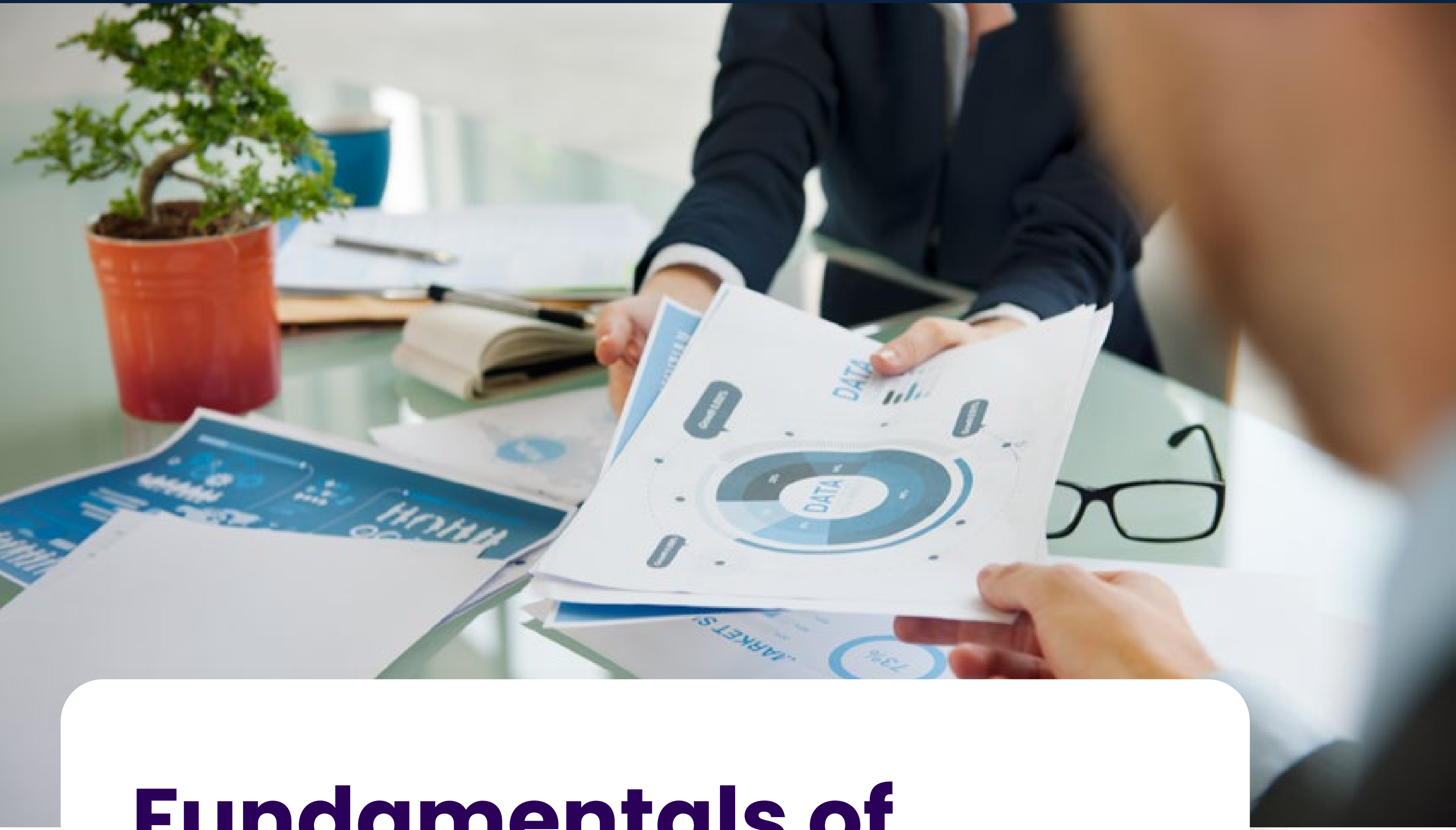
Analyttica TreasureHunt® LEAPS

www.analyttica.com



Fundamentals of Data Analytics

Module 3:

Frequently Applied Statistical Tests

ANALYTTICA
TREASURE HUNT
LEARN . APPLY . SOLVE

Table of Contents

In the previous module, we covered the concepts of probability distributions, the importance of normal distributions in statistics and the concept of statistical inference. In this module, we will aim to cover the various statistical tests one can carry out, to validate a few of the concepts covered in the previous module, like the normality of the distributions of the sample data, or the rejection or acceptance of a null hypothesis (H0).

