Hypothesis

In our daily life, we often hear statements like Dhoni is the better captain than his contemporaries, Or Motor cycle company claiming that a certain model gives an average mileage of 100Km per liter or Tooth paste company claiming to be the number one brand suggested by dentists.

Let's suppose you have to purchase a motorcycle and you heard about the above claim made by the Motor cycle company. Would you just go and buy it or rather look for the proof of it? There must be a parameter based on which one would judge the correctness of statement made. In this case our parameter will be the Average mileage, which you will use to check if statement made is actually true or just a hoax.

Hypothesis is a statement, assumption or claim about the value of the parameter (mean, variance, median etc.).

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

Like, if we make a statement that "Dhoni is the best Indian Captain ever." This is an assumption that we are making based on the average wins and looses team had under his captaincy. We can test this statement based on all the match data.

Simple and Composite Hypothesis

When a hypothesis specifies an exact value of the parameter, it is a simple hypothesis and if it specifies a range of values then it is called a composite hypothesis.

e.g. Motor cycle company claiming that a certain model gives an average mileage of 100Km per liter, this is a case of simple hypothesis.

Average age of students in a class is greater than 20. This statement is a composite hypothesis.

Null Hypothesis

The null hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty We assume innocence until we have enough evidence to prove that a suspect is guilty.

It is denoted by H0.

Alternate Hypothesis

The alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter.

It is denoted by H1.

Let's understand this with an example:

A soap company claims that it's product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis.

Null Hypothesis(H0): Average =99%

Alternate Hypothesis(H1): Average is not equal to 99%.

Note: The thumb rule is that statement containing equality is the null hypothesis.

Hypothesis Testing

When we test a hypothesis, we assume the null hypothesis to be true until there is sufficient evidence in the sample to prove it false. In that case we reject the null hypothesis and support the alternate hypothesis.

If the sample fails to provide sufficient evidence for us to reject the null hypothesis, we cannot say that the null hypothesis is true because it is based on just the sample data. For saying the null hypothesis is true we will have to study the whole population data.

One Tailed and Two Tailed Tests

If the alternate hypothesis gives the alternate in both directions (less than and greater than) of the value of the parameter specified in null hypothesis, it is called **Two tailed test**.

If the alternate hypothesis gives the alternate in only one direction (either less than or greater than) of the value of the parameter specified in null hypothesis, it is called **One tailed test**.

e.g. if H0: mean= 100 H1: mean not equal to 100

here according to H1, mean can be greater than or less than 100. This is an example of Two tailed test

Similarly, if H0: mean>=100 then H1: mean< 100

Here, mean is less than 100, it is called One tailed test.

Critical Region

The critical region is that region in the sample space in which if the calculated value lies then we reject the null hypothesis.

Let's understand this with an example:

Suppose you are looking to rent an apartment. You listed out all the available apartments from different real state websites. You have budget of Rs. 15000/ month. You cannot spend more than that. The list of apartments you have made have price ranging from 7000/month to 30,000/month.

You select a random apartment from the list and assume below hypothesis:

H0: You will rent the apartment.

H1: You won't rent the apartment.

Now, since your budget is 15000, you have to reject all the apartments above that price.

Here all the Prices greater than 15000 becomes your critical region. If the random apartment's price lies in this region, you have to reject your null hypothesis and if the random apartment's price doesn't lie in this region, you do not reject your null hypothesis.

The critical region lies in one tail or two tails on the probability distribution curve according to the alternative hypothesis. Critical region is pre-defined area corresponding to a cut off value in probability distribution curve. It is denoted by **α.**

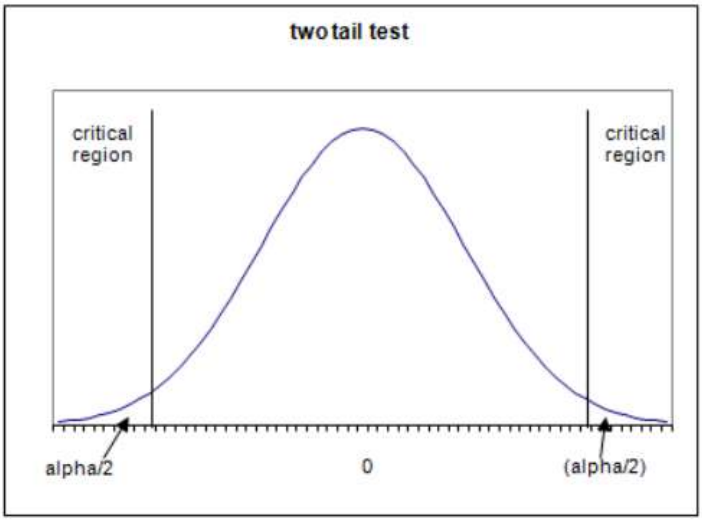
Critical values are values separating the values that support or reject the null hypothesis and are calculated on the basis of alpha.

We will see more examples later on and it will be clear how do we choose **α.**

Based on the alternative hypothesis, three cases of critical region arise:

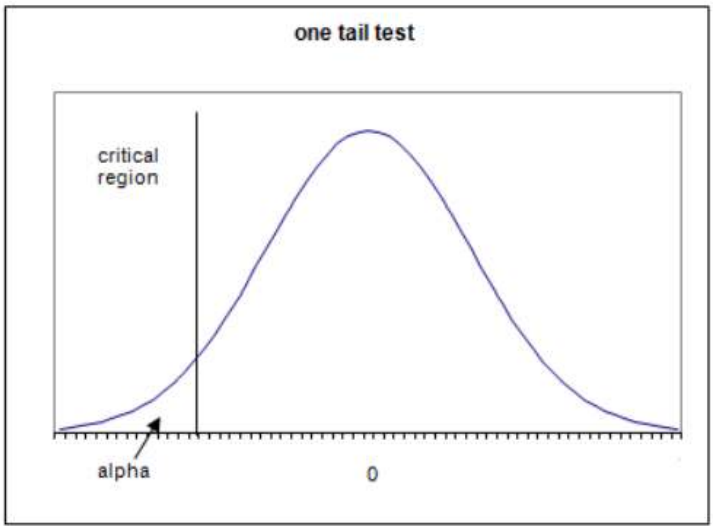
Case 1) This is double tailed test.

