Result
X	18000	= 18000
Mean	20000	= 20000
Standard_dev	$(4000/(20)^{0.5})$	= 894.427191
Cumulative	TRUE	= TRUE

Below the input fields, the calculated result is shown: `= 0.012673659`. A description of the function is provided: 'Returns the normal cumulative distribution for the specified mean and standard deviation.' A note explains the 'Standard_dev' argument: 'Standard_dev is the standard deviation of the distribution, a positive number.' At the bottom, the 'Formula result' is displayed as `0.012673659`. There is a link to 'Help on this function' and two buttons: 'OK' and 'Cancel'.

There is probability close to 1.26% of seeing sample mean of 18,000 when the population mean was 20,000

What does it mean?

It means that it very unlikely that if the sample population was 20,000 and your sample mean would be 18,000 or lower simply because of random by chance.

In other words , your sample is more likely to have come from a population with different mean (lower) than the one you are looking at **or we can also conclude that the averages daily sales has reduced.**

One thing we need to notice that at 40% (when the average sale was 19,000) we concluded that high probability of seeing the sample mean to be from the population and when we got the probability of close to 1.26% (less than 5%) , we are concluding a very low probability of sample mean to be from the population.

	Probability	Conclusion
Random chance of seeing different sample mean from population	High (if probability is 40%)	Difficult to conclude that there is difference between sample and population
	Low (if probability is close to 1.26%)	Conclude that sample is different from population
Note: In statistical terms, the probability of occurrence of any event is called p-value, therefore in the above example p-values are 40% (when the sample mean was 19,800) or 1.26% (when the sample mean was 18,000)		

What we will consider the low probability 40% or less than 5%

To avoid subjectivity a cut off of 5% for low probability is commonly used

What does 5% means?

Only if random chance probability of seeing sample means as extreme or more extreme that is observed is less than 5%, we will conclude that the average daily sales has reduced (or sample is different from the population).

Probability	Sample Mean	Conclusion
40%	19,800	No reduction in sales
1.26, i.e. less than 5 %	18,000	Reduction in sales

In our example, when the sample mean was 19,800 we will conclude that there is no reduction in sales and whatever variations we are observing is just because of random chance and when the sample mean was 18,000, we will conclude that sales has reduced.

What is hypothesis?

- Hypothesis testing is a well defined procedure which helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample.
- There are two types of Hypothesis:
 - ✓ Null Hypothesis (H_0) : Null hypothesis is the hypothesis which is tested for the possible rejection under the assumption that is true
 - ✓ Alternative Hypothesis (H_a) : Alternative hypothesis is the logical opposite of the null hypothesis
- **In the above example:**
 - ✓ H_0 : No reduction in daily sales
 - ✓ H_a : Reduction in daily sales

Basic Statistics P-Value

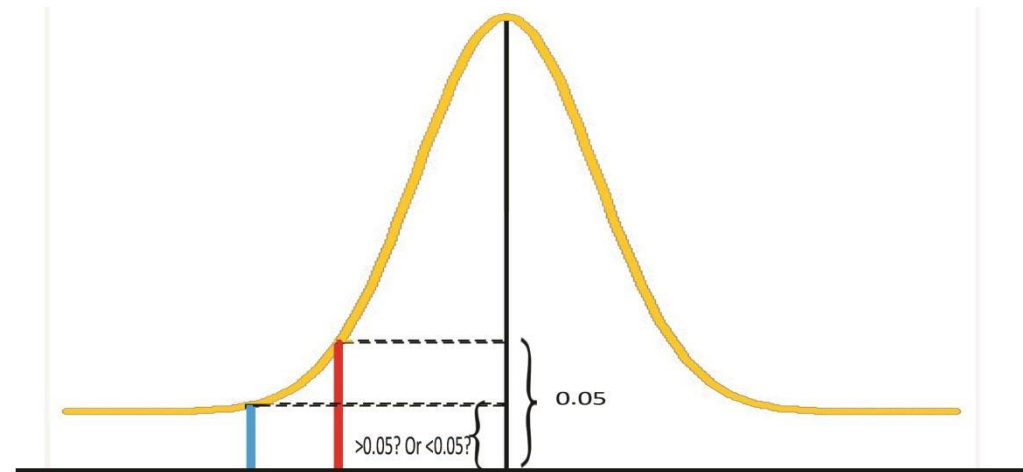
Significance Level (alpha : α): Criterion used for rejecting the null hypothesis.

Red line in figure

P-value: The probability of outcomes more extreme than the observed outcome, assuming the null is true.

Area to the left of the blue line

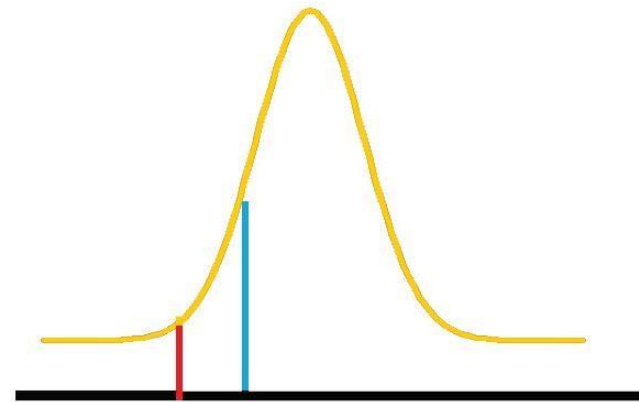
Conclusion: If p-value is less than significance level reject null hypothesis.



More on p-value

- The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event.
- p-value also known as rejection region
- The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.
- A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.
- The rejection region is found by using alpha to find a critical value;
- The rejection region is the area that is more extreme than the critical value.

What if $p\text{-value} > \text{Significance level}$?



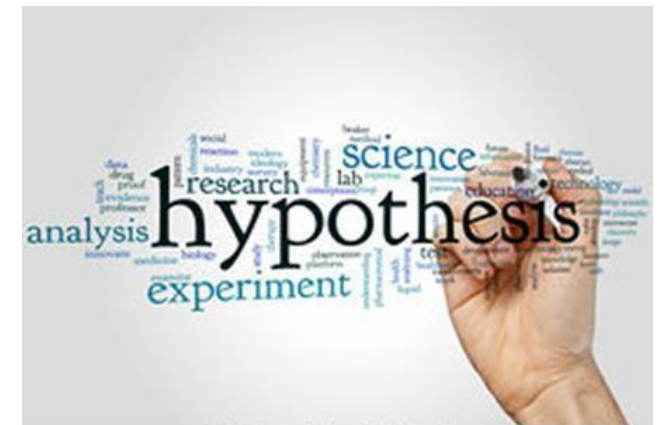
Conclusion?

Conclusion

Basic Setup in Hypothesis Test	When sales is 19,800	When sales is 18,000
Null Hypothesis	Reduction in daily sales	Reduction in daily sales
Alternate Hypothesis	No reduction in daily sales	No reduction in daily sales
Test Distribution	Normal Distribution	Normal Distribution
Significance Level	5%	5%
p-value	0.401	0.012
Conclusion	Accept Null Hypothesis	Reject Null Hypothesis

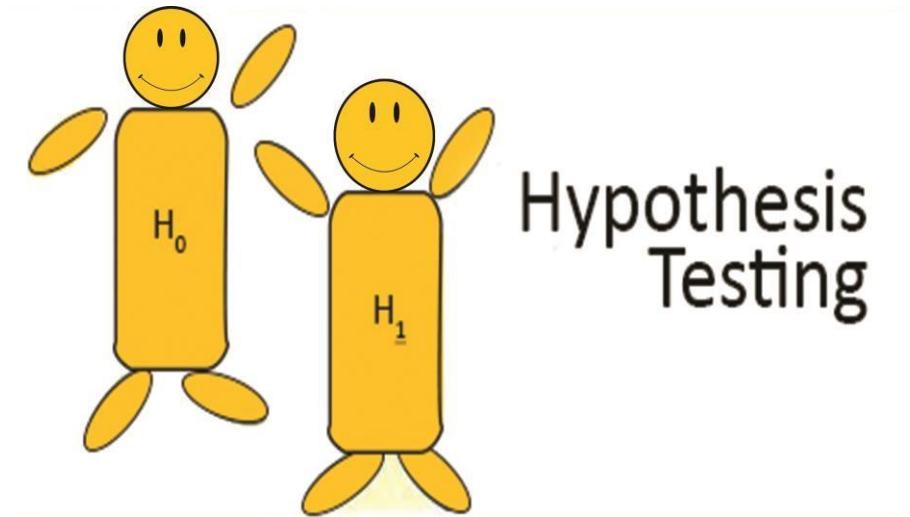
Criteria for hypothesis construction

- Hypothesis should be testable like it is right or wrong
- **Considering the previous example:** Is there a reduction in the daily sales or not?
- It should be specifically and precisely described
- **Considering the previous example:** We see a reduction in the Sample when compared to the population, a precise statement must be described.
- Statements in test should not be contradictory
- **Considering the previous example:** We considered only sales as our variables. It should describe one issue at a time
- **Considering the previous example:** We considered only the reduction of the sales to test the hypothesis.
- It should only specify variables between which relationship to be established



Null Hypothesis

- A null hypothesis proposes that no statistical significance exists in a set of given observations.
- The null hypothesis shows that no variation exists between variables or that a single variable is no different than its mean.
- The hypothesized value of the parameter is called the null hypothesis.
- It is denoted by H_0 symbol
- It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.



Alternative Hypothesis

- The alternative hypothesis is a statement that will be accepted as a result of the null hypothesis being rejected.
- The alternative hypothesis is usually denoted " H_a ".
- The alternative hypothesis is the residual of the null hypothesis
- Alternative Hypothesis includes the outcomes not covered by the null hypothesis.