One Tailed vs. Two-Tailed Tests

The one tailed and two tailed tests in hypothesis testing are alternate ways of computing the statistical significance of a parameter inferred from a data set. Let us consider the supermarket example discussed in the previous module.

For example: In a supermarket, the average sales last year was $260. A market researcher wants to know if the average sales has gone up this year or not.

Now note that the "expectation" is that the average sales has not come down. Suppose, μ_1 is the average sales this year. So, we expect $\mu_1 \geq 260$.

Hence our null hypothesis is H_0: $\mu_1 = 260$ and the alternate hypothesis is H_1: $\mu_1 > 260$.

If we assume sales of an individual follows normal distribution, then the distribution will look as below. The alternate hypothesis expects the value of to be somewhere in the right side of '260' in the axis.

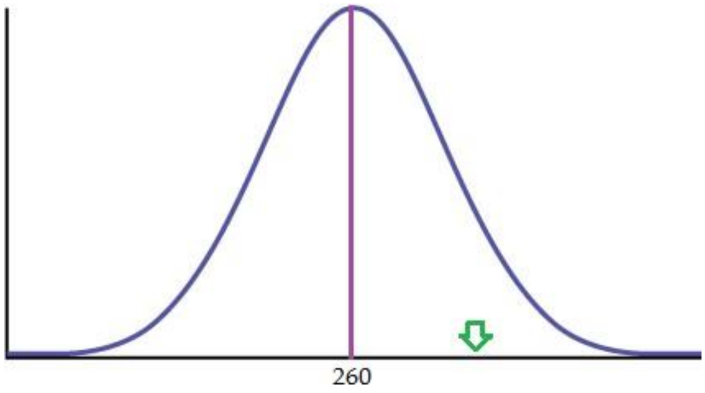


Figure 1: Right-tailed test

This is called a right-tailed test. Similarly, if the alternate hypothesis to be H_2: $\mu_1 < 260$, then it would be a left-tailed test.

For example: Average crime rate per 1000 citizen in cities of a country was previously recorded as 15.3. Few years back, the government funding on police was increased and new technology was introduced. Now, the government wants to see if crime rate has gone down or not.

Clearly, we should not expect any kind of 'increase' in crime rate. Hence it will be a left tailed test.

On the other hand, if the objective of the test would have been to test "if average sales this year is different from previous year sales" then it would be a two-tailed test.

Let's formulate the null hypothesis and alternate hypothesis for the average sales problem. The null hypothesis here is same as before, H_0: $\mu_1 = 260$. The alternate hypothesis is H_0: $\mu_1 \neq 260$.

So, if we look at the distribution of sales in current year, under alternate hypothesis the sales can lie on either side of '260' on the axis.

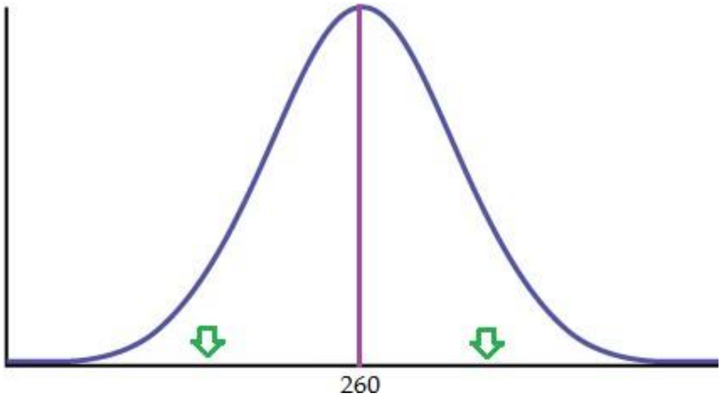


Figure 2: Two-tailed test

Z Test

The Z test is a statistical test in which the z-statistic follows a normal distribution. This test is based used for sample sizes greater than 30, because, as per the central limit theorem, as the number of samples get large, the distribution tends to follow a normal distribution. When conducting a z test, the null and alternate hypothesis must be stated, following which, the z-statistic must be calculated.