$\mathbf{H_0}: \mu = \mu_0$ $\mathbf{H_1}: \mu \neq \mu_0;$



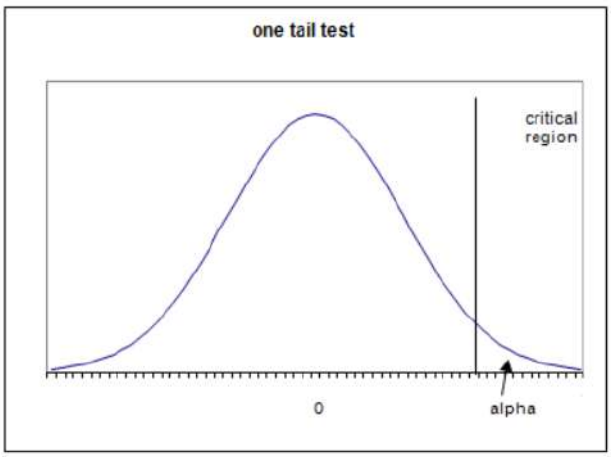
Case 2) This scenario is also called Left-tailed test.

$\mathbf{H_0}: \mu = \mu_0$ $\mathbf{H_1}: \mu < \mu_0;$



Case 3) This scenario is also called Right-tailed test.

$\mathbf{H_0}: \mu = \mu_0 \qquad \mathbf{H_1}: \mu > \mu_0;$



Type I and Type II Error

Decision	H0 True	H0 False
Reject H0	Type I error	Correct Decision
Do not reject H0	Correct Decision	Type II error

A false positive (type I error) — when you reject a true null hypothesis.

A false negative (type II error) — when you accept a false null hypothesis.

The probability of committing Type I error (False positive) is equal to the significance level or size of critical region α.

α= P [rejecting H0 when H0 is true]

The probability of committing Type II error (False negative) is equal to the beta β and is called 'power of the test'.

β = P [not rejecting H0 when h1 is true]

Example:

Person is arrested on the charge of being guilty of burglary. A jury of judges has to decide guilty or not guilty.

H0: Person is innocent

H1: Person is guilty

Type I error will be if the Jury convicts the person [rejects H0] although the person was innocent [H0 is true].

Type II error will be the case when Jury released the person [Do not reject H0] although the person is guilty [H1 is true].

Level of Significance(α) :

It is the probability of type 1 error. It is also the size of the critical region.

Generally, a strong control on α is desired and in tests it is pre fixed at very low levels like 0.05(5%) or 01(1%).

If H0 is not rejected at a significance level of 5%, then one can say that our null hypothesis is true with 95% assurance.

Steps involved in Hypothesis testing

1) Setup the null hypothesis and the alternate hypothesis.
2) Decide a level of significance i.e. alpha = 5% or 1%
3) Choose the type of test you want to perform as per the sample data (z test, t test, chi squared etc.) (we will study all the tests in next section)
4) Calculate the test statistics (z-score, t-score etc.) using the respective formula of test chosen
5) Obtain the critical value for in the sampling distribution to construct the rejection region of size alpha using z-table, t-table, chi table etc.
6) Compare the test statistics with the critical value and locate the position of the calculated test statistics i.e. is it in rejection region or non-rejection region.
7) **I)** If the critical value lies in the rejection region, we will reject the hypothesis i.e. sample data provides sufficient evidence against the null hypothesis and there is significant difference between hypothesized value and observed value of the parameter.
   II) If the critical value lies in the non- rejection region, we will not reject the hypothesis i.e. sample data does not provide sufficient evidence against the null hypothesis and the difference

between hypothesized value and observed value of the parameter is due to fluctuation of the sample.

p-value

Let's suppose we are conducting a hypothesis test at a significance level of 1%.

Where, **H0: mean<X** (we are just assuming a scenario of 1 tail test.)

We obtain our critical value (based on the type of test we are using) and find that our test statistics is greater than the critical value. So, we have to reject the null hypothesis here since it lies in the rejection region. Now if the null hypothesis is getting rejected at 1%, then for sure it will get rejected at the higher values of significance level, say 5% or 10%.

What if we take significance level lower than 1%, would we have to reject our hypothesis then also?

Yes, there might be a chance that the above scenario can happen and here comes "p-value" in play.

p-value is the smallest level of significance at which a null hypothesis can be rejected.

That's why many tests now a days gives p-value and it is more preferred since it gives out more information than the critical value.

For right tailed test:

p-value = P[Test statistics >= observed value of the test statistic]

For left tailed test:

p-value = P[Test statistics <= observed value of the test statistic]

For two tailed test:

p-value = 2 * P[Test statistics >= |observed value of the test statistic|]

Decision making with p-value

The p-value is compared to the significance level(alpha) for decision making on null hypothesis.

If p-value is **greater** than alpha, we **do not reject** the null hypothesis.

If p-value is **smaller** than alpha, we **reject** the null hypothesis.

Confidence Intervals

A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level.

Calculating a Confidence Interval (Theory)

Suppose a group of researchers is studying the heights of high school basketball players. The researchers take a random sample from the population and establish a mean height of 74 inches. The mean of 74 inches is a point estimate of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far away this 74-inch sample mean might be from the population mean. What's missing is the degree of uncertainty in this single sample.

Confidence intervals provide more information than point estimates. By establishing a 95% confidence interval using the sample's mean and standard deviation, and assuming a normal distribution as represented by the bell curve, the researchers arrive at an upper and lower bound that contains the true mean 95% of the time. Assume the interval is between 72 inches and 76 inches. If the researchers take 100 random samples from the population of high school basketball players as a whole, the mean should fall between 72 and 76 inches in 95 of those samples.

If the researchers want even greater confidence, they can expand the interval to 99% confidence. Doing so invariably creates a broader range, as it makes room for a greater number of sample means. If they establish the 99% confidence interval as being between 70 inches and 78 inches, they can expect 99 of 100 samples evaluated to contain a mean value between these numbers. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter. Likewise, a 99% confidence level means that 95% of the intervals would include the parameter.