Examples of hypothesis

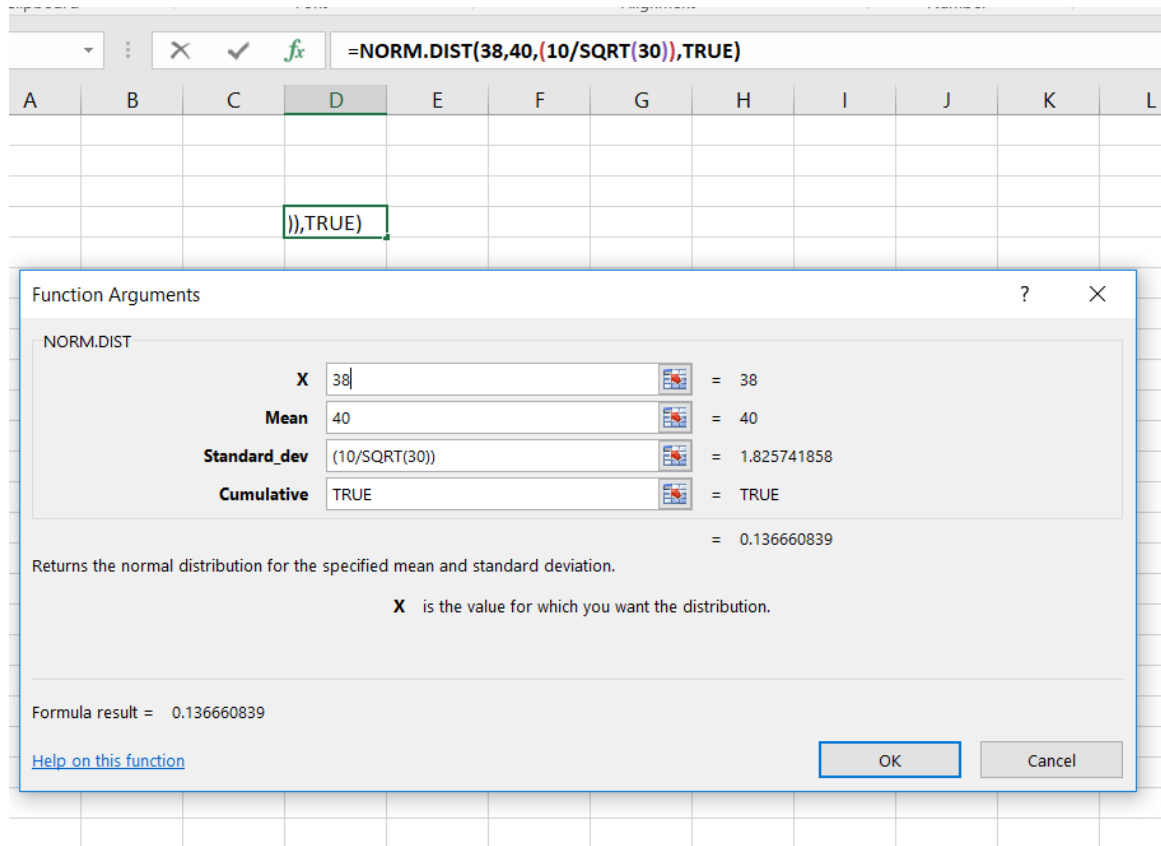
- A researchers investigates that the mean body temperature of people who is 17 year old is greater than 98.6 degrees. What will be the null and alternate hypothesis in this case?
Suppose x = Mean body temperature
 $H_0: x = 98.6$
 $H_a: x > 98.6$
- A pharmaceutical company claims that a new treatment is successful in reducing heart attack in more than 65% of the cases. The treatment was tried on 50 randomly selected cases and 15 were successful.
 $H_0: x = 6.5$
 $H_a: x < 6.5$
- A radio station publicizes that the average number of the local listening audience is greater than 45%
 $H_0: x = 0.45$
 $H_a: x < 0.45$

Another example

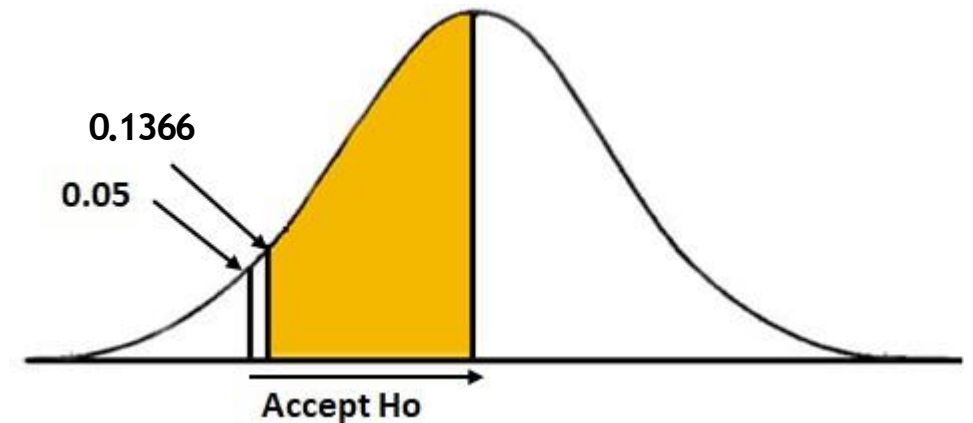
- Indian cricket selection committee has a selection meeting and is confused on one player over the selection in the team. This particular player has a career batting average of 40 with a standard deviation of 10. But in his recent 30 innings he has a batting average of 38. The selection committee has a selection criteria where the average should not be less than 40. Now they want to confirm on which statistics they want to rely on to make the selection?
- Now, what will be your Null hypothesis and the Alternate hypothesis:
 - Null Hypothesis is the batting average is equal to 40
 - Alternate hypothesis is the batting average is less than 40.

Hypothesis Example

- Calculating the Probability using normal distribution, since the data is normally distributed:



- The Probability is 13.66%



Hypothesis testing through Z table

$$Z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{38 - 40}{10/\sqrt{30}} = \frac{-2}{1.825} = -1.09$$

The probability as per z-table is 0.1379 i.e. 13.79%

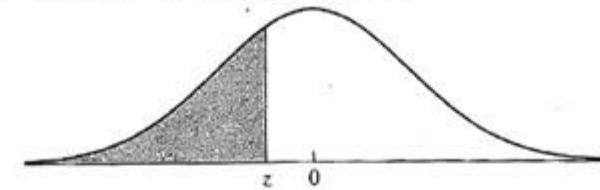
Conclusion: p-value as per table and excel is 0.1379 and 0.1366.

In both the cases p-value > significance level

$$0.1379/0.1366 > 0.05.$$

Hence, we cant reject null hypothesis.

TABLE A.2 Cumulative normal distribution (z table)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

Hypothesis Testing with type 1 and type 2 error

An International cargo company delivers product from Chennai to Singapore Port in 50 hrs. with a standard deviation of 12 hrs. is normally distributed

The Company has appointed you as a Supply Chain Analyst, and you observed the last 35 shipments and found the average delivery time between Chennai and Singapore is 53 hrs. with a standard deviation of 2.5 hours.

How Do you prove that your study is accurate and formulate the Null Hypothesis and Alternate Hypothesis:

Null and Alternate Hypothesis:

Null Hypothesis: There is no change in average delivery time, and the variation in sample is just because of random by chance

Alternate Hypothesis: There is a change in average delivery time and it has increased from the previous observations.

Finding the p-value:

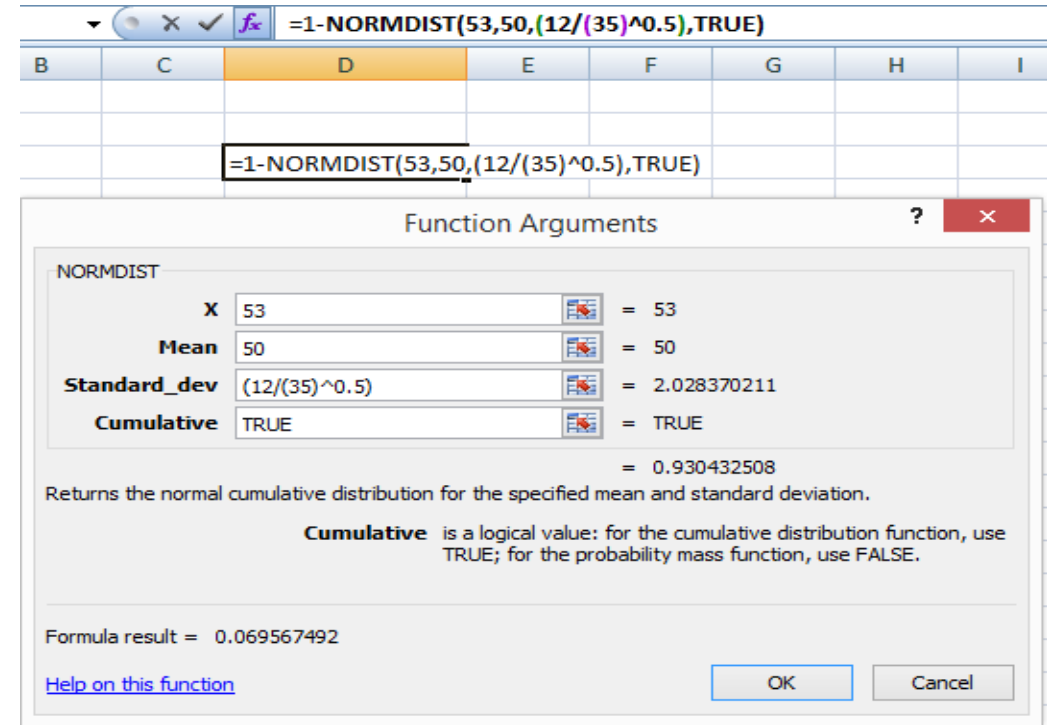
Using the available data we have the

Standard deviation of the sample is : 2

Population mean= 50

Sample Mean= 53

$$=1-\text{NORMDIST}(53,50,12/(35)^{0.5},\text{TRUE})$$



After solving the Problem, we get the P-value of 7%, which is greater than the 5% significance level.

Hence, we fail to reject the Null Hypothesis. Which means that we still have to consider the average delivery time as 50 hrs and not 53 hrs.

Types of Error in Hypothesis

- In Hypothesis testing, there are two kinds of errors that can be made in significance testing
 - (1) Type 1 Error : This error occurs if we reject the null hypothesis H_0 (in favor of alternative hypothesis (H_a)) when the null hypothesis (H_0) is true ,
 - (2) Type 2 Error : This error occurs if we fail to reject the null hypothesis H_0 when the alternative hypothesis H_a is true.
- The risks of these two errors are inversely related and determined by the level of significance and the power of the test. Therefore, you should determine which error has more severe consequences for your situation before you define their risks.
- No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a chance of drawing an incorrect conclusion.

Statistical Decision	True State of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not Reject H_0	Correct	Type II error

Type 1 Error

- When the null hypothesis is true and you reject it, you make a type I error. The rate of this error is called the size of the test.
- The probability of making a type I error is α , which is the level of significance of your hypothesis test.
 $P(\text{type 1 error}) = \text{significance level}$
- It is denoted by Greek letter α (alpha)

Now, considering the previous example, Suppose if we reject the null hypothesis even if it is true, i.e. We start assuming that the delivery time has increased to 53 hrs even if it has not increased, we commit a **Type 1 error**.

Type 2 Error

- When the Null Hypothesis is false, but still you do not reject it. This type of Error is called as Type 2 Error.

Again, considering the previous example, let's assume if you find the p-value less than 5% instead of 7% but still you do not reject the null hypothesis and still consider 50 hrs average delivery time correct, you commit a type two error.

Summarizing Type 1 and Type 2 Error

Let's assume a scenario in which a person is standing in front of a judge who is about to give a judgement. All the evidences prove that person is not guilty but the judge gives a judgement that the person is guilty.

What type of error is it?

This is a type 1 error in which the judge is denying the evidences and giving a judgement against the evidence i.e. the person is guilty.

What could have been a type 2 error?

