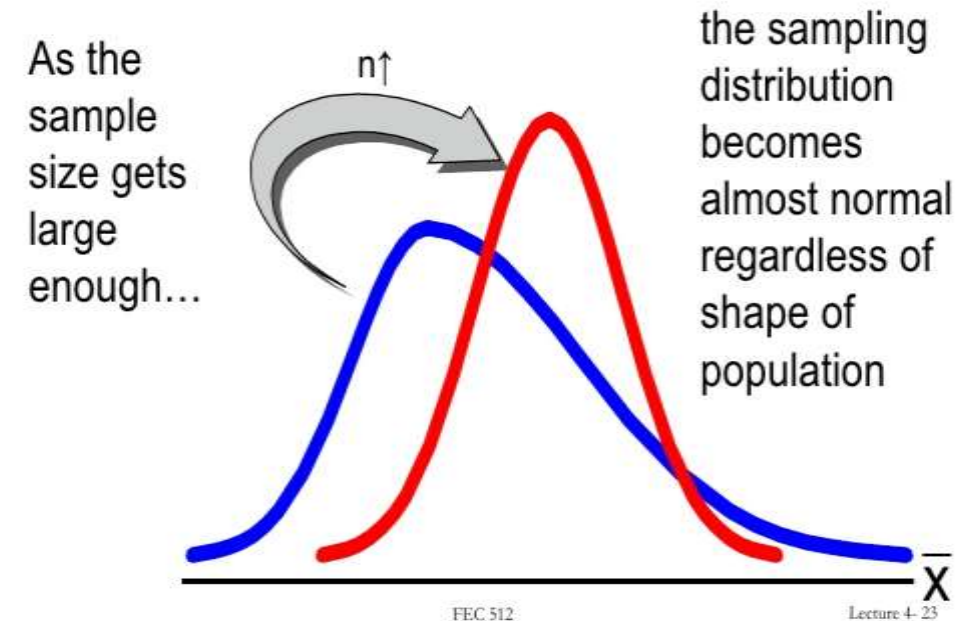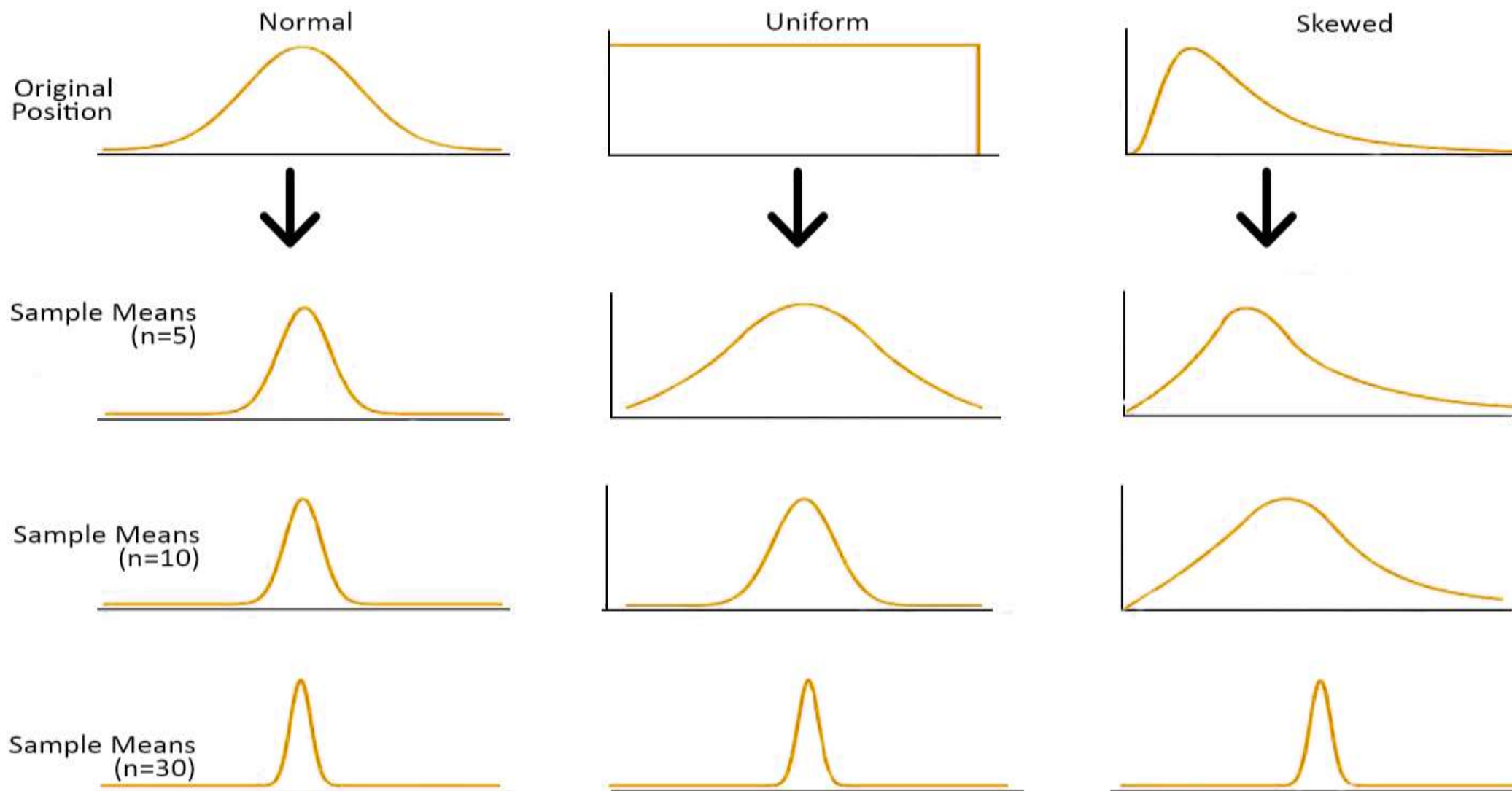# Central Limit Theorem

- 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

As the sample size gets large enough...

n↑

the sampling distribution becomes almost normal regardless of shape of population

$\bar{X}$

FEC 512   Lecture 4– 23

Shapes of Distributions as Sample Size Increases

Test Statistic of Sample Mean

- To determine the probability of sample mean falls within a particular range, we use,

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

- The sample mean $\bar{x}$ is a random variable and its probability distribution is called the sampling distribution of $\bar{x}$

- The sampling distribution is denoted by Z

- Where, Z follows Normal distribution with E($\bar{x}$) and standard deviation $\sigma_{\bar{x}}$

- Here, E($\bar{x}$) = μ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

- Lets understand CLT with an example

Central Limit Theorem- Example

- In a survey of a company, mean salary of employees is 6066 dollars with SD of 2875 dollars. Consider the sample of 40 employees and find the probability their mean salary will be less than 6000 dollars?

- Can we use the normal distribution formulae here? (NORMDIST(6000,6066,2875,TRUE))

   -Is it a right approach?

Font			Alignment
=NORMDIST(6000,6066,2875,TRUE)			
C	D	E	F
0.490842			

Central Limit Theorem- Example

In this example, we have sample size n = 40. We calculate the mean of each sample and hence we get 40 means. These are called as sample means. The values we get our as follows:

Population Mean   6061.28
Sample Mean   6066.38

We can see that population mean and sample mean is almost same.
 In other words the distribution of sample means will have a mean equal to the population mean. Remember it is the mean of sample means and that distribution of sample mean that its mean will be equal to the population mean.

What about the standard deviation?

The population standard deviation can be found out easily.
Population Standard Deviation   2878.92

But is the standard deviation of the sample means the same as the standard deviation of the population?
No!
Sample Standard Deviation   437.30

It turns out that the standard deviation of the sample mean will be a lot lower than the standard deviation of the population. Why?

Central Limit Theorem- Example

Now supposing that we pick up samples of size 40 and someone picking at samples of size 10.

In which of these two set of samples would we expect to see a greater variation in the sample averages.
Intuitively we understand that the variation of sample averages will be a lot higher in the samples of size 10 relative to the samples of size 40.