The Confidence Interval is based on Mean and Standard Deviation and is given as:

For n>30

Confidence interval = $X \pm (z * s/\sqrt{n})$

where z critical value is derived from the z score table based on the confidence level.

X is the sample mean.

s is sample standard deviation.

n is the sample size

Confidence Level	z- value
80%	1.28
85%	1.44
90%	1.64
95%	1.96
98%	2.33
99%	2.58

We obtain these values from the z-score table only, but since the confidence levels are most of the times fixed as the above values, so we can use this table.

For n<30

Confidence interval = $X \pm (t * s/\sqrt{n})$

where t critical value is derived from the t score table based on the confidence level.

X is the sample mean.

s is sample standard deviation.

n is the sample size.

We will see how to create confidence intervals in the examples to follow.

Now that we have got all the theory behind Hypothesis testing, let's see different types of tests that are used for testing. We have already seen examples on finding z-score and t-score, we will see how they are used in the testing scenario.

General points for selection type of tests:

sample size	Population Variance	Normality of Sample	Sample variance	Type of Test
Large (>30)	Known	Normal/Non-Normal		Z-test
Large (>30)	Unknown	Normal	Use this to calculate t-score	t-test
Large (>30)	Unknown	Unknown	Use this to calculate z-score	Z-test
Small (<30)	Known	Normal		Z-test
Small (<30)	Unknown	Normal	Use this to calculate t-score	t-test

Note: We will learn about other non-parametric tests and their cases later

Hypothesis Testing for Large Size Samples

Thumb rule: A sample of size greater than 30 is considered a large sample and as per central limit theorem we will assume that all sampling distributions follows a normal distribution.

We are familiar with the steps of hypothesis testing as shown earlier. We also know, from the above table, when to use which type of test.

Let's start with few practical examples to help our understanding more.

Note: We have learned in previous section how to use the z-score table to calculate probabilities, in this section we have some standard Significance level for which we need to find the critical value(z-score). So instead of going through the whole table, we will just use the below standardized critical value table for calculation purposes.

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
.50	0.00	.050	1.64	.030	1.88	.020	2.05	.010	2.33
.45	0.13	.048	1.66	.029	1.90	.019	2.07	.009	2.37
.40	0.25	.046	1.68	.028	1.91	.018	2.10	.008	2.41
.35	0.39	.044	1.71	.027	1.93	.017	2.12	.007	2.46
.30	0.52	.042	1.73	.026	1.94	.016	2.14	.006	2.51
.25	0.67	.040	1.75	.025	1.96	.015	2.17	.005	2.58
.20	0.84	.038	1.77	.024	1.98	.014	2.20	.004	2.65
.15	1.04	.036	1.80	.023	2.00	.013	2.23	.003	2.75
.10	1.28	.034	1.83	.022	2.01	.012	2.26	.002	2.88
.05	1.64	.032	1.85	.021	2.03	.011	2.29	.001	3.09

Q) A manufacturer of printer cartridge clams that a certain cartridge manufactured by him has a mean printing capacity of at least 500 pages. A wholesale purchaser selects a sample of 100 printers and tests them. The mean printing capacity of the sample came out to be 490 pages with a standard deviation of 30 printing pages.

Should the purchaser reject the claim of the manufacturer at a significance level of 5%?

Ans. population mean = 500

Sample mean = 490

Sample standard deviation = 30

Significance level(alpha) = 5% = 0.05

Sample size = 100

H0: Mean printing capacity >=500

H1: Mean printing capacity < 500

We can clearly see it is one tailed test (left tail).

Here, the sample is large with an unknown population variance. Since, we don't know about the normality of the data, we will use the Z-test (from the table above).

We will use the sample variance to calculate the critical value.

Standard error (SE) = Sample standard deviation/ (sample size) * 0.5

= 30 / (100) *0.5 = 3

Z(test) = (Sample mean - population mean)/ (SE)

= (490-500)/3 = -3.33

Let's find out the critical value at 5% significance level using the above Critical value table.

Z (0.05%) = - 1.645 (since it is left tailed test).

We can clearly see that Z(test) < Z (0.05%), that means our test value lies in the rejection region.

Thus, we can reject the null hypothesis i.e. the manufacturer's claim at 5% significance level.

Using p-value to test the above hypothesis:

p-value = P[T<=-3.33] (we know p(-x) = 1 -p(x) also, remember that the p(x) represents the cumulative probability from 0 to x)

let's use z-table to find the p-value:

2.8	0.9974	0.9975	0.9976	0.9977	0.9977
2.9	0.9981	0.9982	0.9982	0.9983	0.9984
3.0	0.9987	0.9987	0.9987	0.9988	0.9988
3.1	0.9990	0.9991	0.9991	0.9991	0.9992
3.2	0.9993	0.9993	0.9994	0.9994	0.9994
3.3	0.9995	0.9995	0.9995	0.9996	0.9996

p-value = 1 – 0.9996 = 0.0004

Here, p-value is less than the significance level of 5%. So, we are right to reject the null hypothesis.

Q) A company used a specific brand of Tube lights in the past which has an average life of 1000 hours. A new brand has approached the company with new Tube lights with same power at a lower price. A sample of 120 light bulbs were taken for testing which yielded an average of 1100 hours with standard deviation of 90 hours. Should the company give the contract to this new company at a 1% significance level.

Also, find the confidence interval.

Ans. Population mean = 1000

Sample mean = 1010

Significance level = 1% = 0.01

Sample size = 120

Sample standard deviation = 90

H0: average life of tube lights >= 1000

H1: average life of tube lights < 1000

Here, the sample is large with an unknown population variance. Since, we don't know about the normality of the data, we will use the Z-test (from the table above).

Standard error (SE) = Sample standard deviation/ (sample size) * 0.5

= 90 / (120) *0.5 = 8.22

Z(test) = (Sample mean - population mean)/ (SE)

= (1010-1000)/8.22 = 1.22

Let's find out the critical value at 1% significance level using the above Critical value table.

Z (0.01%) = -2.33(since it is left tailed test).

We can clearly see that Z(test) >Z (0.01%), that means our test value doesn't lie in the rejection region.

Thus, we cannot reject the null hypothesis i.e. the company can give the contract at 1% significance level.