Imagine if the evidences prove that the person is guilty but the judge gives a judgement that the person is not guilty. **This is a type 2 error.**

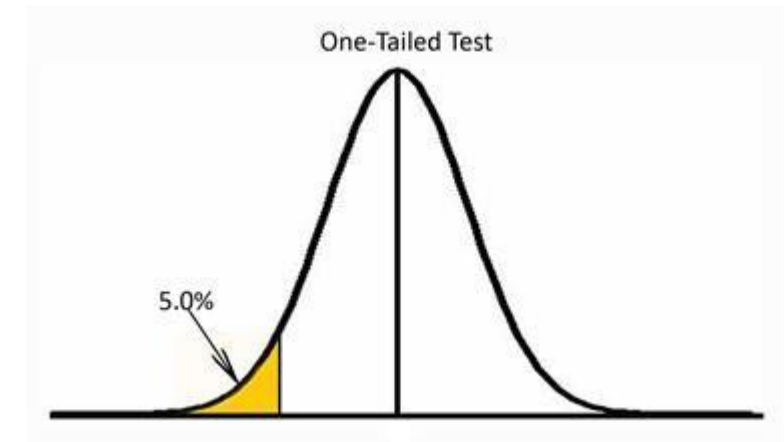
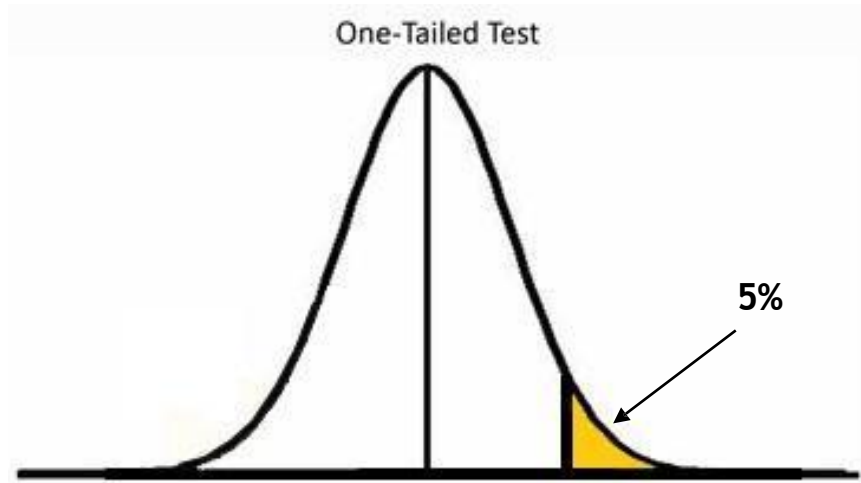
Two tailed and One tailed Test

- You have got one case from a company where the company has come out with a campaign to promote the product. You have to analyse whether the promotional campaign has positive impact on sales.

H_0 : No change in sale

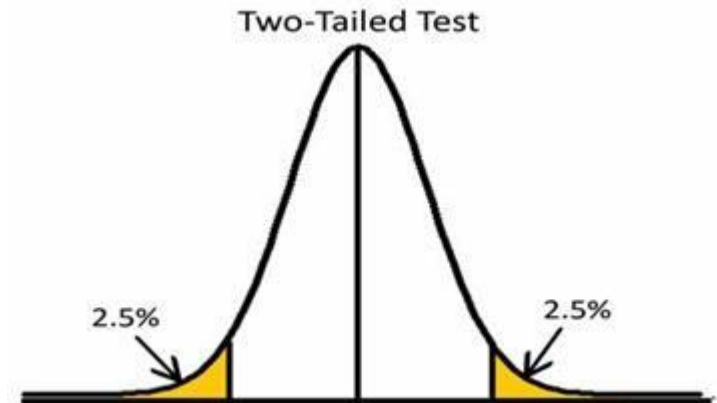
H_1 : Sales has increased

- If you have strong reason to believe that sales has increased then you should use an one tail test.



Two Tail Test

- Some time it may happened that a company has come out with a campaign to promote the product and they have used some controversial issue (ex: religious sentiments or unfair towards women). You have to analyse whether the promotional campaign has positively or negatively impacted the sales.
- In the above example, if we have a strong reason for believing that sample outcome may be more or less than the expected population mean, use a two tail test.



One tailed test Vs two tailed test

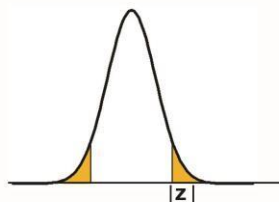
- A **two-tailed test**, also known as a **non directional hypothesis**, is the standard test of significance to determine if there is a relationship between variables in either direction. Two-tailed tests do this by dividing the .05 in two and putting half on each side of the bell curve.
- A **one-tailed test**, also known as a **directional hypothesis**, is a test of significance to determine if there is a relationship between the variables in one direction
- Two tailed test is very stringent because in this test rejection comes at 2.5%.
- One tailed test is most frequently used because we have an idea about the outcome, for example if you increase the advertisement expenses it will have positive impact on sales

Types of Alternative Hypothesis

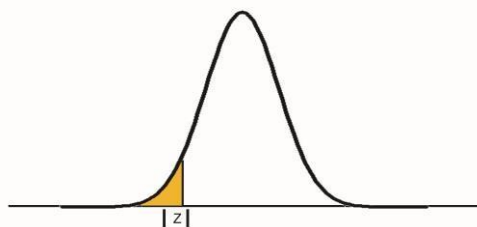
There are three types of alternative hypotheses:

- ✓ Two-sided test – When the population parameter is not equal to a certain value.
 $H_a: p \neq p_0$, or $H_a: \mu \neq \mu_0$
- ✓ Lower or Left-tailed test - When the population parameter is less than a certain value.
 $H_a: p < p_0$, or $H_a: \mu < \mu_0$
- ✓ Upper or Right-tailed test – When the population parameter is greater than a certain value.
 $H_a: p > p_0$, or $H_a: \mu > \mu_0$
Where, p = Proportion
 μ = Mean

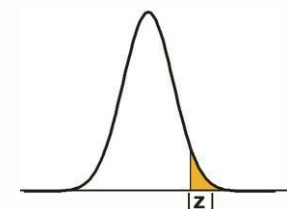
Two Tail test



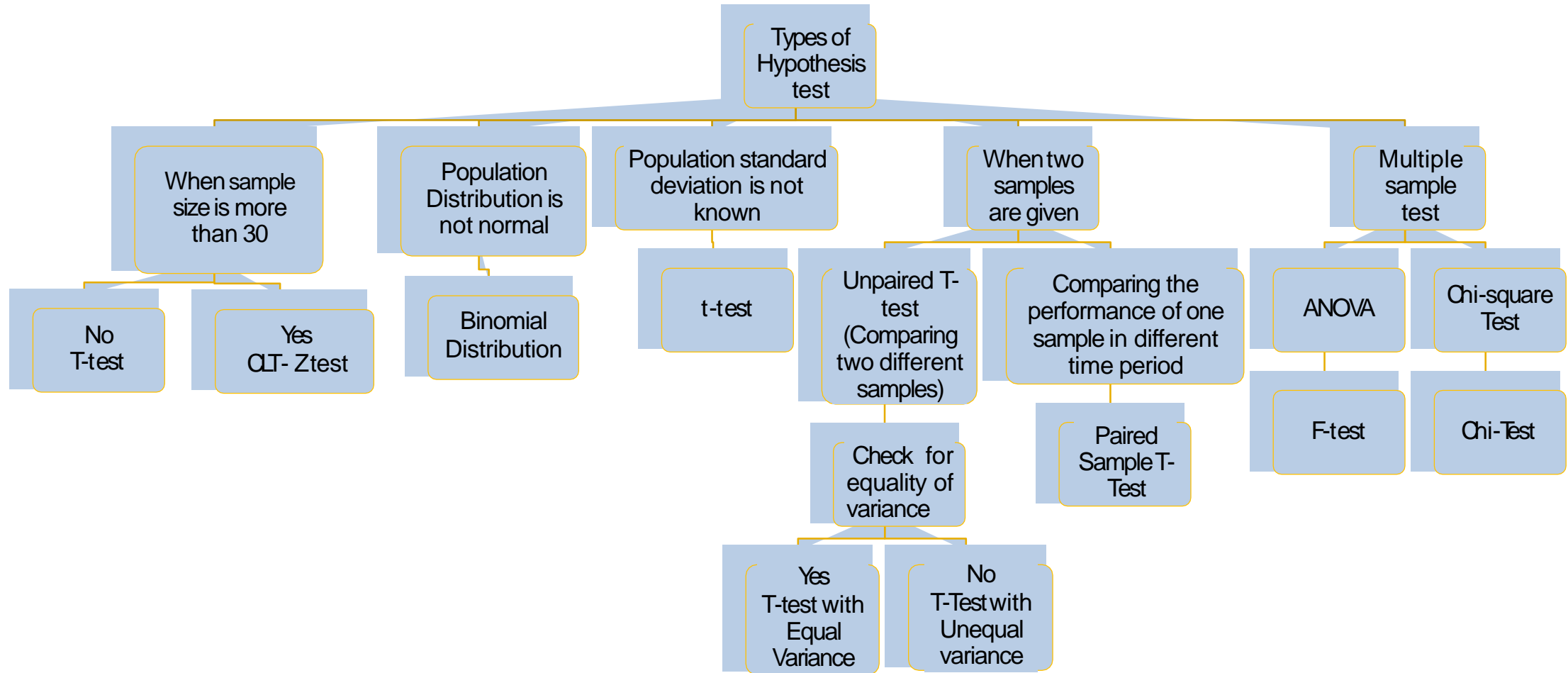
Left tail test



Right tail test



Types of Hypothesis Test



Sample Sizes are Small

Hypothesis testing : When the sample size is less than 30 (t-test)

In a survey report it is observed that the height of Chinese is less than global people. Average height of people is 155 cms. Taking random sample of 15 Chinese, Should we conclude that the height of Chinese is less than the global people ?

Observation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average
Heights	158	160	154	176	162	175	159	179	155	172	162	160	156	161	165	164

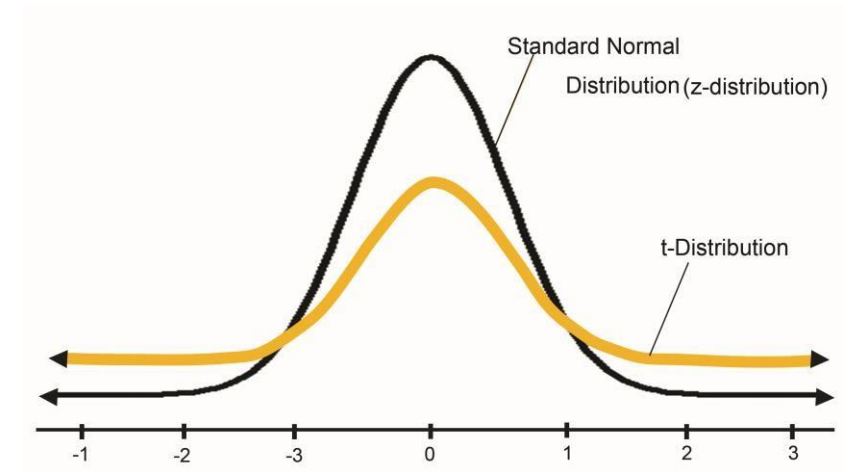
In this example, we cannot use the normal distribution function (NORMDIST) but have to use the T-Test, since the sample population is 30

There are two cases where you must use the t -distribution instead of the Z-distribution.

1. The sample size is small (below 30 or so),
2. When the population standard deviation is not known, and you have to estimate it using the sample standard deviation(s).
3. In both cases, you have less reliable information on which to base your conclusions, so you have to pay a penalty for this by using the t -distribution, which has more variability in the tails than a Z-distribution has.

T- test

- A hypothesis test for a population mean that involves the t -distribution is called a t -test. T



Degrees of Freedom in t-test

Degrees of freedom of an estimate is **the number of independent pieces of information that went into calculating the estimate**. In order to get the df for the estimate, you have to subtract 1 from the number of items.