Let us consider the following example, a market research analyst wants to prove that the shoppers from a certain residential area, say Presidential Height, are above the average spenders at the nearby Macy's shop. A random sample of 30 such shoppers from Presidential Height has an average spend of $112.5. The average spend per shopping trip across all demographics is $100, with a standard deviation of 15.

Null Hypothesis: H_o: $\mu = 100$, as the population mean is 100 as given.

Alternate Hypothesis: H_1: $\mu > 100$, as the analyst claims that shoppers from Presidential Height have above than average spend per shopping trip.

Post formulating the null and alternate hypothesis we follow certain steps:

- We summarize the data into a metric called 'test statistic'
- Then we compare the observed value of the test statistic with respect to its distribution under null hypothesis
- Finally, we find a likelihood of null hypothesis being true in view of the data (what we call a p-value)

Similarly, in the above example the test statistic (assuming spend follows normal distribution) is:

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

Where the sample mean is $\overline{x,}$ μ is the population mean under null hypothesis, σ is the population standard deviation and *n* is the sample size

Under null hypothesis, Z, the test statistic, will follow Normal Distribution with mean zero (0) and standard deviation one (1). This distribution is more popularly called a Standard Normal Distribution.

As per the data,

$\overline{x}$ = $112.5, μ = $100, σ=$15 and n= 30

This gives the observed value of statistic as z= 4.56

One Tailed Z-Test

Now how likely it is that Z~N(0,1) ? For a right tailed test, we try to find

P(Z>4.56, given Z is a standard normal variable)

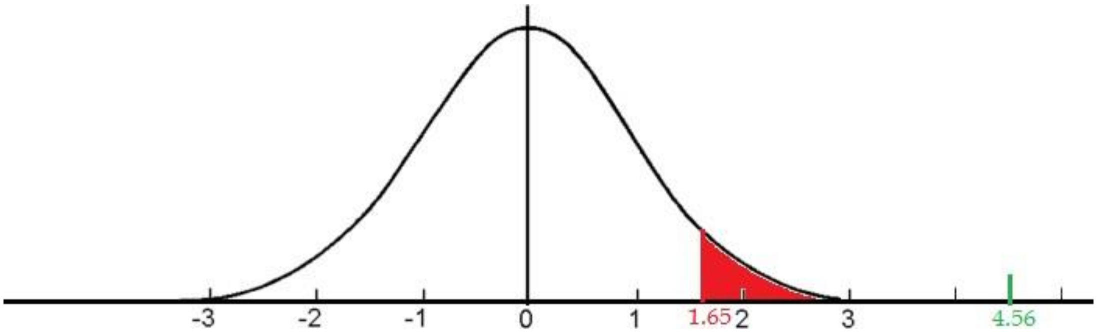


Figure 3: Right-tailed Z test

From the above figure, it is clear that there is almost zero probability that a standard normal variable takes a value 4.56. Hence, we reject the null hypothesis.

One the other hand, we see that P(Z>1.645) = 0.05 (the red filled area). Hence, any value of Z greater that 1.65 can be considered as very unlikely and hence null hypothesis could be rejected with 95% significance level. This value 1.645 is also known as critical value of the test.

As the test statistic follows standard normal distribution under null hypothesis, the test is known as Z- test (In statistical theories, a standard normal distribution is often denoted by Z).

Two Tailed Z-Test

Let's revisit the same problem where the researcher wants to test if the new average spend is different from previous $100 due to marketing campaign or not. This means the alternate hypothesis is $H_1: \mu \neq 100$ and it is a two-tailed test.