Using p-value to test the above hypothesis:

p-value = P[T<1.22]

z	0.00	0.01	0.02
0.0	0.5000	0.5040	0.5080
0.1	0.5398	0.5438	0.5478
0.2	0.5793	0.5832	0.5871
0.3	0.6179	0.6217	0.6255
0.4	0.6554	0.6591	0.6628
0.5	0.6915	0.6950	0.6985
0.6	0.7257	0.7291	0.7324
0.7	0.7580	0.7611	0.7642
0.8	0.7881	0.7910	0.7939
0.9	0.8159	0.8186	0.8212
1.0	0.8413	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888

p-value = 0.88

Here, p-value is greater than the significance level of 1%. So, we do not reject the null hypothesis.

Comparing two population samples mean using Z test

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $X_{1(\text{mean})}$ - $X_{2(\text{mean})}$, and divide by the standard error (shown below) in order to standardize the difference.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation i.e. standard error.

The standard error (SE) is:

$$\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}$$

Z is given as :

$$\frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}}$$

In this comarison case, our null assumpiton is that μ(1) = μ(2)

So, Z becomes = X1(mean)- X2(mean)/ (SE)

Q) In two samples of men from two different states A and B , the height of of 1000 men and 2000 men respectively are 76.5 and 77 inches. If population standard deviation for both states is same and is 7 inches, can we assume that mean hieghts of both sates can be regarde same at 5% level of significance.

Ans. n1 = 1000

n2 = 2000

X1(mean) = 76.5

X2(mean) = 77

S1=S2= 7

Let's μ(1) = μ(2) be the mean heights of men from states A and B

H0: μ(1) = μ(2)

H1: μ(1) is not equal to μ(2)

Standard error(SE) = [((S1)^2/n1)+((S2)^2/n2)]^0.5 = 0.27

Z(test) = X1(mean)- X2(mean)/ (SE) = (76.5-77)/0.27 = -1.85

Since, it is a two tailed test, we need to find critical value for 2.5% on each tail.

Z(2.5%) = 1.96 and Z(-2.5%) = -1.96

We can clearly see, Z(-2.5%) < Z(test) <Z(2.5%)

Thus, we cannot reject the null hypothesis.

Using p-value

p-value = 2* P[Z>=|-1.85|] = 2 * P[Z>=-1.85]

z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842

p-value = 2 * (1- 0.9678) (since we want z> 1.85) = 0.0644

We can clearly see, p-value is greater than 0.05% ,thus we cannot reject the null hypothesis.

Hypothesis Testing for Small Size Samples

In real world scenarios, large sample sizes are possible most of times because of the limited resources such as money. We generally do hypothesis testing based on small samples, only assumption being the normality of the sample data.

We will see how to use t- tests in this section and how to use the t-score table (continued from the topic of student t's distribution).

All the steps involved are similar to the z-test, only we will calculate t-score instead of z-score.

Let's start with an example:

Q) A tyre manufacturer claims that the average life of a particular category of its tyre is 18000km when used under normal driving conditions. A random sample of 16 tyres was tested. The mean and SD of life of the tyres in the sample were 20000 km and 6000 km respectively.

Assuming that the life of the tyres is normally distributed, test the claim of the manufacture at 1% level of significance.

Construct the confidence interval also.

Ans: population mean = 18000 km

Sample mean = 20000 km

Standard deviation = 6000 km

Sample size = 16

H0: population mean = 18000km

H1: population mean is not equal to 18000km (It will be a two tailed test.)

Since sample size is small, population variance is unknown and the sample is normally distributed, we will used t-test for this.

Standard error = [6000/(16)^0.5] = 1500

t-score(test) = (20000 - 18000)/1500 = 1.33

Let's find out the critical t- value, for significance level 1% (two tailed) and degree of freedom = 16-1 = 15

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002
df										
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646

t(0.005) = 2.947 and t(-0.005) = -2.947

We can see that, t (- 0.005) < t-score(test) = 1.33 < t (0.005)

So, the value lies in non-rejection region and we cannot reject our null hypothesis.

Using the p-value

p-value = P[t>|1.33|]

degree of freedom = 15

let's see the p-value from the table for the above values:

***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	
one-tail	0.50	0.25	0.20	0.15	0.10	
two-tails	1.00	0.50	0.40	0.30	0.20	
df						
1	0.000	1.000	1.376	1.963	3.078	6
2	0.000	0.816	1.061	1.386	1.886	2
3	0.000	0.765	0.978	1.250	1.638	2
4	0.000	0.741	0.941	1.190	1.533	2
5	0.000	0.727	0.920	1.156	1.476	2
6	0.000	0.718	0.906	1.134	1.440	1
7	0.000	0.711	0.896	1.119	1.415	1
8	0.000	0.706	0.889	1.108	1.397	1
9	0.000	0.703	0.883	1.100	1.383	1
10	0.000	0.700	0.879	1.093	1.372	1
11	0.000	0.697	0.876	1.088	1.363	1
12	0.000	0.695	0.873	1.083	1.356	1
13	0.000	0.694	0.870	1.079	1.350	1
14	0.000	0.692	0.868	1.076	1.345	1
15	0.000	0.691	0.866	1.074	1.341	1
16	0.000	0.690	0.865	1.071	1.337	1
17	0.000	0.689	0.863	1.069	1.333	1

from the table we can see: $0.20 < p < 0.30$

Here, p > significance level (1%), thus we cannot reject the null hypothesis.

Confidence interval = [20000 – 2.47*1500 , 20000 +2.47*1500]

= [16295, 23705]

Comparing two population samples mean using t test

Just like the case we saw with z-test, t-test is actually more suitable for comparison of two populations samples because in practice population standard deviations for both populations are not always known.

We assume a normal distribution of samples and though the population standard deviations are unknown, we assume them to be equal.

Also, samples are independent to each other.