Let's say you were finding the mean weight loss for a low-carb diet. You could use 4 people, giving 3 degrees of freedom ($4 - 1 = 3$).

Why do we subtract 1 from the number of items? Another way to look at degrees of freedom is that they are **the number of values that are free to vary** in a data set. What does “free to vary” mean? Here's an example using the mean (average):

Q. Pick a set of numbers that have a mean (average) of 10.

A. Some sets of numbers you might pick: 9, 10, 11 or 8, 10, 12 or 5, 10, 15.

Once you have chosen the first two numbers in the set, the third is fixed. In other words, **you can't choose the third item in the set**. The only numbers that are free to vary are the first two. You can pick 9 + 10 or 5 + 15, but once you've made that decision you **must** choose a particular number that will give you the mean you are looking for. So degrees of freedom for a set of three numbers is TWO.

Hypothesis testing:

For hypothesis testing, first of all, we need to check the test statistics & then need to find the critical distance. We need to repeat the same process that we followed in Z statistic, only we need to use sample standard deviation instead of Population standard deviation and use a t- test rather than a z value

Null Hypothesis:

Height of Chinese people is same as the height of Global Population

Alternate Hypothesis

Height of Chinese people is less than the height of Global Population

Considering 5% significance level

Test Statistic:

\bar{X} = Mean of sample

μ = Mean of population

S = Sample Standard Deviation (if sample S is not given, then you can find the SD of samples by using stdev in Excel)

N = Sample size

Hypothesis testing

Solution of Example

SDof Sample = stdev (sample datasets) = 8.05

Test Statistic = $(163.6 - 155) / (8.05 / 15^{0.5}) = 4.13$

Critical Distance:

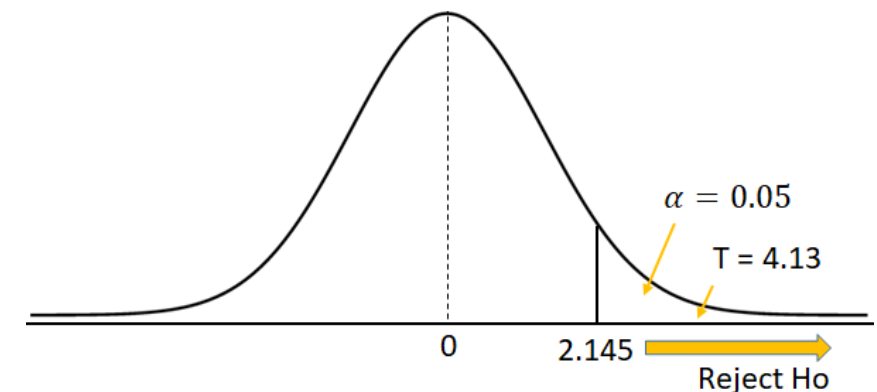
In t-test, we need degrees of freedom also to find the critical distance. In this case , df is 15 (sample size) – 1 = 14 You can find the Critical distance through t-test.

Critical Distance = 2.145

If test statistics (4.13) is farther away from mean (0) than the critical value (2.145), then reject null hypothesis.

Here we will conclude that height of Chinese people is more than the height of global population.

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073



Instead of using t-test for hypothesis testing, we can directly use p-value in Excel

Step 1: Calculate test statistics (Repeat the previous formula)

Test statistics = 4.13

Step 2: Use the T– distance value in Excel with “t.dist function and use df as 14

= 0.9994

p-value = $1 - 0.9994 = 0.0005$

Since the p-value is less than 0.0005 , we will reject the null hypothesis

Function Arguments ?

T.DIST

X	4.13	=	4.13
Deg_freedom	14	=	14
Cumulative	TRUE	=	TRUE
		=	0.999489703

Returns the left-tailed Student's t-distribution.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

Formula result = 0.999489703

[Help on this function](#)

OK Cancel

Population Standard Deviation not
known

Hypothesis testing : Population standard deviation is not known

So far many of the calculations we have population standard deviation, but in many cases the standard deviation is not known.

For Example:

- We test the weight of random sample of 40 city students We find average weight is 48kg with a standard deviation 8 kg. Weight-test are standardized such that mean weight is 45.

Would you be able to state with certainty that this group of students are more heavier than average?

- In the above example, the population standard deviation is not known, therefore sample standard deviation (s) can be substituted for population standard deviation

Hypothesis testing : Population standard deviation is not known

- In this example , we need to calculate
- What is the probability that purely by random chance, we picked a sample that gave us an average of 48 when the actual population average is 45?

Solution:

Ho: Weight of student is same as general population

Ha: Weight of student is greater than general population

Significance level : 5%

Test statistics: $(48 - 45) / (8 / ((40)^{0.5})) = 2.371$

P-value : $1 - \text{T.DIST}(2.37, 39, \text{TRUE}) = 1 - 0.99 = 0.011$

Conclusion: Reject null hypothesis

Two Sample t-test

Manufacturing department in a company A is facing a problem about the production of defective piece. The production department is asked to find a solution on reducing the issue of defective pieces.

Sample 1	Average no of defective pieces before project implementation on sample 1	Sample 2	Average no of defective pieces after project implementation on sample 2
1	5	1	4
2	9	2	7
3	4	3	3
4	8	4	7
5	7	5	5
6	11	6	9
7	10	7	8
8	8	8	6
9	6	9	4
10	7	10	6

Approach

1. Take a sample of defective pieces, note the average number of defective pieces
2. Implement project and note average number of defective pieces for another sample
3. Check if sample means are significantly different: 2 sample t-test

Explanation

1. When mean across sample of two groups are compared, we use 2 sample t-test
2. In most business scenarios, we check for differences across samples rather than a population outcome and a sample outcome.
3. The basic principle is to test the null hypothesis that the means of the two groups are equal.
4. The t-test compares the actual difference between two means in relation to the variation in the data

2 sample t-test

Sample 1	Average no of defective pieces before project implementation on sample 1	Sample 2	Average no of defective pieces after project implementation on sample 2
1	5	1	4
2	9	2	7
3	4	3	3
4	8	4	7
5	7	5	5
6	11	6	9
7	10	7	8
8	8	8	6
9	6	9	4
10	7	10	6
Average	7.5	Average	5.9

2 sample t-test

Hypothesis testing

Null hypothesis

Ho: No difference in the number of defective pieces post implementation of project

Alternate hypothesis

Ha: Reduction in the number of defective pieces post implementation of project

Solution:

Assume equal variance

Mean of defective pieces before project implementation
= 7.5

Mean of defective pieces after project implementation
= 5.9

Std. dev. of defective pieces before project implementation
= 2.2

Std. dev. of defective pieces after project implementation
= 1.9

DF = 10 + 10 - 2 = 18

t-stat = 1.646

Sample 1	Average no of defective pieces before project implementation on sample 1	Sample 2	Average no of defective pieces after project implementation on sample 2
1	5	1	4
2	9	2	7
3	4	3	3
4	8	4	7
5	7	5	5
6	11	6	9
7	10	7	8
8	8	8	6
9	6	9	4
10	7	10	6
Average	7.5	Average	5.9
SD	2.2		1.9

2 sample t-test in excel

Select t-test: Two sample assuming equal variances

Select Data Analysis

The screenshot shows the Excel interface with the **DATA** tab selected. The ribbon includes options like **Sort**, **Filter**, **Text to Columns**, **Flash Fill**, **Remove Duplicates**, **Data Validation**, **Consolidate**, **What-If Analysis**, **Relationships**, **Group**, **Ungroup**, **Subtotal**, and **Data Analysis**. A blue arrow points from the text "Select Data Analysis" to the **Data Analysis** button in the ribbon.

The worksheet contains the following data:

	A	B	C
1		Average no of defective pieces before project implementation	Average no of defective pieces after project implementation
2		5	4
3		9	7
4		4	3
5		8	7
6		7	5
7		11	9
8		10	8
9		8	6
10		6	
11		7	
12	Average	7.5	

The **Data Analysis** task pane is open, showing a list of analysis tools. A blue arrow points from the text "Select t-test: Two sample assuming equal variances" to the option **t-Test: Two-Sample Assuming Equal Variances** in the list.

Data Analysis

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances**
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

Buttons: OK, Cancel, Help

2 sample t-test in excel

Select input range for variable 1 and variable 2 then click on ok

The image shows an Excel spreadsheet with two columns of data. Column B is labeled 'Average no of defective pieces before project implementation' and Column C is labeled 'Average no of defective pieces after project implementation'. The data for Column B is: 5, 9, 4, 8, 7, 11, 10, 8, 6, 7, and the average is 7.5. The data for Column C is: 4, 7, 3, and the average is 7.5. A dialog box titled 't-Test: Two-Sample Assuming Equal Variances' is open, showing the input ranges for Variable 1 and Variable 2. The dialog box also includes options for the hypothesized mean difference, alpha, and output options.

	A	B	C	D	E	F	G
1		Average no of defective pieces before project implementation	Average no of defective pieces after project implementation				
2		5	4				
3		9	7				
4		4	3				
5		8					
6		7					
7		11					
8		10					
9		8					
10		6					
11		7					
12	Average	7.5					

t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

☐ Labels

Alpha:

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

2 sample t-test output Equal variance

As the alternate hypothesis is the reduction in the number of defective pieces. Hence, we are using a p-value of a one tail test to check the hypothesis.

The p-value of one tail test = $0.04 < \text{Significance level } (0.05)$. Hence, we reject null hypothesis.

Tstat approach,
critical value = 1.73
t-stat = 1.74

t-Stat > t-critical value, Reject Null Hypothesis.

In Business point of view: we are 95% confident that there is a statistically significant reduction in the defective pieces post project implementation.

t-Test: Two-Sample Assuming Equal Variances		
	Average no of defective pieces before project implementation	Average no of defective pieces after project implementation
Mean	7.5	5.9
Variance	4.72	3.65
Observations	10	10
Pooled Variance	4.18	
Hypothesized Mean Difference	0	
df	18	
t Stat	1.74	
P(T<=t) one-tail	0.04	
t Critical one-tail	1.73	
P(T<=t) two-tail	0.09	
t Critical two-tail	2.10	

2 sample t-test output - Unequal variances

- In previous slide, the variance of both the samples is not equal. (Not equal means when the difference in variance of both the sample is more than 15%)
- Hence, we have to use t-test with unequal variance.

t-Test: Two-Sample Assuming Unequal Variances		
	Average no of defective pieces before project implementation	Average no of defective pieces after project implementation
	n	n
Mean	7.5	5.9
Variance	4.72	3.65
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	1.74	
P(T<=t) one-tail	0.04	
t Critical one-tail	1.73	
P(T<=t) two-tail	0.09	
t Critical two-tail	2.10	

T-Test with unequal number of observations

In the above example we have same number of observations.

If the number of observations are unequal, we can use the t-test, but the degrees of freedom should be 1 less the small sample size

Paired Sample t-test

Hypothesis testing Paired sample t-test

Paired Two Sample For Means is used when your sample observations are naturally paired. The usual reason for performing this test is when you are testing the same group twice. For example, if you are testing a new drug, you'll want to compare the sample before and after they take the drug to see if the results are different. This particular t-test in Excel used a paired two-sample test to determine if the before and after observations are likely to have been derived from distributions with equal population means.