Therefore it has to be the case that the standard deviation of the averages of samples has to be inversely proportional to sample size.
The larger the sample, the more likely less variation in sample averages. The smaller the sample, the more likely greater variation in sample averages and therefore the standard deviation of the sample averages will be equal to standard deviation of the population/square root of sample size.

Hence the formula, $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

Central Limit Theorem- Example

- The correct formulae being, **the standard deviation of the population divided by the square root of the sample size.**

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

Microsoft Excel
Worksheet

Thus, the revised formulae seems to be,

Z= 6066.38-6061.28/(2878.92/40^0.5)
Z= 0.0112

And finally, calculating the p-value:
0.5040

Based on the P-value we conclude accept H0
Note: The values taken in z-value formula are based on dataset

	A	B
1	Sample Mean	6066.384
2	Population Mean	6061.283
3	Population Standard Deviation	2878.928
4	Observations (n)	40
5		
6	Sample Mean - Population Mean	5.101
7	Population Standard deviation/ sqrt(n)	455.19848
8	Z value	0.0112061
9		

Introduction to Hypothesis

Example 1: You are an Analytics advisor of a Retail company and find from the population that the average daily sales is 20,000 USD (approx.) and Standard Deviation is USD 4000.

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?"

- **H0 – mu = 20000**

- **Ha – mu<20000**

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

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

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 form 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 (Ho) : Null hypothesis is the hypothesis which is tested for the possible rejection under the assumption that is true
  - ✓ Alternative Hypothesis (Ha) : Alternative hypothesis is the logical opposite of the null hypothesis
- **In the above example:**
  - ✓ Ho : No reduction in daily sales
  - ✓ Ha: 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.

0.05

>0.05? Or <0.05?
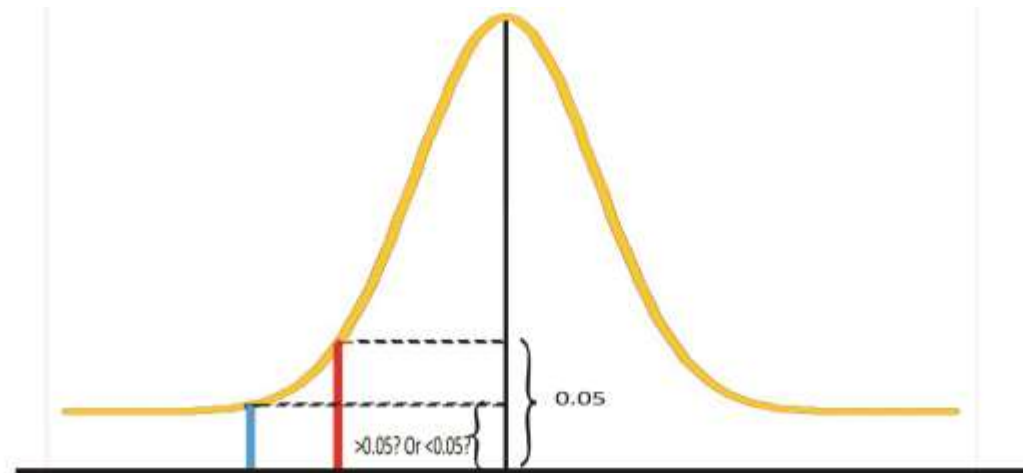
More on p-value

- 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.

P-values

What if p-value > Significance level?
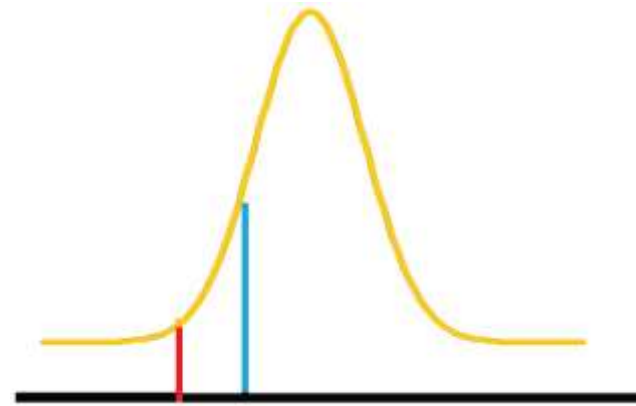
Conclusion?

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.
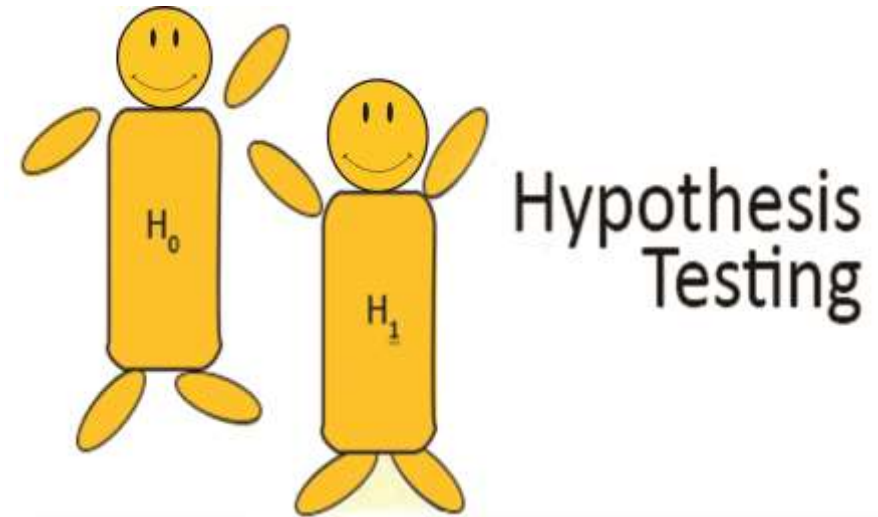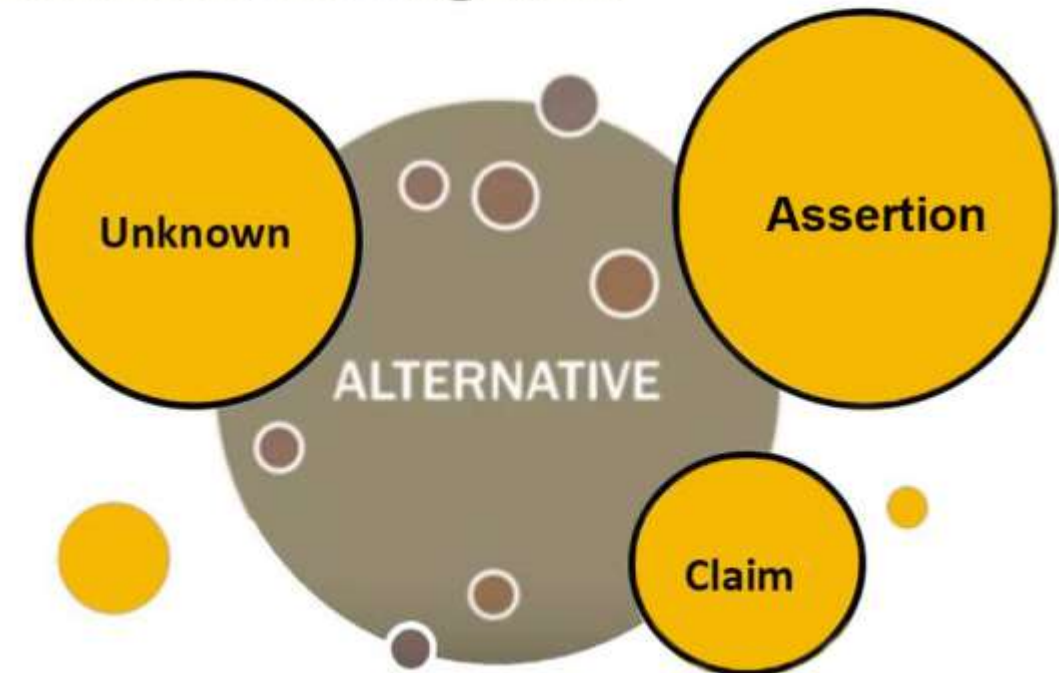
Alternative Hypothesis

- 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.