Let's assume two independent samples with size n1 and n2:

Degree of freedom = n1 + n2 -2

Standard Error(SE):

$$\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}$$

Variance(Sample) = (∑[X-X(mean)]^2 + ∑[Y-Y(mean)]^2))/(n1 + n2 -2)

Test statistic t in this case is given as:

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2}{SE_{\overline{X}_1 - \overline{X}_2}}$$

Q) The means of two random samples of sizes 10 and 8 from two normal popultaion is 210.40 and 208.92. The sum of sqaures od deviation from their means is 26.94 and 24.50 respectively.Assuming population with equal variances, can we consider the normal populatiojns have equal mean?(Significance level =5%)

Ans.

n1 =10 , n2= 8 , X(mean) = 210.40 , Y(mean) = 208.92

std. Deviation(sample) =[(26.94 + 24.50)/(10 + 8 - 2)]^0.5 = 1.79

H0: Population means are equal

H1: Population means are not equal (two tailed test)

Standard errror = 1.79 * (1/10 + 1/8)^0.5 = 0.84

t(test) = X(mean)- Y(mean)/0.84 = 1.48/.84 = 1.76

Degree of freedom = 10 +8 -2 = 16

Let's look for critical value in the t-table for significance 5%(two tailed) and d.o.f 16:

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.

t(0.005) = 2.120 and t(-0.005) = -2.120

We can see that, t (- 0.005) < t-score(test) = 1.76 < t (0.005)

So, the value lies in non-rejection region and we cannot reject our null hypothesis.

Paired Sample t-Tests

A paired t-test is used to compare two population means where you have two samples which are not independent e.g. Observations recorded on a patient before and after taking medicine, weight of a person before and after they started working out etc.

Now, instead of two separate populations, we create a new column with difference of the populations, and instead of testing equality of two population mean we test the hypothesis that mean of the population difference is zero. Also, we assume the samples are of same size. Population variances are not known and not necessarily equal.

Standard error = Deviation of differences/(n^0.5)

t= D(mean)/ standard error, where D(mean) is the men of the differences.

Q) A group 20 students were tested to see how many of them have improved marks after a special lecture on the subject.

marks before the lecture	marks after the lecture	Difference(D)	(D-mean)^2
18	22	4	3.24
21	25	4	3.24
16	17	1	1.44
22	24	2	0.04
19	15	-4	38.44
24	26	2	0.04
17	20	3	0.64
21	23	2	0.04
13	18	5	7.84
18	20	2	0.04
15	15	0	4.84
16	15	-1	10.24
18	21	3	0.64
14	16	2	0.04
19	22	3	0.64
20	24	4	3.24

12	18	6	14.44
22	25	3	0.64
14	18	4	3.24
19	18	-1	10.24
		44	103.2
		Difference mean = 2.2	5.43157895
		Standard Deviation	2.33057481

H0: Difference mean >= 0

H1: Difference mean < 0

Standard error = 2.33 /(20)^0.5 = 0.52

t= 2.2 / 0.52 = 4.23

D.o.f = 19

On significance level 5%. 19 d.o.f and a one tail test, let's calculate our critical level:

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05
two-tails	1.00	0.50	0.40	0.30	0.20	0.10
df						
1	0.000	1.000	1.376	1.963	3.078	6.314
2	0.000	0.816	1.061	1.386	1.886	2.920
3	0.000	0.765	0.978	1.250	1.638	2.353
4	0.000	0.741	0.941	1.190	1.533	2.132
5	0.000	0.727	0.920	1.156	1.476	2.015
6	0.000	0.718	0.906	1.134	1.440	1.943
7	0.000	0.711	0.896	1.119	1.415	1.895
8	0.000	0.706	0.889	1.108	1.397	1.860
9	0.000	0.703	0.883	1.100	1.383	1.833
10	0.000	0.700	0.879	1.093	1.372	1.812
11	0.000	0.697	0.876	1.088	1.363	1.796
12	0.000	0.695	0.873	1.083	1.356	1.782
13	0.000	0.694	0.870	1.079	1.350	1.771
14	0.000	0.692	0.868	1.076	1.345	1.761
15	0.000	0.691	0.866	1.074	1.341	1.753
16	0.000	0.690	0.865	1.071	1.337	1.746
17	0.000	0.689	0.863	1.069	1.333	1.740
18	0.000	0.688	0.862	1.067	1.330	1.734
19	0.000	0.688	0.861	1.066	1.328	1.729
20	0.000	0.687	0.860	1.064	1.325	1.725
21	0.000	0.686	0.859	1.063	1.323	1.721

t(5%) = -1.729

Since, t is greater than critical t, thus it lies in non-rejection region and hence we cannot reject the null hypothesis.

Testing of Hypothesis for population Variance Using Chi-Squared test

Till now we were dealing with hypothesis testing for the means of various samples, but sometimes it is also necessary or desired to test the variances of the population under study i.e. let's we obtained certain variance for a sample which is different than the population variance, now we need to find out if the variances are within acceptable limit or does it varies more than the desired variance of the population.

The chi-square test for variance is a non-parametric statistical procedure with a chi-square-distributed test statistic that is used for determining whether the variance of a variable obtained from a particular sample has the same size as the known population variance of the same variables.

The test statistic of the chi-square test for variance is calculated as follows:

$$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$$

where, n is sample size, s is sample deviation, σ is population std. deviation

As similar with other tests, the critical value is obtained through a chi table on the basis of degree of freedom and significance level.

Degree of Freedom	Probability of Exceeding the Critical Value								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

We will see about it with an example:

Q) The variance of a certain size of towel produced by a machine is 7.2 over a long period of time. A random sample of 20 towels gave a variance of 8. You nee to check if the variability for towel has increased at 5% level of significance, assuming a normally distributed sample.

Ans.

n = 20

sample variance = 8

population variance = 7.2

H0: variance <= 7.2

H1: variance > 7.2 (Right tailed test)

Using chi squared test,

χ-square = (20-1) * 8/7.2 = 21.11

Critical value for D.o.f = 19 and 5% significance level,

Degree of Freedom	Probability of Exceeding the Critical Value							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92

Critical value = 30.14

Here, chi value is less than the critical value, thus we do not reject the null hypothesis.

Chi-Squared Test for Categorical Variables

The chi-square test is widely used to estimate how closely the distribution of a categorical variable matches an expected distribution (the **goodness-of-fit test**), or to estimate whether two categorical variables are independent of one another (**the test of independence**).

In mathematical terms, the χ2 variable is the sum of the squares of a set of normally distributed variables.