Assumptions for paired sample T-test:

To run the paired sample t-test data

- ✓ It should have a normal distribution
- ✓ Check whether it is a large dataset
- ✓ No outlier should present

Hypothesis testing Paired sample t-test example

A College has launched some mock interviews that claims that it will help students to get more job offers.

College record average of job offers received by students before and after the mock interview for 10 students for 24 weeks. Test the hypothesis that the student have a positive impact of mock interview with 95% confidence level

Calculate the mean and standard error of the student selection before mock interview differences for each pair and then use that for the test statistic

Test statistic:

$$t = \frac{\overline{d}}{\frac{S_d}{\sqrt{n}}}$$

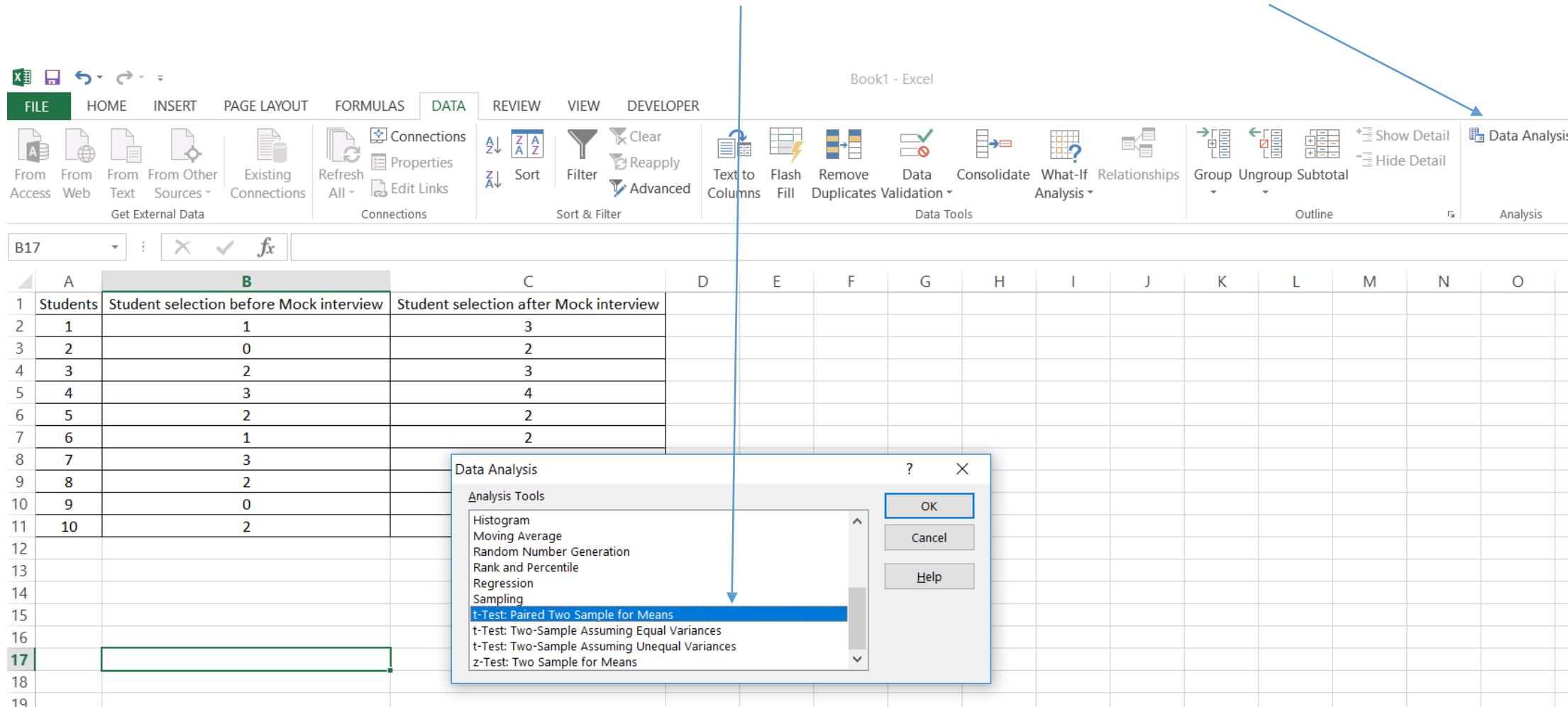
Where d is the difference in the score

Students	Job offers before Mock interview	Job offers after Mock interview
1	1	3
2	0	2
3	2	3
4	3	4
5	2	2
6	1	2
7	3	4
8	2	3
9	0	1
10	2	3

Hypothesis testing Paired sample t-test example in excel

Select t-test : paired two sample for means

Select Data Analysis



Hypothesis testing Paired sample t-test example in excel

- The p-value for one tail test is $8.66386E-05 < \text{Significance level (0.05)}$.
- Hence we reject null hypothesis.
- **Business Point of view:** The students have positive impact of job offers after the mock interview.

t-Test: Paired Two Sample for Means		
	Student selection before Mock interview	Student selection after Mock interview
Mean	1.6	2.7
Variance	1.15	0.9
Observations	10	10
Pearson Correlation	0.84	
Hypothesized Mean Difference	0	
df	9	
t Stat	-6.12	
P(T<=t) one-tail	8.66E-05	
t Critical one-tail	1.83	
P(T<=t) two-tail	0.000173277	
t Critical two-tail	2.262157163	

Population Distribution is not Normal

Population Distribution Not Normal

Population Distribution not normal for every hypothesis test. In the below example, the **Binomial distribution** is used.

Example: A Pollster conducts an survey before the general election and claims that 3 out of 10 people will vote for party A. A random sample of 100 people results in 15 people preferring party A or vote for A, is the pollster claim justified?

Solution: H_0 : Preference for Party A 30% ($\frac{3}{10} * 100$)

H_1 : Preference for Party A is less than 30%

Significance level : 5%

The outcome in the population follows Binomial distribution (Vote/ No Vote)

Binomial Distribution:

p-value: $\text{BINOM.DIST}(15, 100, 0.3, \text{TRUE}) = 0.0004$

Since p-value < Significance level Here, $0.0004 < 0.05$. We reject null hypothesis.

Conclusion:

Pollster claim is NOT justified, and preference for Party A is actually less than 30%, at a 95% confidence level

Population Distribution Not Normal

Alternate use a **normal distribution** if we have a binomially distributed random variable

Then,

$$\text{approx. mean} = n * p$$

$$\text{approx. std. deviation} = \sqrt{n * p * q}$$

Here,

$$\text{mean} = 100 * 0.3 = 30$$

$$\text{Std. deviation} = \sqrt{0.3 * 100 * 0.70} = 4.58$$

Normal Distribution Formula:

$$p\text{-value} = \text{NORM.DIST}(15, 30, 4.58, \text{TRUE}) = 0.0005.$$

Conclusion:

Since $p\text{-value} < \text{Significance level}$ Here, $0.0005 < 0.05$. We reject null hypothesis.

ANOVA

ANOVA

ANOVA

- Nike sports has launched a product in three different Metro cities Delhi, Mumbai and Bangalore. They allocated same marketing budget to these cities, but got different sales figure from each city. To understand this difference, the sales team wants to analyse whether the sales is dependent on cities or as an Analytics Consultant you want to analyse whether the sales figures are statistically different in each city.

	Delhi	Mumbai	Bangalore
Week 1	980	1123	1084
Week 2	776	1357	1025
Week 3	923	1152	1114
Week 4	1498	921	1182
Week 5	999	959	1022

Whether Sales is dependent on cities?

ANOVA

- Now suppose if we are considering only 2 cities to find out the sales i.e. Delhi & Mumbai and the sales team wants to analyse whether the sales is dependent on cities.

	Delhi	Mumbai
Week 1	980	1123
Week 2	776	1357
Week 3	923	1152
Week 4	1498	921
Week 5	999	959

- In this case we can use t-test to find out the difference between them
- But we have to find out the difference between 3 samples, then instead of t-test we have to use ANOVA.**

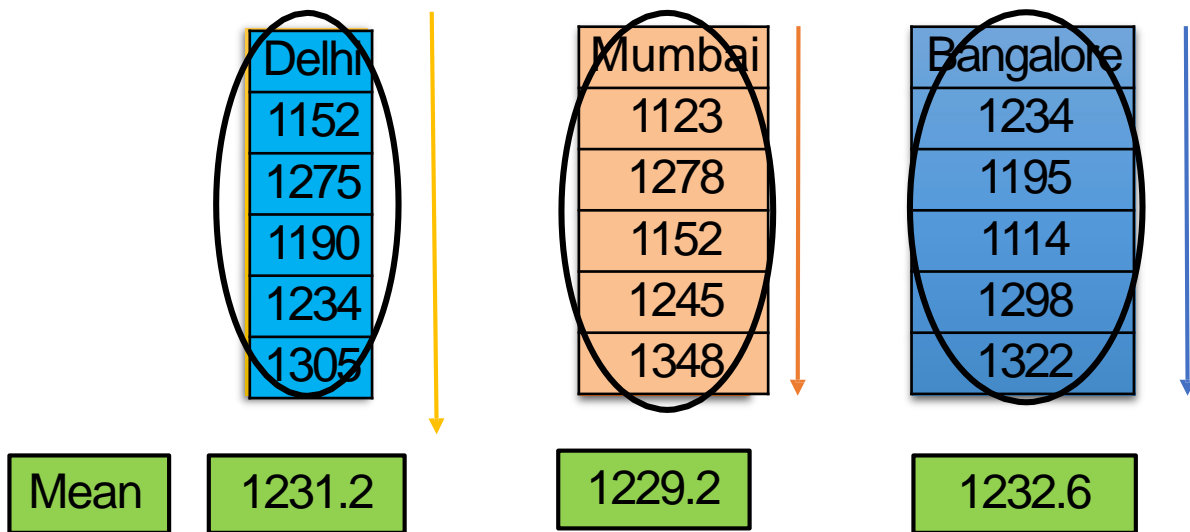
ANOVA

- The Hypothesis of ANOVA:
 - ✓ **Null Hypothesis (H_0):** The mean sales for each cities are same.
 - ✓ **Alternate Hypothesis (H_a):** The mean sales for each city is different.
- Total variation in sample data can be account of two component
 - ✓ Variance between the sample: This is because of sales according to cities are different
 - ✓ Variance within the sample: This is because of the people in each group have different sales in each week.
- ❖ For more understanding, lets go through some examples in the next slides

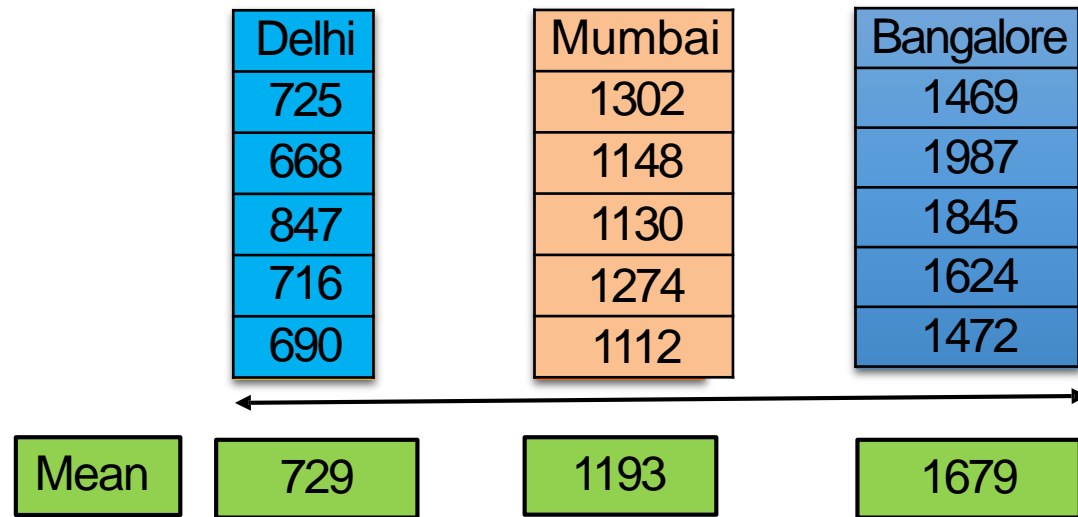
Variance Within Samples (SSE) and Variance Between Samples(SSC)

- In example 1, we can see that the samples are much alike there is not much variation between the samples, but there is a lot of variation within each city.
- In Example 2, all the sales within each group is very close to one another but the samples are very different from one another.
- There is a lot of difference between cities.