Null and Alternate Word Salad
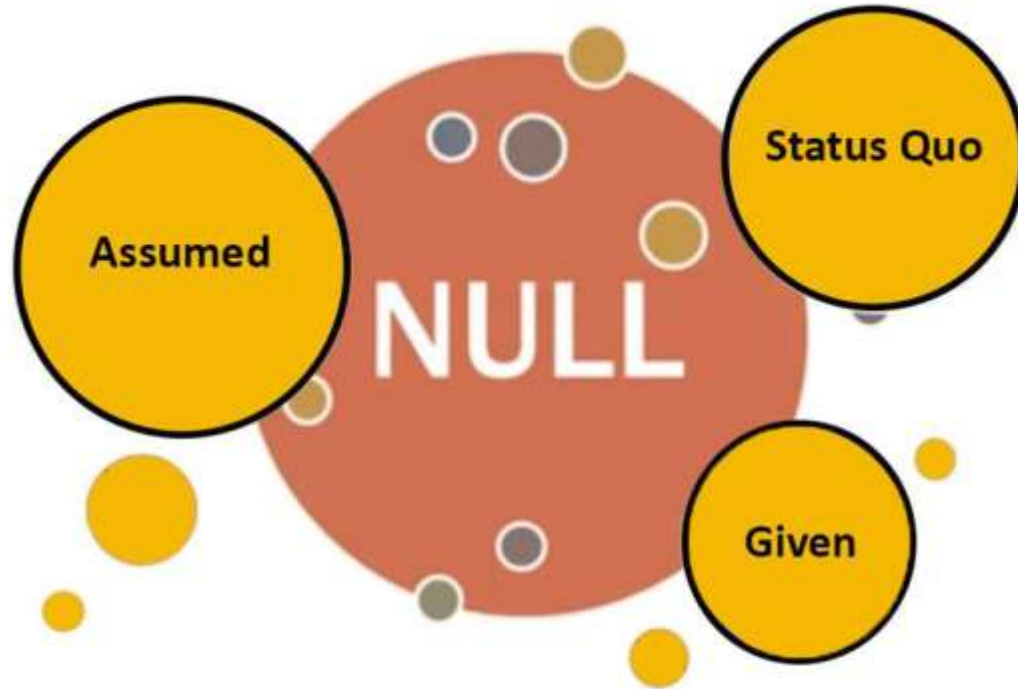
Examples of hypothesis

- 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

  Ho: x = 98.6

  Ha: x < 98.6 [Alternative Hypothesis includes the outcomes not covered by the null hypothesis.]

- 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.

  Ho: x = .65

  Ha: x < .65

- A radio station publicizes that the average number of the local listening audience is greater than 45%

  Ho: x = 0.45

  Ha: x < 0.45

Another example

- A 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:
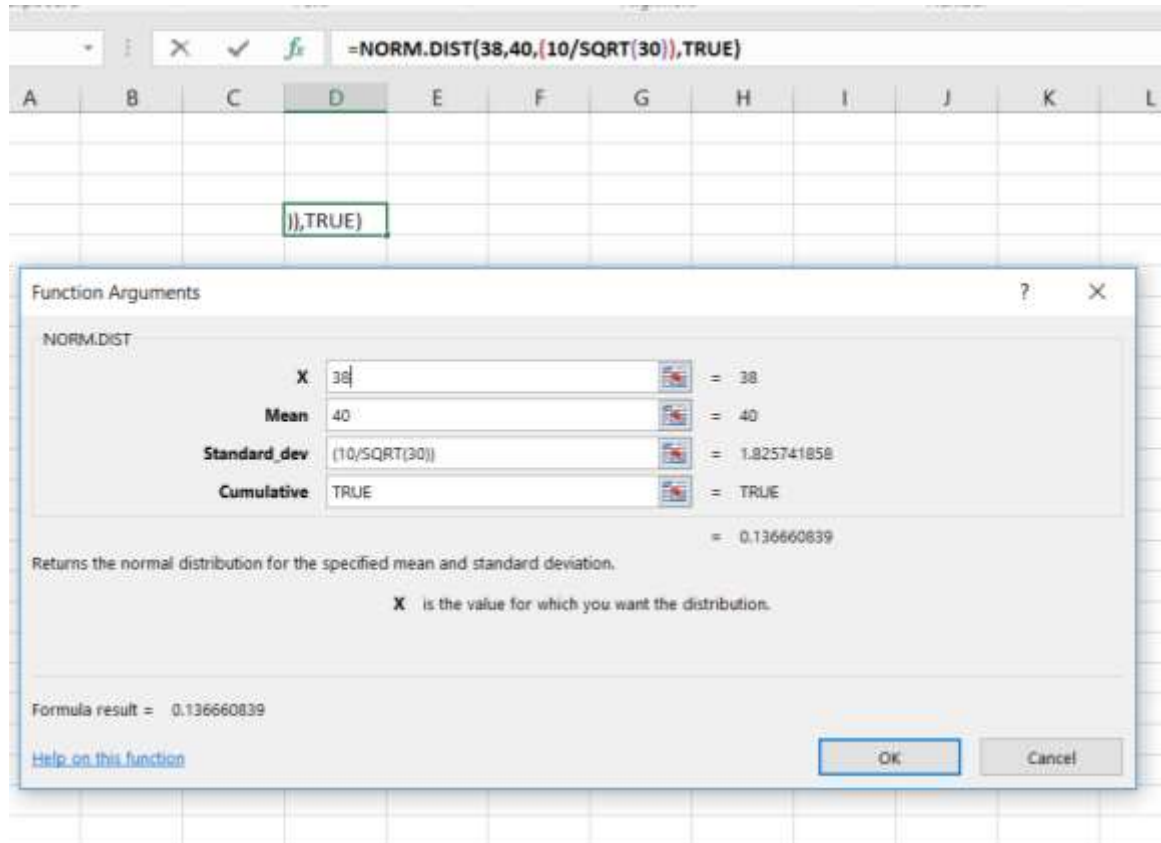
- The Probability is 13.66%
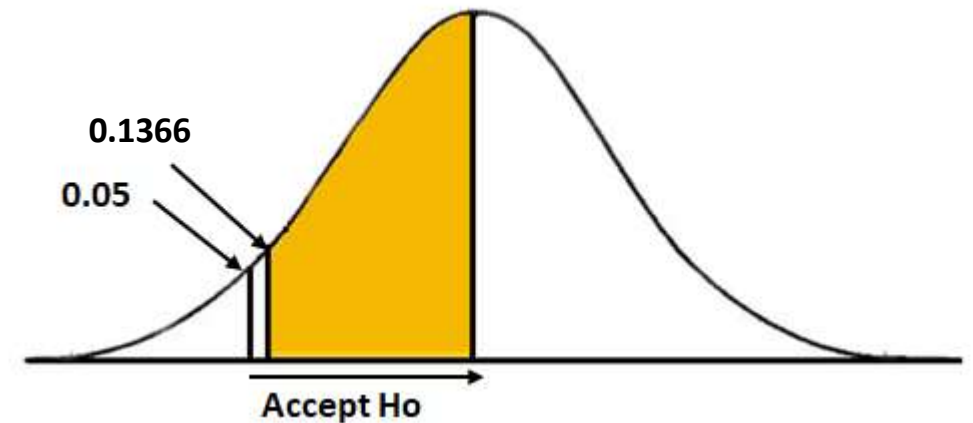
Hypothesis testing through Z table

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

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

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)

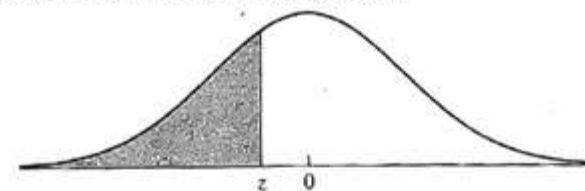| z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.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