Suppose that a particular value Z1 is randomly selected from a standardized normal distribution. Then suppose another value Z2 is selected from the same standardized normal distribution. If there are **d** degrees of freedom, then let this process continue until **d** different Z values are selected from this distribution. The χ2 variable is defined as the sum of the squares of these Z values

$$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 + \cdots + Z_d^2$$

This sum of squares of **d** normally distributed variables has a distribution which is called theχ2distribution with d degrees of freedom.

Chi Squared test For Goodness Of fit

Chi Square test for testing goodness of fit is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

A goodness of fit test is a test that is concerned with the distribution of one categorical variable.

The null and alternative hypotheses reflect this focus:

H0: The population distribution of the variable is the same as the proposed distribution

HA: The distributions are different

The chi-square statistic is calculated as:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

Where, Observed= actual count values in each category

Expected= the predicted (expected) counts in each category if the null hypothesis were true.

Let's see an example for better understanding:

Q) A survey conducted by a Pet Food Company determined that 60% of dog owners have only one dog, 28% have two dogs, and 12% have three or more. You were not convinced by the survey and decided to conduct your own survey and have collected the data below,

Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs

Determine whether your data supports the results of the survey by the pet.

Use a significance level of 0.05

Ans: E(1 dog) =0.60

E(2 dog) = 0.28

E(3 dogs) = .12

H0: proportions of dogs is equal to survey data

H1: proportions of dogs is not equal to survey data

	1 Dog	2 Dog	3 Dog	Total
Observed	73	38	18	129
Expected	0.60*129 = 77.4	0.28 *129=36.12	0.12 * 129 = 15.48	129
Observed -Expected	-4.4	1.88	2.52	
(Observed -Expected) ^2	19.36	3.53	6.35	

Chi statistics = 19.36/77.4 + 3.53/36.12 + 2.52/15.48 = 0.7533

Let's see the critical value using d.o.f 2 and significance 5%:

Degree of Freedom	Probability of Exceeding the Critical Value							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81

Critical chi = 5.99

Here, our chi statistic is less than the critical chi. Thus, we will not reject the null hypothesis.

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other by analyzing comparisons of variance estimates. ANOVA checks the impact of one or more factors by comparing the means of different samples.

When we have only two samples, t-test and ANOVA give the same results. However, using a t-test would not be reliable in cases where there are more than 2 samples. If we conduct multiple t-tests for comparing more than two samples, it will have a compounded effect on the type 1 error.

Assumptions in ANOVA

1) Assumption of Randomness: The samples should be selected in a random way such that there is no dependence among the samples.
2) The experimental errors of the data are normally distributed.
3) Assumption of equality of variance (Homoscedasticity) and zero correlation: The variance should be constant in all the groups and all the covariance among them are zero although means vary from group to group.

One Way ANOVA

When we are comparing groups based on only one factor variable, then it said to be one-way analysis of variance (ANOVA).

For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.

The ANOVA model:

Mathematically, ANOVA can be written as:

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

where x are the individual data points (i and j denote the group and the individual observation), ε is the unexplained variation and the parameters of the model (μ) are the population means of each group. Thus, each data point (x_{ij}) is its group mean plus error.

Let's understand the working procedure of One-way Anova with an example:

Sample(k)	1	2	3	Mean
1	x_{11}	x_{12}	x_{13}	X_{m1}
2	x_{21}	x_{22}	x_{23}	X_{m2}
3	x_{31}	x_{32}	x_{33}	X_{m3}
4	x_{41}	x_{42}	x_{43}	X_{m4}

Suppose we are given with the above data set; we have an independent variable x and 3 samples with different values of x and each sample has its respective mean as shown in last column.

Grand Mean

Mean is a simple or arithmetic average of a range of values. There are two kinds of means that we use in ANOVA calculations, which are separate sample means and the grand mean.

The **grand mean (Xgm)** is the mean of sample means or the mean of all observations combined, irrespective of the sample.

$X_{gm} = (X_{m1} + X_{m2} + X_{m3} + X_{m4} + \ldots\ldots\ldots X_{mk})/k$ where, k is the number of samples

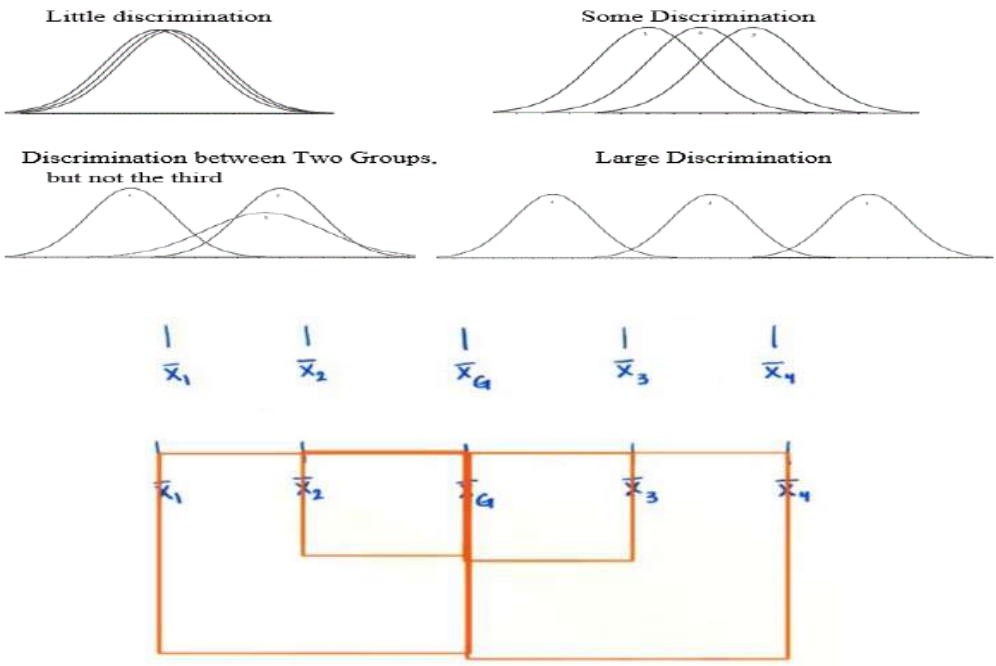
For our dataset, k = 4

$X_{gm} = (X_{m1} + X_{m2} + X_{m3} + X_{m4})/4$

Between Group Variability (SST)

It refers to variations between the distributions of individual groups (or levels) as the values within each group are different.