Example 1:



Example 2:



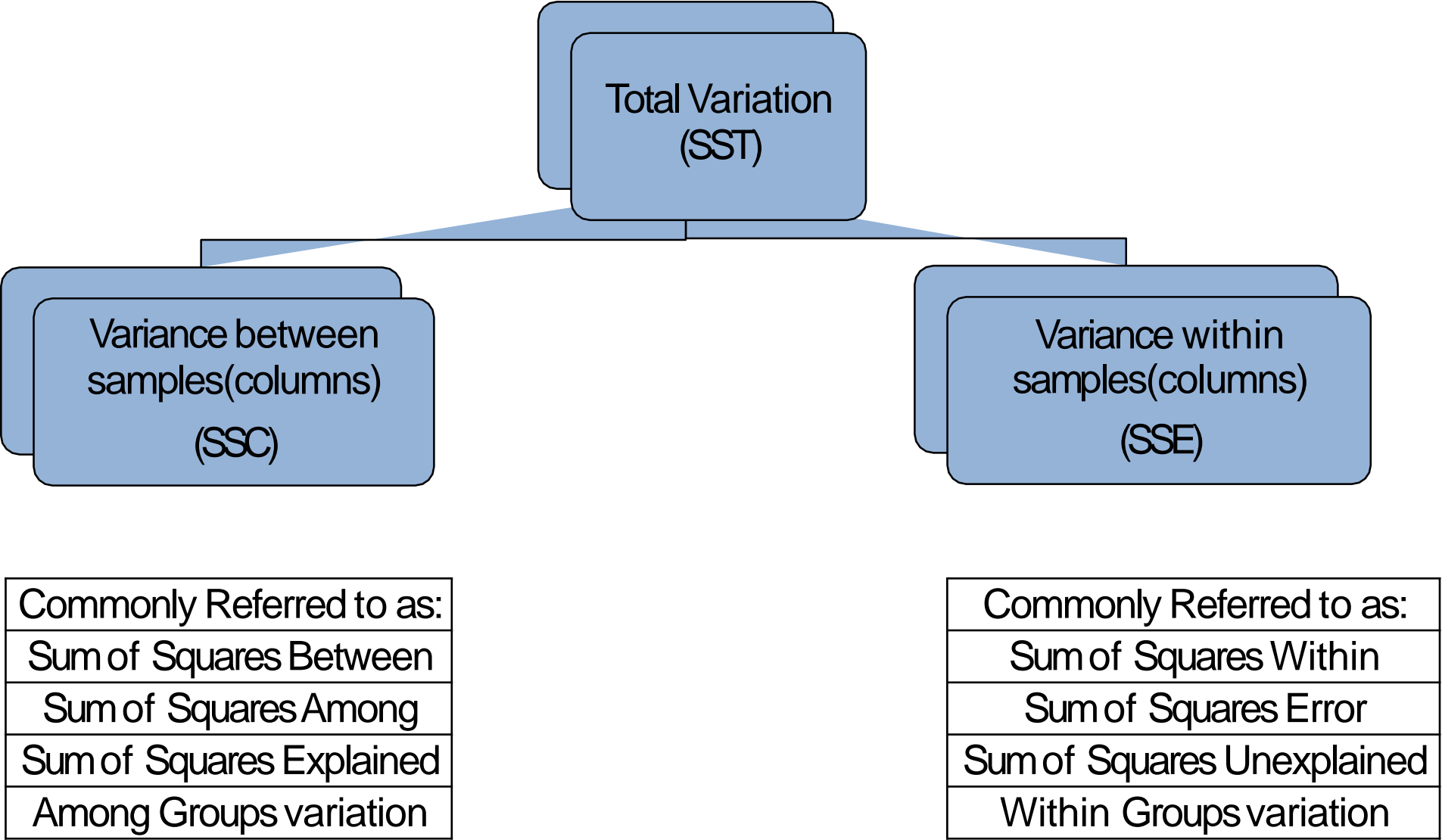
F - Test

- The F-ratio is the ratio of Mean square column (MSC) to the Mean square error (MSE).
- F-test is used to find out the total variance comes from the variance between samples and variance within samples.

$$F = \frac{\left(\begin{array}{l} \text{(Mean square column) or } \frac{SS}{\text{Degrees of freedom}} \end{array} \right)}{\left(\begin{array}{l} \text{(Mean Square error) or } \frac{SS}{\text{Degrees of freedom}} \end{array} \right)}$$

- The numerator degrees of freedom is k-1 i.e. samples - 1
- The denominator degrees of freedom is n-k i.e. observations – samples.
- The larger the ratio, the more likely it is the groups have different means.
- Let's understand the step by step process of finding one way ANOVA in the coming slides

Partition of Total Variation



Total Sum of Square

Delhi
980
776
923
1498
999
Mumbai
1123
1357
1152
921
959
Bangalore
1084
1025
1114
1182
1022

Total Sum of Squares (SST)

=

Delhi
980
776
923
1498
999



Mumbai
1123
1357
1152
921
959



Bangalore
1084
1025
1114
1182
1022



+

Delhi
980
776
923
1498
999

Mumbai
1123
1357
1152
921
959

Bangalore
1084
1025
1114
1182
1022



Variance Within Samples /
Sum of Squares within Samples
(SSE)

Variance Between Samples /
Sum of Squares Between
Samples (SSC)

Variance Within Samples (SSE)

- The procedure of calculating the variance within samples is as below:
 - ✓ Calculate the mean of each sample
 - ✓ Calculate the difference of each observation in k samples from the mean values of the respective samples.
 - ✓ Square all the difference obtained in step 2 and calculate the total of all these squared difference.
- **Variance Within Samples (SSE) = 298315 + 121175 + 17819 = 437309**

	Delhi			
	Sales (S)	Mean (M)	D= S- M	Square of D
Week 1	980	1035	-55	3047
Week 2	776	1035	-259	67185
Week 3	923	1035	-112	12589
Week 4	1498	1035	463	214184
Week 5	999	1035	-36	1310
Total				298315

	Mumbai			
	Sales (S)	Mean (M)	D= S- M	Square of D
Week 1	1123	1102	21	424
Week 2	1357	1102	255	64821
Week 3	1152	1102	50	2460
Week 4	921	1102	-181	32906
Week 5	959	1102	-143	20564
Total				121175

	Bangalore			
	Sales (S)	Mean (M)	D= S- M	Square of D
Week 1	1084	1085	-1	2
Week 2	1025	1085	-60	3648
Week 3	1114	1085	29	818
Week 4	1182	1085	97	9332
Week 5	1022	1085	-63	4020
Total				17819

Variance Between Samples (SSC)

- The procedure of calculating the variance between the sample is as follows:
 - Calculate mean of each sample
 - Calculate a Grand mean
 - Calculate the difference between the mean of each sample and grand mean
 - Square the difference between the mean of each sample and grand mean
 - Calculate the sum of each sample mean
 - Multiply sum by the number of observation in each sample

Variance Between Samples (SSC)

$$= 5(1531 + 788 + 122) = 12208$$

Delhi	
Week 1	980
Week 2	776
Week 3	923
Week 4	1498
Week 5	999
Mumbai	
Week 1	1123
Week 2	1357
Week 3	1152
Week 4	921
Week 5	959
Bangalore	
Week 1	1084
Week 2	1025
Week 3	1114
Week 4	1182
Week 5	1022
Grand Mean (GM)	1074

	Delhi	Mumbai	Bangalore
Week 1	980	1123	1084
Week 2	776	1357	1025
Week 3	923	1152	1114
Week 4	1498	921	1182
Week 5	999	959	1022
Mean (M)	1035	1102	1085
D = M - GM	-39	28	11
Square of D	1531	788	122
SSC = Sum (Number of observation * sum of Squares)	12208		

Total Sum of Squares (SST)

- $SST = SSE + SSC$
- The calculation of SST using other methods is mentioned in excel
- The procedure of calculating the total sum of square is as follows:
 1. Calculate a Grand Mean
 2. Calculate the difference between the observation and Grand Mean.
 3. Square the difference between the observation and Grand Mean.
 4. Calculate the sum of the square of observation and Grand Mean.

	Observation (O)	Grand Mean (M)	D= O- GM	Square of D
Delhi				
Week 1	980	1074	-94	8899
Week 2	776	1074	-298	89003
Week 3	923	1074	-151	22902
Week 4	1498	1074	424	179493
Week 5	999	1074	-75	5675
Mumbai				
Week 1	1123	1074	49	2368
Week 2	1357	1074	283	79900
Week 3	1152	1074	78	6032
Week 4	921	1074	-153	23511
Week 5	959	1074	-115	13302
Bangalore				
Week 1	1084	1074	10	93
Week 2	1025	1074	-49	2434
Week 3	1114	1074	40	1573
Week 4	1182	1074	108	11592
Week 5	1022	1074	-52	2739
Grand Mean (GM)	1074			
Total				449517

F Ratio

- The F ratio is the ratio of Mean square column (MSC) to the Mean square error (MSE).

MSC

- ✓ The Mean square column is the ratio of Variance between sample to the (k-1) degrees of freedom.


$$MSC = \frac{SS}{k-1}$$

- ✓ The Degrees of freedom is (k-1) i.e. Number of samples(k) - 1.
- We have calculated Variance Between Samples (SSC) in previous slides i.e. 12208.
- In our Example, There are 3 samples i.e. Delhi, Mumbai and Bangalore.