A International cargo company delivers product from Mumbai to Dubai 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 Mumbai and Dubai 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:

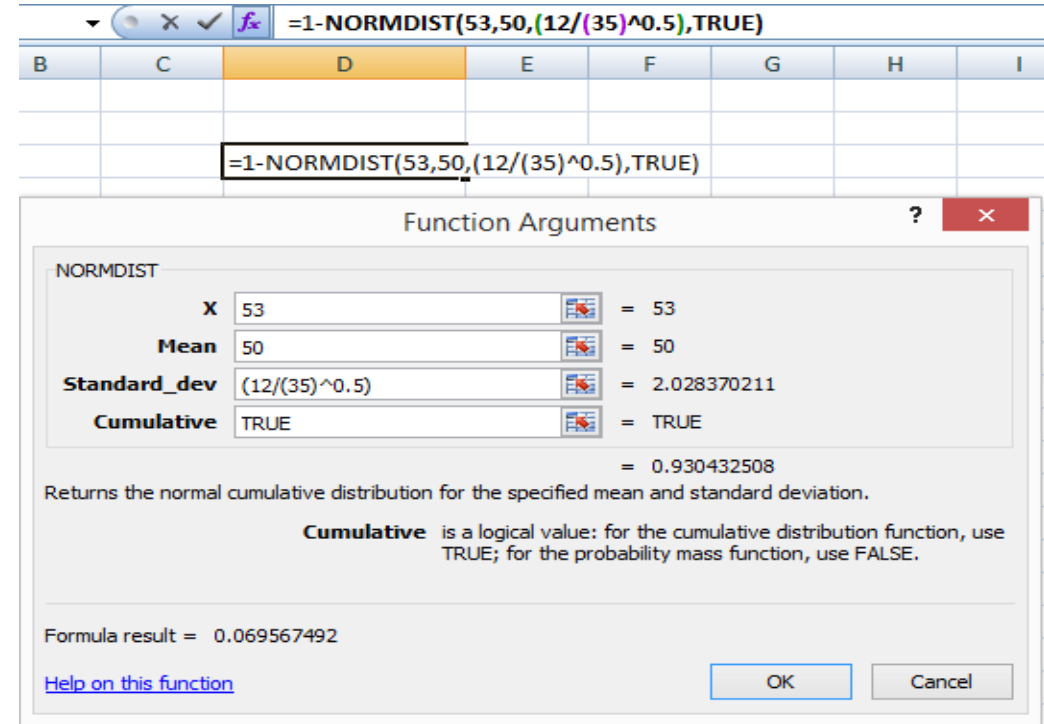Using the available data we have, the

Standard deviation of the sample is : 2
Population mean= 50
Sample Mean= 53

=1-NORMDIST(53,50,12/(35)^0.5,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 Ho (in favor of alternative hypothesis (Ha)) when the null hypothesis (Ho) 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

An Initiative

- 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, lets 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

Lets 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.

    Ho : No change in sale

    H1: Sales has increased

- If you have strong reason to believe that sales has increased then you should use an one tail test.
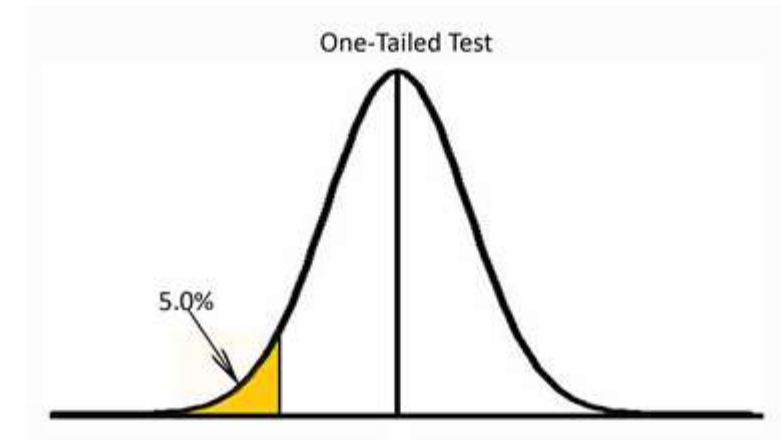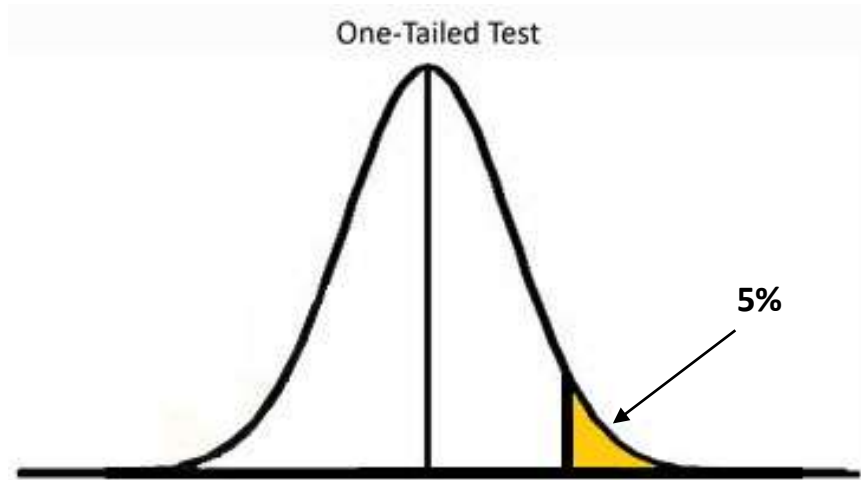
- Some time it may happened that a company has come out with a campaign to promote the product and they have used some controversial issue (ex: religious sentiments or unfair towards women). You have to analyse whether the promotional campaign has positively or negatively impacted the sales.

- 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.

Two-Tailed Test

2.5%   2.5%
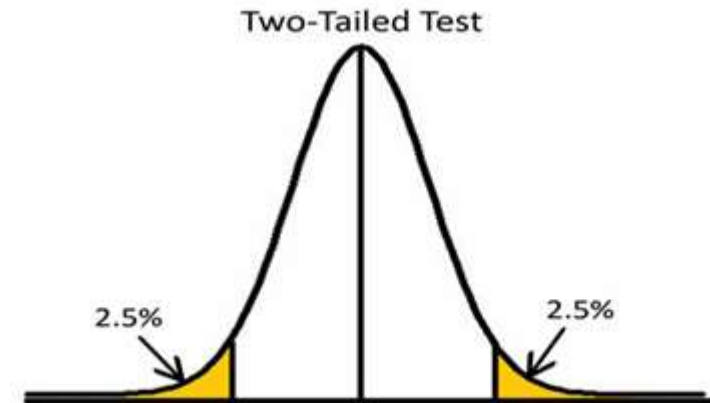
One tailed test Vs two tailed test

- 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.
  $Ha: p \neq p0$, or $Ha: \mu \neq \mu0$

- ✓ Lower or Left-tailed test - When the population parameter is less than a certain value.
  $Ha: p < p0$, or $Ha: \mu < \mu0$