Each sample is looked at and the difference between its mean and grand mean is calculated to calculate the variability. If the distributions overlap or are close, the grand mean will be similar to the individual means whereas if the distributions are far apart, difference between means and grand mean would be large.



Let's calculate Sum of Squares for between group variability:

$SS_{between} = n_1 * (X_{m1} - X_{gm})^2 + n_2 * (X_{m2} - X_{gm})^2 + n_3 * (X_{m3} - X_{gm})^2 + \ldots\ldots\ldots\ldots + n_k * (X_{mk} - X_{gm})^2$

where, n_1, $n_{2,\ldots.}$,n_k are the number of observations in each sample

Degree of freedom for between group variability = number of samples – 1 = k-1

$Mean_{SSbetween} = SS_{between}/k-1$

In our dataset example we have k =4 and n_k = 3, so for our dataset:

$SS_{between} = 3 * (X_{m1} - X_{gm})^2 + 3 * (X_{m2} - X_{gm})^2 + 3 * (X_{m3} - X_{gm})^2 + 3 * (X_{m4} - X_{gm})^2$

$Mean_{SSbetween}(MSST) = SS_{between}/ (4-1) = SS_{between}/3$

Within Group Variability (SSE)

It refers to variations caused by differences within individual groups (or levels) as not all the values within each group are the same. Each sample is looked at on its own and variability between the individual points in the sample is calculated. In other words, no interactions between samples are considered.

We can measure Within-group variability by looking at how much each value in each sample differs from its respective sample mean. So, first, we'll take the squared deviation of each value from its respective sample mean and add them up. This is the sum of squares for within-group variability.

$$SS_{within} = \Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2 + \ldots + \Sigma(x_{ik} - \bar{x}_k)^2$$
$$= \Sigma(x_{ij} - \bar{x}_j)^2$$

Note: x_{i1} is the ith value from the first sample, x_{i2} is the ith value from the second sample, and so on all the way to x_{ik}, the ith value from the kth sample. x_{ij} is therefore the ith value from the jth sample.

Degree of Freedom for within variability:

$$df_{within} = (n_1 - 1) + (n_2 - 1) + \ldots (n_k - 1) = n_1 + n_2 + n_3 + \ldots. + n_k - k(1) = N - k$$

$$MS_{within} = \sum(x_{ij} - \overline{x_j})^2 / (N - k)$$

Where, N is the total number of observations.

In our dataset example we have k =4 and N =12, so for our dataset:

$SS_{within} = (X_{11} - X_{m1})^2 + (X_{12} - X_{m1})^2 + (X_{13} - X_{m1})^2 +$

$(X_{21} - X_{m2})^2 + (X_{22} - X_{m2})^2 + (X_{23} - X_{m2})^2 +$

$(X_{31} - X_{m3})^2 + (X_{32} - X_{m3})^2 + (X_{33} - X_{m3})^2 +$

$(X_{41} - X_{m4})^2 + (X_{42} - X_{m4})^2 + (X_{43} - X_{m4})^2$

Degree od Freedom = N-k = 12 -4 = 8

$Mean_{SSwithin}(MSSE) = SS_{within}/ 8$

<u>Total Sum of Squares (TSS)</u>

$TSS = SS_{between} + SS_{within} = SST + SSE$

Hypothesis In ANOVA

The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don't have any significant difference. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:

$$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_L \qquad \textit{Null hypothesis}$$

$$H_1: \ \mu_l \neq \mu_m \qquad \textit{Alternate hypothesis}$$

where μ_1 and μ_m belong to any two sample means out of all the samples considered for the test. In other words, the null hypothesis states that all the sample means are equal or the factor did not have any significant effect on the results. Whereas, the alternate hypothesis states that at least one of the sample means is different from another.

To test the null hypothesis, test statistics is given by the F-statistic.

F-Statistic

The statistic which measures if the means of different samples are significantly different or not is called the F-Ratio. Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.

$F = Mean_{SSbetween} / Mean_{SSwitihn}$

F = MSST / MSSE with k-1 and N-k degrees of freedom.

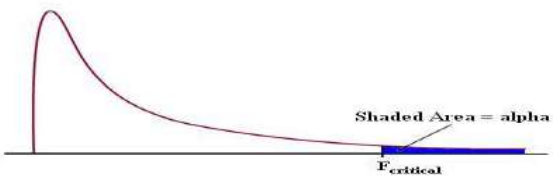
This above formula is pretty intuitive. The numerator term in the F-statistic calculation defines the between-group variability. As we read earlier, as between group variability increases, sample means grow further apart from each other. In other words, the samples are more probable to be belonging to totally different populations.

This F-statistic calculated here is compared with the F-critical value for making a conclusion.

F-critical is calculated using the F-table, degree of freedoms and Significance level.

If observed value of F is greater than the F-critical value then we reject the null hypothesis.

Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.



Let's see an example on One-way ANOVA analysis:

Q) In a survey conducted to test the knowledge of Mathematics among 4 different schools in city. The sample data collected for the marks of students out of 10 is below:

School	Marks						
School 1	8	6	7	5	9		
School 2	6	4	6	5	6	7	
School 3	6	5	5	6	7	8	5
School 4	5	6	6	7	6	7	

Ans:

H0: All the schools have equal means

H1: Difference in means of schools is significant

k = 4

N = 24

	School 1(S1)	School 2(S2)	School 3(S3)	School 4(S4)	(S1) - S1_mean)^2	(S2- S2_mean) ^2	(S3- S3_mean)^2	(S4- S4_mean) ^2
	8	6	6	5	1	0.111111 556	0	1.361095 556
	6	4	5	6	1	2.777775 556	1	0.027775 556
	7	6	5	6	0	0.111111 556	1	0.027775 556
	5	5	6	7	4	0.444443 556	0	0.694455 556
	9	6	7	6	4	0.111111 556	1	0.027775 556
		7	8	7		1.777779 556	4	0.694455 556
			5				1	
Total	35	34	42	37	10	5.333333 333	8	2.833333 334
Mean	7	5.6666666 67	6	6.166666 667				
Grand mean	6.208333 333							