The test statistic and calculation of its observed value remains same as before, just the calculation of likelihood changes.

We know Normal distribution is symmetric around its mean. Hence Z, the test statistic, should be symmetrically distributed around zero (0) under null hypothesis. Hence, we calculate,

P (Z>4.56 or Z < -.4.56, given Z is a standard normal variable)

Or

P (|Z| >4.56, given Z is a standard normal variable)

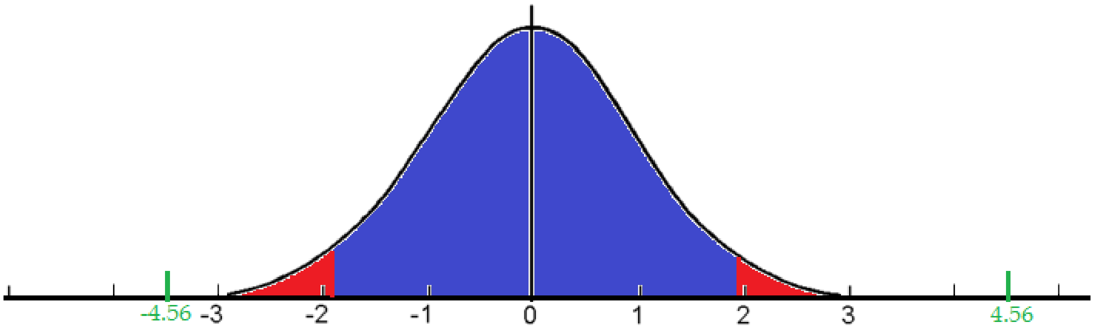


Figure 4: Two-tailed Z test

Like in the previous case, it is clear that there is almost zero probability that a standard normal variable takes a value 4.56. Hence, we reject the null hypothesis.

One the other hand, we see that P(Z>1.96) = 0.025 (the right red filled area) and P(Z<-1.96) = 0.025 (the left red filled area). Hence, any value of Z greater that 1.96 or lesser than -1.96 can be considered as very unlikely and hence null hypothesis could be rejected with 95% significance level. 1.96 is the critical value of the test.

Assumptions for Z-Test

If we note very carefully, we have made the following assumptions for a Z Test

- The data under consideration is continuous
- The observations follow a Normal distribution, even approximately
- The population variance of the Normal distribution is known
- The observations are independent

Some Standard Critical Values for Z-Test

The following table shows critical values for some standard significance level for a two-tailed test.

Table 1:

Significance	α = Tail Area	Central Area = 1 – 2α	Critical Value
80%	0.1	0.8	$z_{.10}$ = 1.28
90%	0.05	0.9	$z_{.05}$ = 1.645
95%	0.025	0.95	$z_{.025}$ = 1.96
98%	0.01	0.98	$z_{.01}$ = 2.33

Two Sample Z-Test

Suppose we want to test the equality of means of two independent samples, we can perform a two-sample z test, if we know the population standard deviations.

Hence, if $\mu_1$and μ_2 are the means of the two populations under consideration, we test the null hypothesis H_o: $\mu_1 = \mu_2$.
With $\sigma_1^2$and σ_2^2as the population variances, the test statistic is:

$$Z = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The test statistic follows standard normal distribution under null hypothesis.

Consider an example where we want to test if the "Daily Log Return" of Company A is same as that of Company B. The data was taken for last 30 working days for both the companies. The sample means were found to be 0.10% and 0.09% respectively.

It is also known that the population variance of "Daily Log Return" for Company A and Company B are 0.027% and 0.037% respectively.

This gives the test statistic Z as 0.0197 which has a p-value of 0.984. This means we accept the null hypothesis, i.e. the average daily log returns are same for both the companies.

Equality of Proportion

We can use the same Z-test, one sample as well as two sample, to test equality of proportion, provided the sample size is huge.

For example: We want to check to see if the proportion of customer complaints getting escalated is more than the acceptable value of 20%.