- ✓ Degrees of Freedom = Samples - 1
 $= 3 - 1 = 2$

$$MSC = \frac{SS}{k-1} = \frac{12208}{2} = 6104$$

1	2	3
Delhi	Mumbai	Bangalore
980	1123	1084
776	1357	1025
923	1152	1114
1498	921	1182
999	959	1022



F Ratio

$$MSE = \frac{SSE}{n - k}$$

MSE

- ✓ The Mean square error is the ratio of variance within sample to the (n-k) degrees of freedom.
 - ✓ The Degrees of Freedom is (n-k) i.e. Total number of observations(n) – Number of Samples(k).
- We have calculated Variance Within Samples (SSE) in previous slides i.e. 437309.
 - In our Example, The total number of Observations is 15 and Total number of samples is 3 i.e. Delhi,
 - Mumbai and Bangalore.
 - ✓ Degrees of Freedom = Total Number of Observations – Number of Samples

$$= 15 - 3 = 12$$

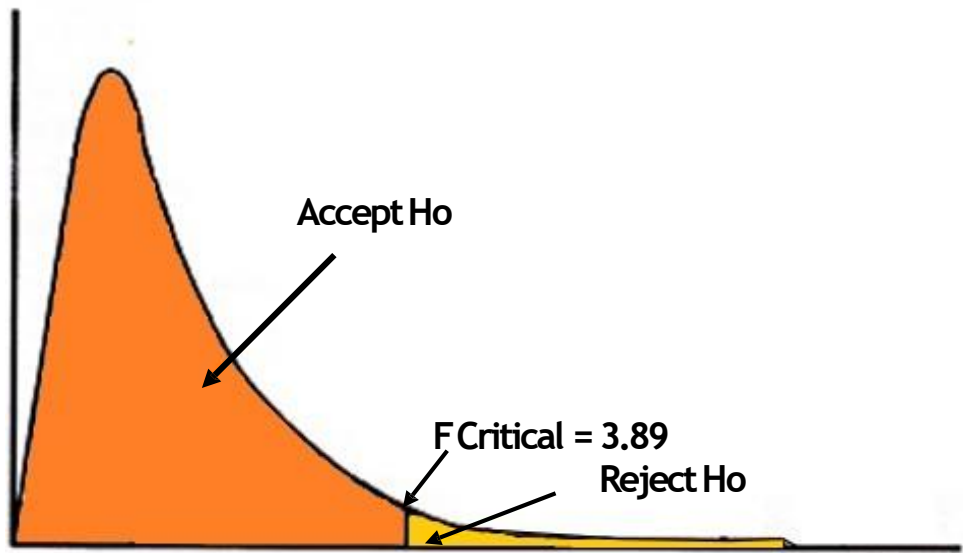
$$MSE = \frac{SSE}{n - k} = \frac{437309}{12} = 36442$$

Delhi	Mumbai	Bangalore
980	1123	1084
776	1357	1025
923	1152	1114
1498	921	1182
999	959	1022

F Ratio

• $F - \text{Test} = \frac{MSC}{MSE} = \frac{6104}{36442} = 0.167$

Therefore, referring to the table, F-critical Value= 3.89



$\nu_1 \backslash \nu_2$	1	2	3	4	5	6	7	8
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94

Hence, considering the values, The F-stat value has to be greater than the F—Critical Value. Which in this case, is smaller. i.e (F-stat<F-Critical Value).

Thus, we cannot reject the Null Hypothesis and find that the mean sales for each cities are same.

One-Way ANOVA in Excel

Data -> Data Analysis -> ANOVA: Single Factor -> Input Range -> Columns.

	A	B	C	D
1		Delhi	Mumbai	Bangalore
2	Week 1	980	1123	1084
3	Week 2	776	1357	1025
4	Week 3	923	1152	1114
5	Week 4	1498	921	1182
6	Week 5	999	959	1022
7				
8				
9				
10				
11				

Anova: Single Factor

Input
Input Range:

Grouped By:
☒ Columns
☐ Rows

☒ Labels in First Row

Alpha:

Output options
☐ Output Range:
☒ New Worksheet ply:
☐ New Workbook

OK
Cancel
Help

One-Way ANOVA in Excel

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Delhi	5	5176	1035.2	74578.7		
Mumbai	5	5512	1102.4	30293.8		
Bangalore	5	5427	1085.4	4454.8		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	12208.13	2	6104.06	0.167	0.847	3.885
Within Groups	437309.2	12	36442.43			
Total	449517.33	14				

Hence, considering the values, The P-value is 0.847 which is greater than the significance level of 0.05 Hence, we failed to reject the null hypothesis.

Thus, we cannot reject the Null Hypothesis and find that the mean sales for each cities are same.

Two Way ANOVA

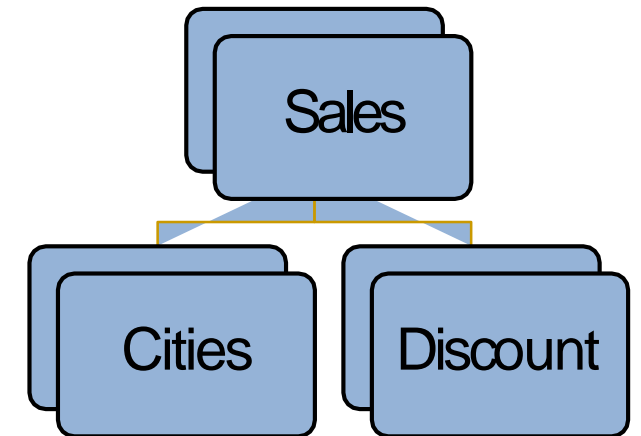
- In the previous example we have analysed sales considering only cities but now one more factor “discount” is added along with cities. Now there are two factors to analyse a sales. Hence, it is called Two Way ANOVA
- Two way ANOVA has 2 types:
 - ✓ Two way ANOVA with replication: It is performed when you have two factors and individuals within that factors are doing more than one thing
 - ✓ Two way ANOVA without replication: It can compare a group of individuals performing more than one task.

Two Way ANOVA with Replication

Discount	Delhi	Mumbai	Bangalore
10%	133	84	94
	107	114	123
	114	138	108
	150	155	148
	85	78	93
20%	130	138	120
	86	119	127
	130	146	150
	112	158	106
	125	137	127

Two Way ANOVA without Replication

Discount	Delhi	Mumbai	Bangalore
10%	133	84	94
20%	130	138	120



Two-Way ANOVA Example

Two Way ANOVA With Replication

Nike sports recently launched discount schemes at their premium and best performing stores in Delhi, Mumbai and Bangalore. Now, the Analytics team wants to analyze whether there is an impact of discounts and location on the sales. Do they need to keep different discount structure for different cities?

They collected the data of previous 5 days of sales at different discount rates and location.

Null Hypothesis for sample(H_0): The impact of discount on the sales is same.

Null Hypothesis for column(H_0): The impact of cities on sales is same.

Null hypothesis for Interaction: The combination of the discount and cities has no impact on sales.

Discount	Delhi	Mumbai	Bangalore
10%	133	84	94
	107	114	123
	114	138	108
	150	155	148
	85	78	93
20%	130	138	120
	86	119	127
	130	146	150
	112	158	106
	125	137	127

Two-Way ANOVA

- Two-way ANOVA as its name signifies, is a hypothesis test wherein the classification of data is based on two factors.
- For instance, the two bases of classification for the sales made by the firm is
 - ✓ Sales in various Cities
 - ✓ Sales when discount
- It is a statistical technique used by the researcher to compare several levels (condition) of the two independent variables involving multiple observations at each level.
- Two-way ANOVA examines the effect of the two factors on the continuous dependent variable. It also studies the inter-relationship between independent variables influencing the values of the dependent variable, if any.

Discount	Delhi	Mumbai	Bangalore
10%	133	84	94
	107	114	123
	114	138	108
	150	155	148
	85	78	93
20%	130	138	120
	86	119	127
	130	146	150
	112	158	106
	125	137	127

Assumptions of two-way ANOVA:

- ✓ Normal distribution of the population from which the samples are drawn.
- ✓ Measurement of dependent variable at continuous level.
- ✓ Two or more than two categorical independent groups in two factors.
- ✓ Categorical independent groups should have the same size.
- ✓ Independence of observations
- ✓ Homogeneity of the variance of the population.

Two Way ANOVA in Excel

- Data -> Data Analysis -> ANOVA: Two-Factor with Replication -> OK.

The screenshot shows the Excel interface with the 'Data' tab selected. The ribbon includes 'Get External Data', 'Get & Transform', 'Connections', 'Sort & Filter', and 'Data Tools'. The data table is as follows:

	A	B	C	D
1		Delhi	Mumbai	Bangalore
2	Discount 10%	133	84	94
3		107	114	123
4		114	138	108
5		150	155	148
6		85	78	93
7	Discount 20%	130	138	120
8		86	119	127
9		130	146	150
10		112	158	106
11		125	137	127

The 'Data Analysis' dialog box is open, showing the 'Analysis Tools' list. 'Anova: Two-Factor With Replication' is selected. The 'OK' button is highlighted.

Two Way ANOVA in Excel

Select the input range -> Rows per sample -> OK

	A	B	C	D
1		Delhi	Mumbai	Bangalore
2	Discount 10%	133	84	94
3		107	114	123
4		114	138	108
5		150	155	148
6		85	78	93
7	Discount 20%	130	138	120
8		86	119	127
9		130	146	150
10		112	158	106
11		125	137	127

Anova: Two-Factor With Replication

Input
Input Range:
Rows per sample:
Alpha:

Output options
☐ Output Range:
☒ New Worksheet Ply:
☐ New Workbook

OK
Cancel
Help

Two-Way ANOVA Example

Anova: Two-Factor With Replication						
SUMMARY	Delhi	Mumbai	Bangalore	Total		
Discount 10%						
Count	5	5	5	15		
Sum	589	569	566	1724		
Average	117.8	113.8	113.2	114.93		
Variance	618.7	1113.2	527.7	650.07		
Discount 20%						
Count	5	5	5	15		
Sum	583	698	630	1911		
Average	116.6	139.6	126	127.4		
Variance	346.8	203.3	253.5	325.11		
Total						
Count	10	10	10			
Sum	1172	1267	1196			
Average	117.2	126.7	119.6			
Variance	429.51	770.01	392.71			
ANOVA						
Source of Variation	SS	df	MS	F	P-value	Fcrit
Sample	1165.63	1	1165.63	2.28	0.14	4.26
Columns	488.07	2	244.03	0.48	0.63	3.40
Interaction	911.67	2	455.83	0.89	0.42	3.40
Within	12252.8	24	510.53			
Total	14818.17	29				

The p-value of null is greater than 0.05, We fail to reject Null hypothesis, which means that, The impact of discount on the sales is the same.

One Way ANOVA VS Two Way ANOVA

One Way ANOVA	Two Way ANOVA
A hypothesis test that enables us to test the equality of three or more means simultaneously using variance.	A statistical technique in which the interrelationship between factors, influencing variable can be studied for effective decision making.
There is only one factor or independent variable	There are two factors or two independent variables.
One-way ANOVA, compares three or more levels (conditions) of one factor.	Two-way ANOVA compares the effect of multiple levels of two factors
In one-way ANOVA, the number of observations need not be same in each sample	In two-way ANOVA, it should be same in the case of two-way ANOVA
One-way ANOVA need to satisfy only two principles of design of experiments, i.e. replication and randomization.	Two-way ANOVA, meets all three principles of design of experiments which are replication, randomization, and local control.

Chi Square Test

Chi Square Test

Honda automotive started a new showroom in a Tier 2 city in India and wants to calculate whether the sales transaction of the new showroom are reflecting the same trend when compared to the national average of sales.

Considering a particular set of sedan cars, The transaction of each type of sedan car in showrooms all over India is calculated by the company.

Also, a random sample of 200 sales transactions for these cars is taken from the new showroom.

National sales figure

Cars	Sales transaction
Honda City	46%
Honda Amaze	23%
Honda Accord	21%
Honda CR-V	10%

Sample of 200 transactions from the new showroom

Cars	No of transactions	Sales Transaction
Honda City	98	49%
Honda Amaze	40	20%
Honda Accord	52	26%
Honda CR-V	10	5%

Can we use Anova or t-test to compare the national sales figure with the sample figures?