$SS_{between}$ = 5 * (7-6.21)^2 + 6 * (5.7 – 6.21)^2 + 7 * (6-6.21)^2 + 6 * (6.17 – 6.21)^2

= 4.99

MSST = 4.99 / (4-1) = 1.66

SS_{within} = 10 + 5.33 + 8 + 2.83 = 26.16

MSSE = 26.16/(N-k) = 26.16/20 = 1.308

F-statistics = MSST/MSSE = 1.66 / 1.308 = 1.27

Critical F-value

At 5% significance and degree of freedom (3, 20):

F - Distribution (α = 0.05 in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom						
Denominator Degrees of Freedom	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638

F-critical = 3.098

Clearly, our F-statistics is less than F-critical. So, we cannot reject our null hypothesis.

Two Way ANOVA

Two-way ANOVA allows to compare population means when the populations are classified according to two independent factors.

Example: We might like to look at SAT scores of students who are male or female (first factor) and either have or have not had a preparatory course(second factor).

The Two-way ANOVA model:

Mathematically, ANOVA can be written as:

$$x_{ij} = \mu_{ij} + \varepsilon_{ij}$$

where x are the individual data points (i and j denote the group and the individual observation), ε is the unexplained variation and the parameters of the model (μ) are the population means of each group. Thus, each data point (x_{ij}) is its group mean plus error.

Just like one-way model, we will calculate the sum of squares between, in this case there will be two SSTs for both the categories and sum of squares of errors (within).

We calculate F-statistics for both the MSST and see which once greater value than F-critical and compare them to find the effect of both categories on our assumption.

Example:

Below given is the data of yield of crops based on temperature and salinity. Calculate the ANOVA for the table.

Temperature (in F)	Categorical variable salinity				
	700	1400	2100	Total	Mean(temp)
60	3	5	4	12	4
70	11	10	12	33	11
80	16	21	17	54	18
Total	30	36	33	99	11
Mean(salanity)	10	12	11	11	

Ans:

Hypothesis for Temperature:

H0: Yield is same for all temperature

H1: yield varies with temperature with significant difference

Hypothesis for Salinity:

H0: Yield is same for all Salinity

H1: yield varies with temperature with significant Salinity

Grand mean = 11

N = 9, K =3, n_t= 3, n_s = 3

$SS_{between\_temp}$ = 3 *(4-11)^2 + 3*(11-11)^2 + 3*(18-11)^2 = 294

$MSST_{temp}$ = 294 / 2 = 147

$SS_{between\_salanity}$ = 3 *(10-11)^2 + 3*(12-11)^2 + 3*(11-11)^2 = 6

$MSST_{salainity}$ = 6 /2 = 3

In such question calculating SSE can be tricky, so instead of calculating SSE let's calculate TSS then we can subtract SST values from it and get SSE.

To calculate Total sum of squares, we need to find sum of the squares of difference of each value from the grand mean.

TSS = (3-11)^2 + (5-11)^2 + (3-11)^2 +(4-11)^2 +(11-11)^2 +(10-11)^2 +(12-11)^2 +(16-11)^2 +(21-11)^2 +(17-11)^2

TSS = 312

SSE = TSS - $SS_{between\_temp}$ - $SS_{between\_salanity}$ = 312 – 294-6 = 12

Degree of freedom for SSE = (n_t-1$_)($ n_s-1) =(3-1)(3-1) = 4

MSSE = SSE/4 = 3

<u>F-Test For temperature</u>

F_{temp} = $MSST_{temp}$/ MSSE = 14/3 = 49

<u>F-Test For Salinity</u>

$F_{salinity}$ = $MSST_{salinity}$/MSSE = 3/3 = 1

F-critical for 5% significance and degree of freedom (k-1, (p-1) (q-1)) i.e. (2,4):

F-critical = 10.649

Clearly, we can see that F_{temp} is greater than F-critical, so we reject the null hypothesis and support that temperature has a significant effect on yield.

On the other hand, $F_{salinity}$ is less than the F-critical value, so we do not reject the null hypothesis and support that salinity doesn't affect the yield.

Confusion Metrics

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Once you understand these four parameters then we can calculate Accuracy, Precision, Recall and F1 score.

Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

Accuracy = TP+TN/TP+FP+FN+TN

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate.

Precision = TP/TP+FP

Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

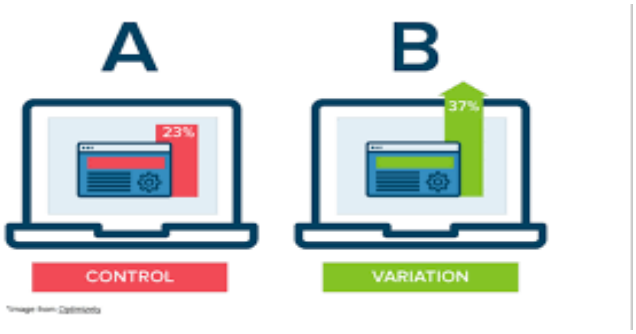
Recall = TP/TP+FN

F1 score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

A/B Testing



An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

To put this in more practical terms, a prediction is made that Page Variation #B will perform better than Page Variation #A, and then data sets from both pages are observed and compared to determine if Page Variation #B is a statistically significant improvement over Page Variation #A.

This process is an example of statistical hypothesis testing.

The null hypothesis for the A/B test might be something like this:

- The difference in conversion between Version A and Version B is caused by random variation. It's then the job of the trial to disprove the null hypothesis. If it does, we can adopt the alternative explanation:
- The difference in conversion between Version A and Version B is caused by the design differences between the two.

To determine whether we can reject the null hypothesis, we use certain statistical equations to calculate the likelihood that the observed variation could be caused by chance, which include Student's t test, χ-squared and ANOVA.

Using different tests, as we did in several other examples, the test statistics is calculated (t-score, p-value, chi score or F-score) and it is compared to the critical value as per the significance level. We then accept or reject our null hypothesis based on the Test score.