Hence, if p is the proportion of complaints getting escalated, we need to test

H_o: $p = p_0$ vs. H_1: $p < p_0$. Here, p_0 is 0.2. The test statistic will be,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

If we have data on 250 customer complaints, out which 60 is escalated, then$\hat{p}$, the sample proportion is 0.24 (=60/250).

For example: In a marketing campaign, a retail bank reached out to 20,000 people offering their home loan to transfer from other banks at a cheap interest rate. 500 of them came and availed the offer. During the same phase, a control population of 1,000 people was also kept where no communication was made, but the same offer was given to people who asked for a home loan transfer. 15 out of the 1,000 control population came forward and asked to transfer their home loan.

Now, the marketing analyst wants to know if there was any incremental response due to the campaign.

So, here we want to test if the proportion of mailed population that responded (say $\boldsymbol{p_1}$) and proportion of control population that responded (say $\boldsymbol{p_2}$) are significantly different or not. Hence, we test

H_o: $p_1 = p_2$ vs. H_1: $p_1 > p_2$

The test statistic we use here is:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\hat{p}_1$ and $\hat{p}_2$ are sample response rates for mailed and control population respectively. $\hat{p}$ is the pooled response rate, i.e. considering both the samples together.

In the above example, $\hat{p}_1$ is 500/20,000 = 0.025, $\hat{p}_2$ is 15/1,000= 0.015 and $\hat{p}$ is (500+15)/(20,000 + 1,000)= 0.0245. This gives the observed value of test statistic as 1.995.

t Tests

Student’s t-Test

In the previous example of Z-test, the standard deviation was known to be $15. Now, in most real-life scenarios, this is not known. In such cases, we replace the σ, the population mean, by *s*, the sample mean.

$$z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \quad \Rightarrow \quad T = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$$

The above test statistic, T, follows a t-distribution with degrees of freedom *(n-1)*. Degrees of freedom reduces from *n* to *n-1* as we are estimating *s* and using it in the statistic.

Now to test H_o: $\mu = 100$ vs. H_1: $\mu > 100$, we calculate observed value of T. If *s*= 14.25 then T = 4.80

Now we find P (T > 4.80, T ~ t_{29}), which is very low, almost zero. Hence, we reject null hypothesis.

If we test H_o: $\mu = 100$ vs. H_1: $\mu \neq 100$, then, we would be calculating, P (|T| > 4.80, T ~ t_{29})

This test is known as Student’s t-Test.

Two Sample t-Test

Often, we need to test the equality of means of two independent populations. We use two sample t-test in those cases.

For example: In an analysis of HR department of a software organization, the researcher wants to study if employees from a computer science (Com Sc.) or information technology (IT) engineering background perform better than employees with engineering in other fields. She selected 15 people with Com Sc. or IT randomly and 20 people from other background (e.g. Mechanical, Tele Communication, Electronics, etc.). All of them were rated on a scale of 10 by their line managers. (Refer to Table 2).

What will be her null and alternate hypothesis? Suppose μ_1 is the average performance of Population 1 (i.e. Com Sc or IT Engineers) and μ_2 is the average performance of Population 2 (i.e. Engineers with other background).

The researcher needs to test H_o: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 > \mu_2$ There will be a right tailed test as objective is to test if “Population 1” performs better than “Population 2”

Here, we need to make the first assumption that the data comes from a normal distribution. Secondly there can be two instances; viz. the variance of the two populations are same or they are different. Test statistics differ across two scenarios.

Table 2:

Population 1: Engineers with Com Sc or IT		Population 2: Engineers with other subjects	
5.16	9.31	6.95	4.98
4.05	8.84	8.22	6.53
5.62	6.74	8.09	3.56
5.46	9.16	7.08	4.83
7.38	8.39	4.75	3.87
6.8	6.99	5.08	5.03
5.36	9.24	5.73	3.64
5.67		7.49	7.31
		4.18	4.76
		6.42	5.39

Two Sample t-Test with Equal Variance

Here, we assume the two populations have same variation, only their means can be different. Hence, we test if the mean is same or not.