Chi-Square

Z Test : Z-test is used to determine whether two population means are different when the variances are known and the sample size is large .

ANOVA : ANOVA used when data is numeric and number of samples are more than 2.

Chi Square

Chi square is used when we have categorical data

- The Chi Square test is a type of Non-Parametric test, which deals with multiple samples with categorical data.
- Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis.
- The Chi Square Test is a statistical test which consists of three different types of analysis
 - 1) Test for Homogeneity
 - 2) Chi Square test of Association or Test of Independence
 - 3) Goodness of fit

Test of Homogeneity

Test of Homogeneity

- A test of homogeneity compares the proportions of responses from two or more populations with regards to a dichotomous variable (e. g., male/female, yes/no) or variable with more than two outcome categories.
- The chi-square test of homogeneity is the nonparametric test used in a situation where the dependent variable is categorical.
- The chi-square test of homogeneity statistic is computed in exactly the same manner as chi-square Test of Independence statistic.
- The difference between these two tests consists of stating the null hypothesis, the underlying logic, and the sampling procedures.

Example of Chi Square test - Test of Homogeneity

Cars	Sales transaction
Hyundai i10	46%
Hyundai i20	23%
Hyundai Verna	21%
Hyundai Creta	10%

Expected Transaction

Cars	No of transactions	Sales Transaction
Hyundai i10	98	49%
Hyundai i20	40	20%
Hyundai Verna	52	26%
Hyundai Creta	10	5%

Observed transaction

Null Hypothesis: There is no difference between the overall sales transaction share and the new showroom sales.

Alternate Hypothesis: There is a difference between the overall sales transaction and the new showroom sales.

Example of Chi Square test

Cars	Sales	Sales transaction
Hyundai i10	92	46%
Hyundai i20	46	23%
Hyundai Verna	42	21%
Hyundai Creta	20	10%

Expected Transaction(Comparing with 200)

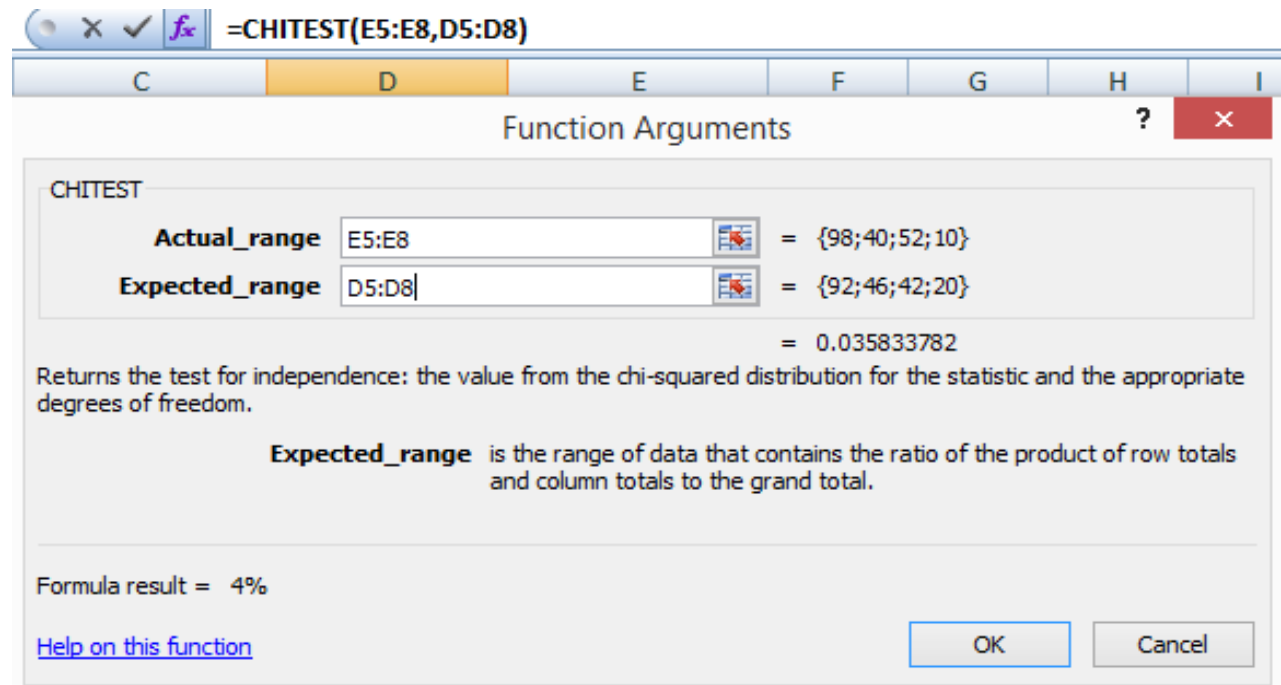
Cars	Sales	Sales transaction
Hyundai i10	98	49%
Hyundai i20	40	20%
Hyundai Verna	52	26%
Hyundai Creta	10	5%

Observed Sales transaction

Example of Chi Square test

Solving the example in Excel, using the formula =CHITEST(actual_range, expected_range)

We get the P-Value of 4%, which is less than that of the 5% significance level.



Hence, we reject the Null hypothesis and conclude that the sales transaction of the new showroom is different than the overall transaction percentage.

Test of Independence or Test of Association

- An important application of chi square test involves using the sample data to test for the independence of two categorical variable
- This test enables us to test whether or not two variables are associated.
- The null hypothesis for this test is that the two categorical variables are independent. The test is referred to as a test of independence
- When null hypotheses is rejected it can be concluded that there is a significant association between two variables

Chi Square test Example- Test of Association

Hyundai Automotive is planning to launch three cars in a new state. The research team conducted a survey in the state to know the buying preferences of the people.

They conduct a survey on random sample of people from urban, rural and semi urban areas.

Now, Hyundai Automotive wants to know whether there is a specific preference for cars in different areas?

Observed Sample				
Car Type	Urban	Semi Urban	Rural	Total
Hyundai i20	74	52	48	174
Hyundai Verna	29	21	12	62
Hyundai Creta	17	34	39	90
Total	120	107	99	326

Null Hypothesis: There is no specific preference for cars in different areas.

Alternate Hypothesis: There is a specific preference for cars in different areas.

Chi Square test Example

Now, since the expected value is not given, we can calculate that by using,

$$E_{ij} = \frac{T_i \times T_j}{N}$$

Where, E_{ij} is the expected frequency for the cell in the i th row and the j th column, T_i is the total number of subjects in the i th row, T_j is the total number of subjects in the j th column, and N is the total number of subjects in the whole table.

You can think of this equation more simply as (row total * column total) / grand total.

Expected Values				
Car Type	Urban	Semi Urban	Rural	Total
Hyundai i20	64	57	53	174
Hyundai Verna	23	20	19	62
Hyundai Creta	33	30	27	90
Total	120	107	99	326

Chi Square test Example

After calculating the expected values, we can directly use the Excel formulae, **CHITEST(actual_range, expected_range)** to find the p-value.

After calculating the p-value of 0.00047, which is less than the significance level of 5%.
Hence, we reject the Null Hypothesis and suggest that people give preference to specific cars in different areas.

Goodness of Fit

Chi Square Test- Goodness of Fit Example

An automobile company has a defect rate of 10%. A sample of 100 deliveries is taken, and it is found that there is a 16% defect rate. Has the defect rate increased?

Number of Defects	Observed Frequency
0	84
1	16

Chi Square Test- Goodness of Fit

- In Chi-Square goodness of fit-test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution.
- Chi-Square goodness of fit-test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the distribution.
- In Chi-Square goodness of fit-test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

When to Use the Chi-Square Goodness of Fit-test

- The chi-square goodness of fit-test is appropriate when the following conditions are met.
- The sampling method is simple random sampling.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

Chi Square Test- Goodness of Fit Example

Hypothesis of the mentioned Problem:

Null Hypothesis(Ho): There is no change in the Defects rate.

Alternate Hypothesis(Ha): There is a change in the Defects rate.

QUARTILE =BINOMDIST(B6,1,0.1,FALSE)*100

A	B	C	D	E	F
		Expected			
	Number of Defects	Frequency			
	0	90			
	1	=BINOMDIST(B6,1,0.1,FALSE)*100			

BINOMDIST(number_s, trials, probability_s, cumulative)

Calculating the probability, by using binomial distribution

Chi Square Test- Goodness of Fit Example

The screenshot shows an Excel spreadsheet with the following data:

Expected		Observed	
Number of Defects	Frequency	Number of Defects	Frequency
0	90	0	84
1	10	1	16

The formula bar at the top displays: `=CHITEST(F5:F6,C5:C6)`

A tooltip for the CHITEST function is shown at the bottom, indicating the syntax: `CHITEST(actual_range, expected_range)`.

Thus, Calculating the p-value, using the CHITEST function in excel.


Since the p-value is 0.045, which is less than the value of 0.05, we reject the null hypothesis and find that there is a change in the defect rate.

Function Arguments

CHITEST

Actual_range


F5:F6



= {84;16}

Expected_range

C5:C6



= {90;10}

= 0.04550027

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Expected_range

is the range of data that contains the ratio of the product of row totals and column totals to the grand total.

Formula result = 0.04550027

[Help on this function](#)

OK

Cancel

Chi Square Test- Goodness of Fit Example

An automobile company has a defect rate of 10%. A sample of 100 deliveries is taken, and it is found that there is a 16% defect rate. Has the defect rate increased?

Number of Defects	Observed Frequency
0	84
1	16

Chi Square Test- Goodness of Fit

- In Chi-Square goodness of fit-test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution.
- Chi-Square goodness of fit-test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the distribution.
- In Chi-Square goodness of fit-test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

When to Use the Chi-Square Goodness of Fit-test

- The chi-square goodness of fit-test is appropriate when the following conditions are met.
- The sampling method is simple random sampling.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

Chi Square Test- Goodness of Fit Example

Hypothesis of the mentioned Problem:

Null Hypothesis(Ho): There is no change in the Defects rate.

Alternate Hypothesis(Ha): There is a change in the Defects rate.

QUARTILE =BINOMDIST(B6,1,0.1,FALSE)*100					
A	B	C	D	E	F
		Expected			
	Number of Defects	Frequency			
	0	90			
	1	=BINOMDIST(B6,1,0.1,FALSE)*100			
		BINOMDIST(number_s, trials, probability_s, cumulative)			

Calculating the probability, by using binomial distribution

Chi Square Test- Goodness of Fit Example

=CHITEST(F5:F6,C5:C6)					
B	C	D	E	F	G
	Expected			Observed	
Number of Defects	Frequency		Number of Defects	Frequency	
0	90		0	84	
1	10		1	16	
=CHITEST(F5:F6,C5:C6)					
CHITEST(actual_range, expected_range)					

Function Arguments

CHITEST

Actual_range

F5:F6

= {84;16}

Expected_range

C5:C6

= {90;10}

= 0.04550027

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Expected_range

 is the range of data that contains the ratio of the product of row totals and column totals to the grand total.

Formula result = 0.04550027

[Help on this function](#)

OK

Cancel

Thus, Calculating the p-value, using the CHITEST function in excel.

Since the p-value is 0.045, which is less than the value of 0.05, we reject the null hypothesis and find that there is a change in the defect rate.