- ✓ Upper or Right-tailed test – When the population parameter is greater than a certain value.
  $Ha: p > p0$, or $Ha: \mu > \mu0$
  Where, p = Proportion
  μ = Mean
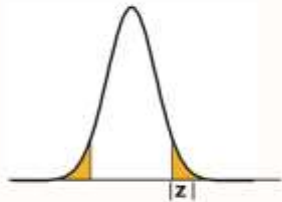
Two Tail test	Left tail test	Right tail test
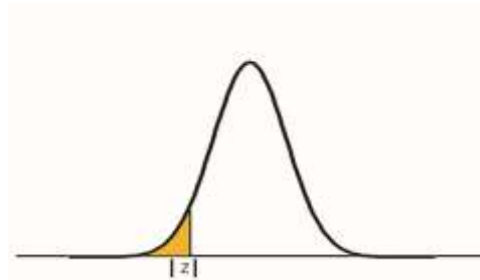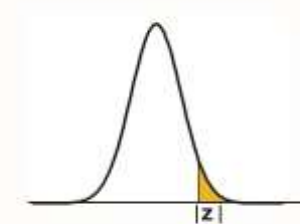
Types of Hypothesis Test
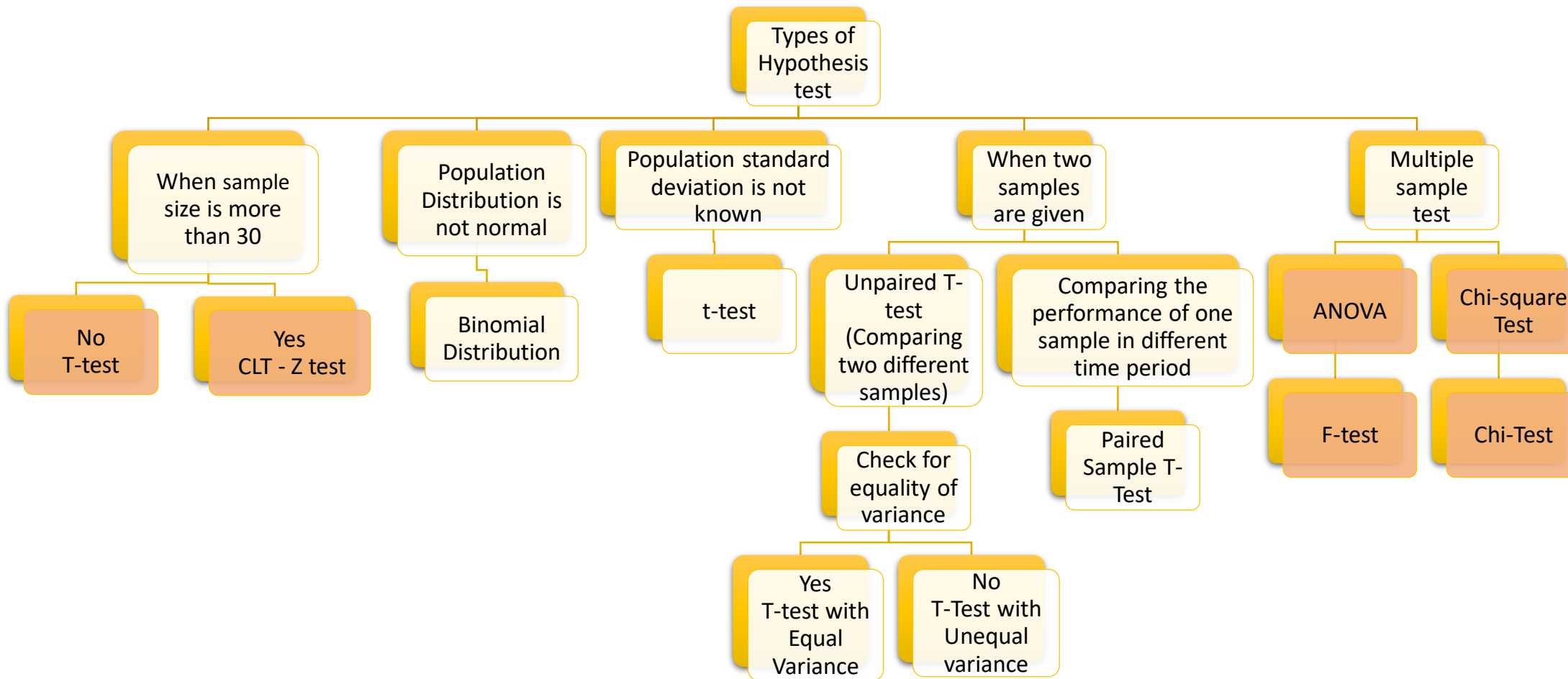
Types of Hypothesis Test

```
                          Types of
                          Hypothesis
                          test
    ┌──────────────┬──────────────┼──────────────────┬───────────────────┐
When sample      Population    Population standard   When two          Multiple
size is more     Distribution is  deviation is not   samples           sample
than 30          not normal    known                 are given         test
  ┌────┴────┐        │              │            ┌──────┴────────┐    ┌──────┴──────┐
No         Yes     Binomial       t-test    Unpaired T-    Comparing the  ANOVA    Chi-square
T-test     CLT - Z test Distribution          test       performance of one        Test
                                            (Comparing   sample in different
                                            two different time period
                                            samples)           │            │          │
                                                │          Paired        F-test     Chi-Test
                                            Check for      Sample T-
                                            equality of    Test
                                            variance
                                          ┌────┴────┐
                                        Yes       No
                                        T-test with T-Test with
                                        Equal      Unequal
                                        Variance   variance
```

Sample Sizes are Small

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

- A hypothesis test for a population mean that involves the *t*-distribution is called a *t*-test. The formula for the test statistic in this case is:

$$t_{n-1} = \frac{\overline{x} - \mu_0}{s / \sqrt{n}},$$

- where t_{n-1} is a value from the *t*-distribution with $n-1$ degrees of freedom.

- As the sample size increases and approaches to 30, the t-test approximates a normal distribution
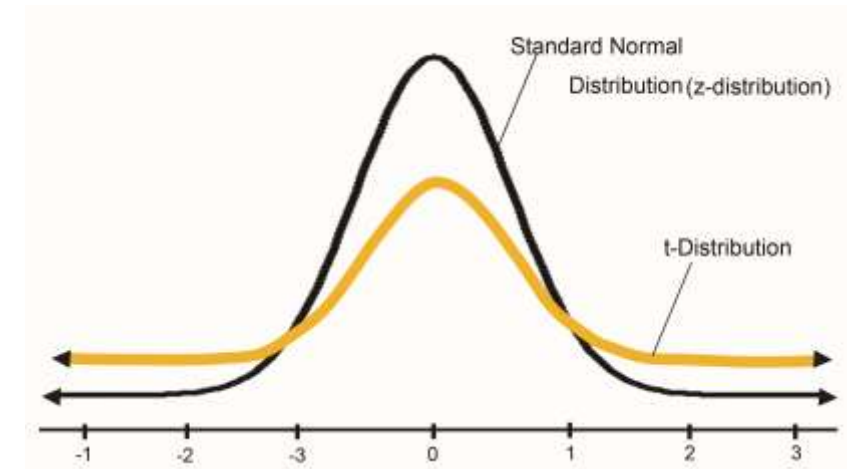
Degrees of Freedom in t-test

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 European people is same as the height of Global Population

Alternate Hypothesis
Height of European people is less than the height of Global Population

Considering 5% significance level

Test Statistic:
X = Mean of sample
U = Mean of population
S = Sample Standard Deviation (if sample Sd is not given, then you can find the SD of samples by using stdev in Excel)
N= Sample size

Hypothesis testing

Solution of Example

SD of 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 European 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

$\alpha = 0.05$

T = 4.13

0   2.145

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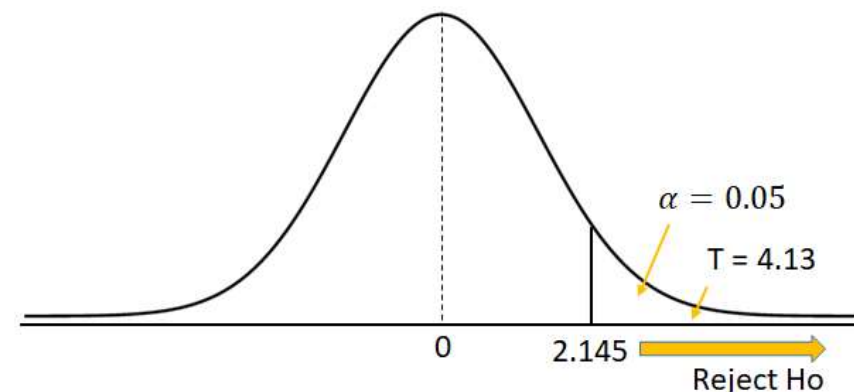
Instead of using t-test for hypothesis testing, we can directly use p-value in Excel