Note what happens if the means are same – both the populations follow normal distribution, both have same variance and same mean. This means there is absolutely no difference between the two population and they can be assumed to come from same population. On the other hand, if means are different, they are from a normal distribution with same variance, only mean is different. Hence, just the populations differ with respect to scale. (Please refer to Figure 5)

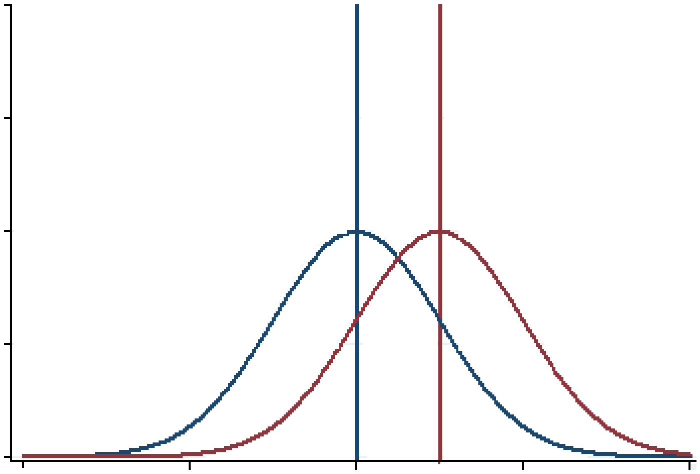


Figure 5: Distribution of two population for Two Sample t-test with Equal Variance

For two sample t-test with equal variance, the test statistics is as below:

$$T_e = \frac{(\overline{x}_1 - \overline{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

There are *n₁* observations from "Population 1" which gives sample mean as $\overline{x}_1$ and sample standard deviation s_1. Similarly, n_2, $\overline{x}_2$and s_2 are defined.

s is called as pooled variance, which is given by as below:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Under null hypothesis, the test statistic, $T_e$follows a t-distribution with degrees of freedom $(n_1 + n_2 - 2)$

For example: In the above example, we can calculate and get:

n_1= 15, n_2= 20, $\overline{x}_1$=6.94, $\overline{x}_2$= 5.69, s_1= 1.719, s_2= 1.469, n_1+ n_2-2= 33

These values give us s = 1.628 and the observed value of t-statistic as 2.248

As we are doing a right tailed test, the p-value is 0.016. Hence, though at 95% significance level we reject the null hypothesis, but at 99% significance we accept the null hypothesis.

Two Sample t-Test with Unequal Variance

If the variances of the two populations under consideration are different, the test statistic must be modified.

The revised test statistic is given as below:

$$T_{ue} = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Though the above formula looks simple, the complication arises in calculation of degrees of freedom for the above statistics. According to Satterthwaite's correction, the degrees freedom is the nearest integer in 'm', where 'm' is given by:

$$m = \frac{(n_1 - 1)(n_2 - 1)}{(n_1 - 1)c^2 + (n_1 - 1)(1 - c^2)}$$

Where,

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

But, if n_1 and n_2 are close to each other, the degrees of freedom can be approximated to

$$n_1 + n_2 - 2$$

For example: In the above example, we can calculate and get:

n_1= 15, n_2= 20, $\overline{x}_1$=6.94, $\overline{x}_2$= 5.69, s_1= 1.719, s_2= 1.469

These values give us the observed value of t-statistic as 2.248 and c = 0.646. We get m = 32.34.

The nearest integer of 32.34 is 32, hence the test statistic will follow a t-distribution with d.f. 32 under null hypothesis. This gives us p-value as 0.015.

Hence our decision remains same even under un-equal variance test. We reject the null hypothesis at 95% significance level, but at 99% significance we accept the null hypothesis.

Paired t-Test

Suppose in a clinical research, the researcher wants to test the effectiveness of a new medicine to reduce blood sugar level on diabetic patients. 100 diabetic patients were selected at random. Their blood sugar levels were measured and then the medicine was applied on each of them. After two hours, the blood sugar level of the same set of patients were measured. Now the researcher wants to see if there is any significant decrease in the blood sugar level or not.

Note that here we want to test if the average blood sugar level pre-medication and post-medication stage are same or not. Hence, it is testing equality of means. But we cannot apply two-sample t-test here because the two samples, i.e. pre-medication blood sugar level and post-medication blood sugar level, are not independent as they are measured on the same individuals. We use paired t-test in this scenario.

In other words, when same set of individuals (here, the patients) were measured on same metric (here, blood sugar level) in two different times (here, pre-medication and post-medication) and we want to test the equality of means, we use “paired t-test”.

For example: An FMCG company ran some promotional campaigns across various cities of a country. They have measured their market share in 15 cities both pre-campaign and post-campaign. They want to test if the market share has significantly gone up due to the campaigns or not.

Table 3:

City	Pre-Campaign	Post Campaign	City	Pre-Campaign	Post Campaign
City A	19.02	13.53	City I	22.32	27.2
City B	16.56	26.85	City J	25.51	35.99
City C	27.61	23.86	City K	26.59	36.37
City D	45.24	45.69	City L	37.45	37.21
City E	25.87	26.92	City M	19.74	25.23
City F	29.77	30.86	City N	37.06	40.12
City G	27.7	29.76	City O	23.89	25.72
City H	24.82	23.85			

Suppose, X_i is the market share of i^{th} city pre-campaign and Y_i is that post the campaign. If μ_x and μ_y are mean of X and Y, then, we want to test $H_o: \mu_y = \mu_x$ vs. $H_1: \mu_y > \mu_x$

The test statistic is given by:

$$T_p = \frac{\bar{d}}{\left(\frac{s}{\sqrt{n}}\right)}$$

Were $d_i = Y_i - X_i$ and $\bar{d}$ & s are mean and standard deviation of d_i respectively.

Under null hypothesis, the test statistic follows a t-distribution with (n-1) degrees of freedom.

For example: In the given example, $\bar{d}$ = 2.66 and s = 4.80. Hence, observed value of t-statistic is 2.15. The p-value of the test is 0.024.

Two Sample t-Test vs. Paired t-Test

It is very important to note the difference between two sample t-test & paired t-test and apply the right test at the right scenario. Two sample t-test is applied when the samples are independent and one does not have any relation with the other. In paired t-test, the same individuals are measured in different time frames. Hence, in paired-t-test the individuality of data points is important.

Here are some sample comparisons of two types of tests (Table 4):

Two Sample t-Test	Paired t-test
Testing if the score in an examination is different for girls and boys	Testing if average score of students has changed from year 1 to year 2
Testing the average sales of customers are different for two different customer segments	Testing if the average sales has changed after last campaign
Testing if average claim amount in a motor insurance portfolio is more for customers with prior claim history or not	Testing if average claim amount is more this year of those who claimed both this year and last year

Assumptions for t-Test

Be it a one sample t-Test, two sample t-Test or paired t-Test, we make the following assumptions:

- The data under consideration is continuous
- The observations follow a Normal distribution, even approximately
- The population variance of the Normal distribution is known
- The observations are independent

t-Test vs Z-Test

If the degrees of freedom of a t-distribution is high, more than 30, the t-distribution becomes similar to standard normal distribution. Hence, for large samples we can use z-test in place of t-test, especially Student's t test and paired t-test.

Test of Equality of Variance or F Test

In two sample t-Test, the test statistic differs if the variance of two populations are same or not. Hence, it becomes important to test if the population variances are same or not.

Under the same assumptions, as in t-Test, we test $H_o: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$

We calculate the sample variances for the two populations s_1^2 and s_2^2 respectively. Then we define the test statistic as below:

$$F = {s_1}^2 / {s_2}^2$$

The above statistic follows a F distribution with degrees of freedom $(n_1 - 1)$ and $(n_2 - 1)$, where n_1 and n_2 are the sample sizes from the two samples respectively.