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

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

- The distribution of sample means will follow T $(\mu, \frac{s}{\sqrt{n}})$

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- T.DIST(2.37,39, TRUE) =1- 0.99 = 0.011

Conclusion: Reject null hypothesis

ANOVA

ANOVA

- Adidas 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 (Ho)**: The mean sales for each cities are same.
  - ✓ **Alternate Hypothesis (Ha)**: 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:

Delhi	Mumbai	Bangalore
1152	1123	1234
1275	1278	1195
1190	1152	1114
1234	1245	1298
1305	1348	1322

Delhi	Mumbai	Bangalore
725	1302	1469
668	1148	1987
847	1130	1845
716	1274	1624
690	1112	1472

Mean	1231.2	1229.2	1232.6

Mean	729	1193	1679
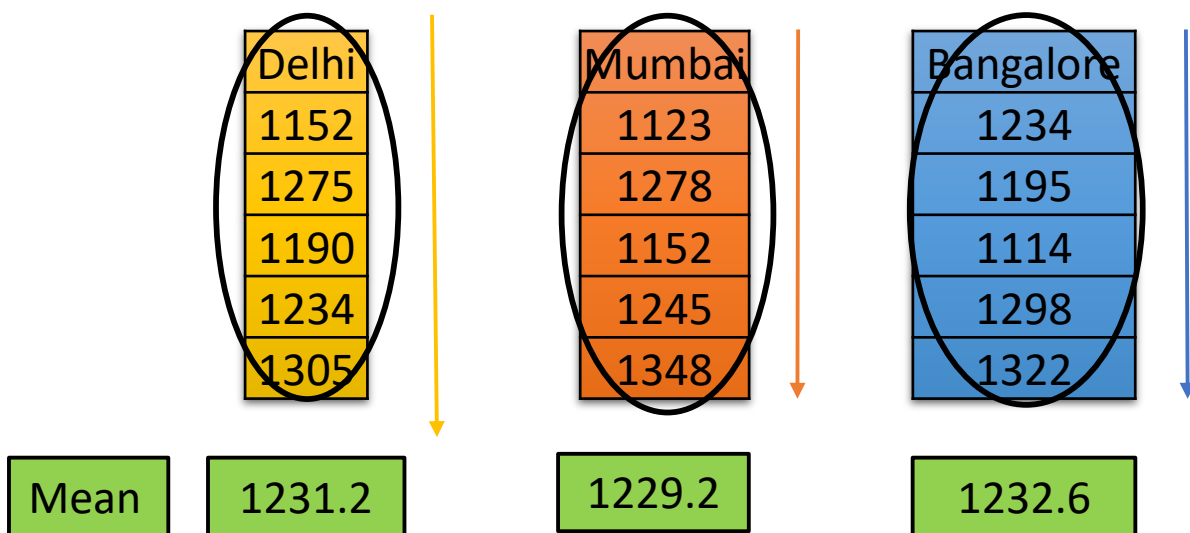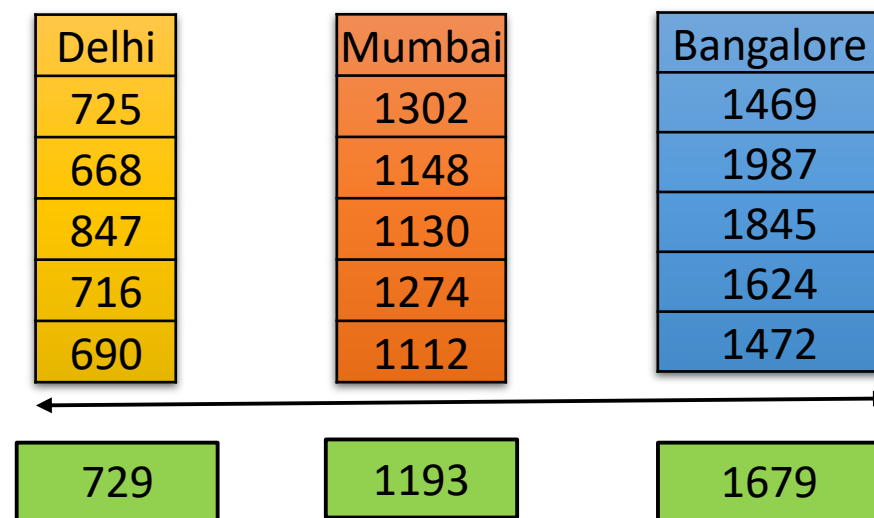
F - Test

- 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.

$$F = \frac{(\text{Mean square column}) \text{ or } \left(\dfrac{SSC}{\text{Degrees of freedom}}\right)}{(\text{Mean Square error}) \text{ or } \left(\dfrac{SSE}{\text{Degrees of freedom}}\right)}$$

- The numerator degrees of freedom is k-1 i.e. samples - 1
- The numerator degrees of freedom is n-k i.e. observations – samples.
- The larger the ratio, the more likely it is the groups have different means.

- Lets understand the step by step process of finding one way ANOVA in the coming slides

Partition of Total Variation

Total Variation
(SST)

Variance between samples(columns)

(SSC)

Variance within samples(columns)

(SSE)
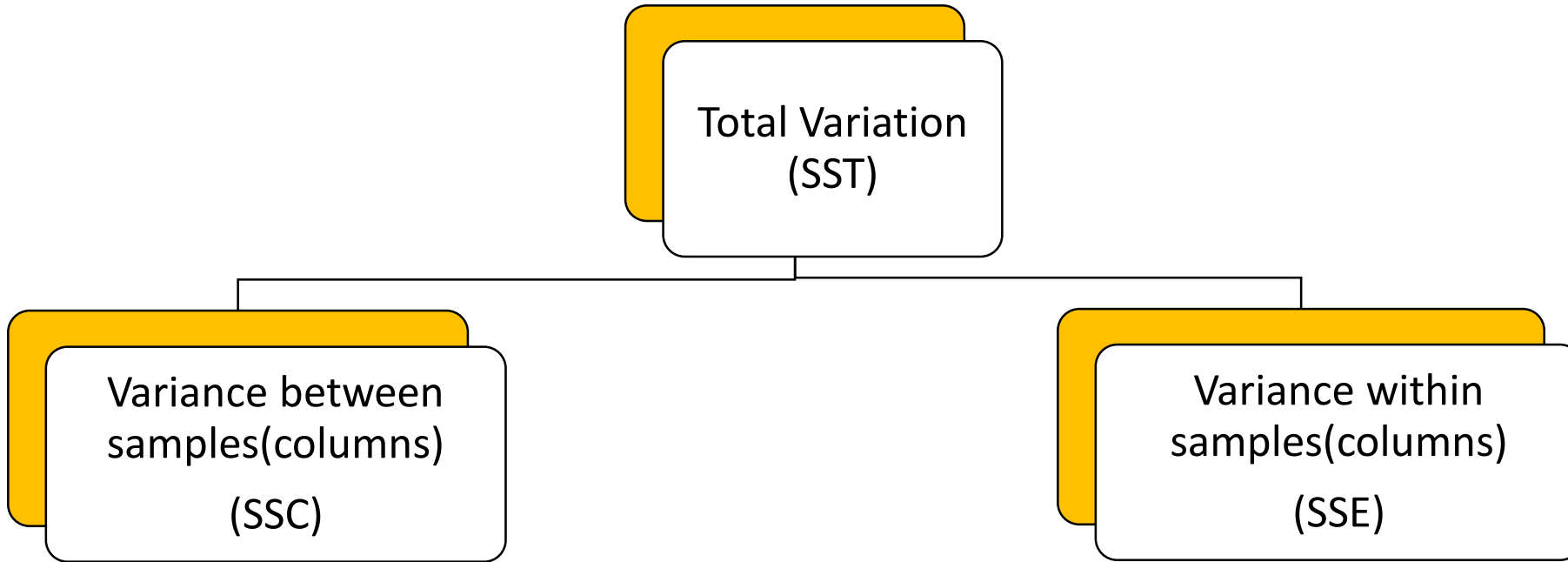
Commonly Referred to as:
Sum of Squares Between
Sum of Squares Among
Sum of Squares Explained
Among Groups variation

Commonly Referred to as:
Sum of Squares Within
Sum of Squares Error
Sum of Squares Unexplained
Within Groups 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

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

+

Delhi
980
776
923
1498
999

Mumbai
1123
1357
1152
921
959

Bangalore
1084
1025
1114
1182
1022

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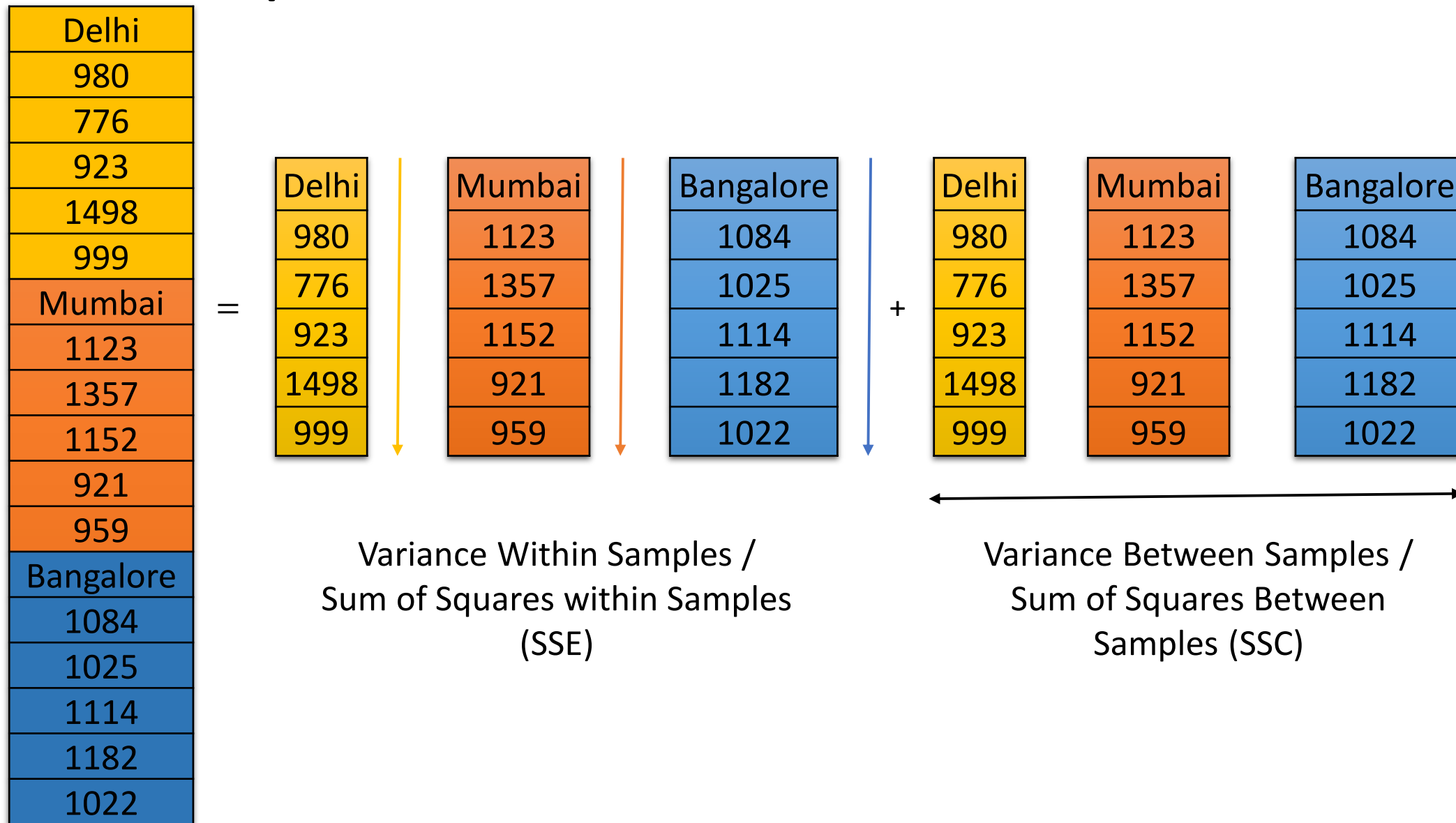
Variance Within Samples (SSE)

- 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:
1. Calculate mean of each sample
1. Calculate a Grand mean
2. Calculate the difference between the mean of each sample and grand mean
3. Square the difference between the mean of each sample and grand mean
4. Calculate the sum of each sample mean
5. 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 method 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.

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

- ✓ 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

$$= 3 - 1 = 2$$

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

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

F Ratio

MSE

✓ The Mean square error is the ratio of variance within sample to the (n-k) degrees of freedom.

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

✓ 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
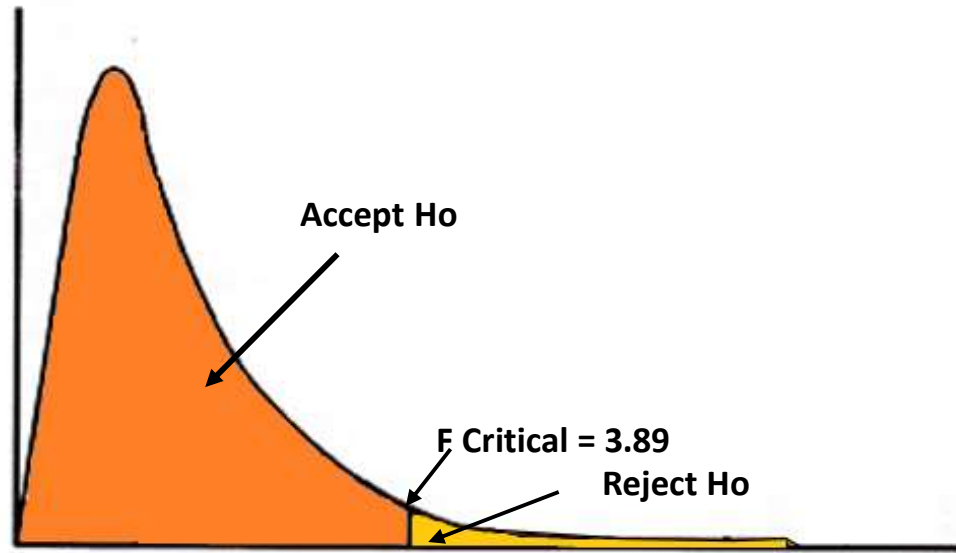 $$= 15 - 3 = 12$$

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

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

F Ratio

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

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

Accept Ho

F Critical = 3.89

Reject Ho

ν_2 \ ν_1	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.

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.

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 Test

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

Expected Transaction

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

Observed transaction

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

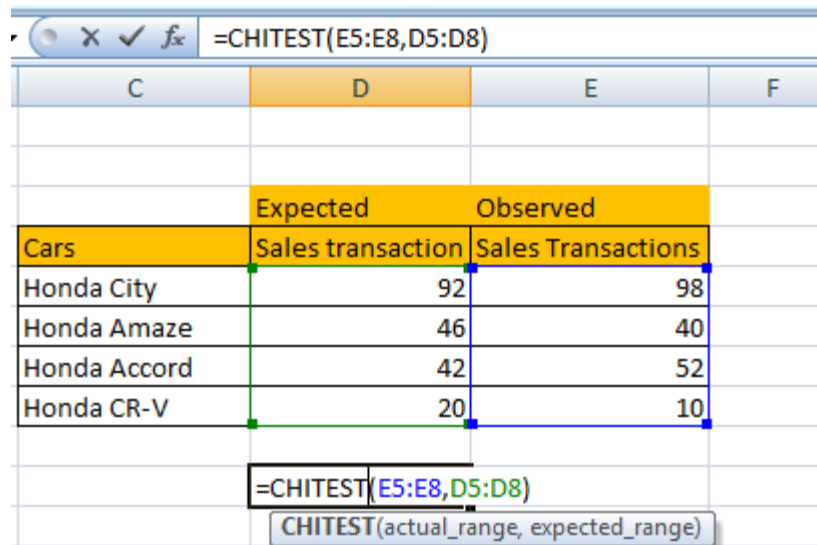**Alternate Hypothesis: There is a difference between the overall sales transaction and the new showroom sales.**

Example of Chi Square test

Cars	Sales	Sales transaction
Honda City	92	46%
Honda Amaze	46	23%
Honda Accord	42	21%
Honda CR-V	20	10%

Cars	Sales	Sales transaction
Honda City	98	49%
Honda Amaze	40	20%
Honda Accord	52	26%
Honda CR-V	10	5%

Expected Transaction(Comparing with 200)

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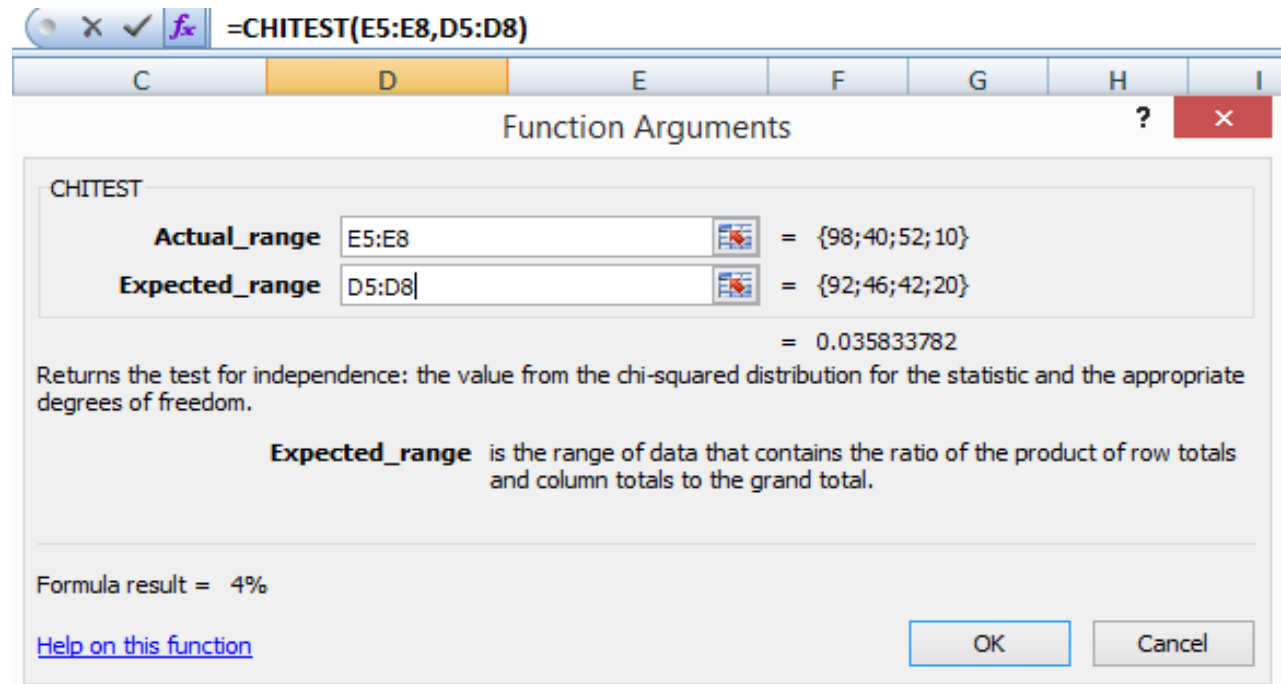
Hence, we reject the Null hypothesis and conclude that the sales transaction of the new showroom is different than the overall transaction percentage.

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 Distribution

IF x1, x2,……Xn are independent normal variates and each is distributed normally with mean zero an standard deviation unity, then $X1^2 + X2^2 + \cdots + Xn^2 = \sum Xi$ Is distributed as chi square with n degrees of freedom (d.f.) Where n is large. The chi square curve for d.f. N = 2,4 and 10 is as follows.

If degree of freedom >2 : distribution is bell shaped
If degree of freedom = 2 : Distribution is L Shaped with maximum ordinate at zero
Id degree of freedom <2 (>0) : Distribution L shaped with infinite ordinate at the origin.
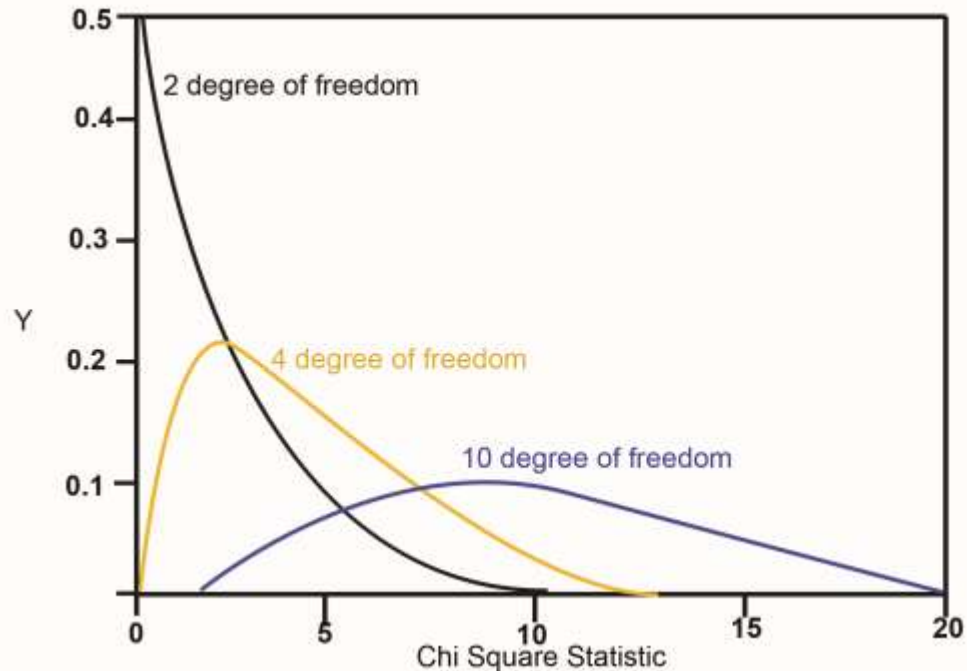
Chi-Square test Formulae,

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The subscript "c" are the degrees of freedom. "O" is your observed value and E is your expected value.

Test of Association

Chi Square test Example– Test of Association

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

Observed Sample				
Car Type	Urban	Semi Urban	Rural	Total
Honda City	74	52	48	174
Honda Accord	29	21	12	62
Honda CRV	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 ith row and the jth column, T_i is the total number of subjects in the ith row, T_j is the total number of subjects in the jth 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
Honda City	64	57	53	174
Honda Accord	23	20	19	62
Honda CRV	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.

| TILE | ▾ | × ✓ f_x | =CHITEST(C5:E7,I5:K7) |

	B	C	D	E	F	G	H	I	J	K	L
	Observed Sample						**Expected Values**				
	Car Type	Urban	Semi Urban	Rural	Total		Car Type	Urban	Semi Urban	Rural	Total
	Honda City	74	52	48	174		Honda City	64	57	53	174
	Honda Accord	29	21	12	62		Honda Accord	23	20	19	62
	Honda CRV	17	34	39	90		Honda CRV	33	30	27	90
	Total	120	107	99	326		Total	120	107	99	326

=CHITEST(C5:E7,I5:K7)
CHITEST(actual_range, expected_range)

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 E-commerce 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.

Calculating the probability, by using binomial distribution

Chi Square Test- Goodness of Fit Example

	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(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.

Exercise for Practice

**Exercise for Basic
Stat 3**

Thank You