For example: In the previous example (given in two-sample t-test), we can calculate and get:

$n_1 = 15$, $n_2 = 20$, $\bar{x}_1 = 6.94$, $\bar{x}_2 = 5.69$, $s_1 = 1.719$, $s_2 = 1.469$, $n_1 + n_2 - 2 = 33$

So, F = 1.369. Now, for a right tailed test, the p-value is 0.2577. Hence, there is 25.77% likelihood that the null hypothesis is correct.

This test is also known as test of homogeneity between two groups.

The assumptions of the test are same as t-test or Z-test

Analysis of Variance

In the previous section, we learned how to test the equality of means of two independent population. We used a two-sample t-test. But this becomes difficult to test equality of means between more than two populations. In that case we must run independent t-tests for each pair of populations.

Suppose a retail shop wants to test if the average transaction value of customers is different among their 10 major outlets. Now, if we want to do this through two sample t-test, we must perform 45 tests (45 pairs of stores).

To avoid that we use "Analysis of Variance" (ANOVA) to test if at all there is any difference or not.

Analysis of Variance Analyses Means, NOT Variance

One can easily get mislead that Analysis of variance, more popularly known as ANOVA test the equality of variance of population through its name. But, one should remember that it has nothing to do with variance of the population. Actually, it analyses means instead. To be very specific, it analyses variation of means across multiple groups or sub-groups of population. Hence the name.

Test of Homogeneity

In ANOVA, we not only assume the data follows normality, we also assume that the standard deviation is same across population or sub-populations under consideration. ANOVA becomes null and void if this is not true.

Hence, it is recommended that we first do the test of homogeneity, i.e. testing if the standard deviation is same or not, across the groups. Then, if we accept the null hypothesis (that they are same), we do ANOVA. This is why, in every statistical software the output of any ANOVA comes with a test of homogeneity and gives a warning message if the test fails.

For example: In an IT company, various kinds of IT tickets are raised. Each IT ticket is classified into any one of the three categories, viz. Category-A, Category-B & Category-C. You, as an analyst, want to see if average time to resolve a ticket (in Hour) is different for different categories.

Below is the data on 15 observations from each category (Refer to Table 5).

If μ_A, μ_B & μ_C are the average times to resolve a ticket (in Hour) then we would like to test H_o: $\mu_A = \mu_B = \mu_C$ with alternate hypothesis being: H_1: At least one on μ_A, μ_B, μ_C is different from rest

Table 5:

Category-A	Category-B	Category-C	Category-A	Category-B	Category-C
12.75	11.90	12.50	13.75	14.25	14.00
12.50	11.90	12.50	13.75	13.00	14.00
15.00	13.50	14.50	13.50	12.75	13.51
14.00	12.25	14.00	13.50	12.50	13.50
13.75	12.25	14.00	13.00	12.50	13.50
13.59	12.00	13.90	13.00	12.40	13.25
13.25	12.00	13.75	13.00	12.30	13.00

ANOVA Table – A Tool to Do ANOVA

We create a table called ANOVA table, which measures how subgroup means are different from over all mean.

For a data with 'k' groups and 'n' observations from each group, below is a quick reference:

Table 6:

Source	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F- Statistic**
Among Group Variation	k-1	$SSA = \sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2$	$MSA = {SSA}/{k-1}$	$F = {MSA}/{MSE}$
Within Group variation	nk-k	$SSE = SST - SSA$	$MSE = {SSE}/{nk-k}$	
Total Variation	nk-1	$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2$		

Just as a quick understanding, SST measures the total variation of data (note that it is the sample variation without the denominator). SSA is the variation across the subgroups. The remaining part of variation is within group variation. The test statistic, F, follows F distribution with degrees of freedom (k-1, nk-k).

The above table is known as ANOVA Table.

What Next After ANOVA

If we reject the null hypothesis through ANOVA table, we may then wish to test which group(s) is different from others. In that case we go back to two-sample t-test.

Correlation Test

Like mean or variance, we can perform tests for correlation coefficient also.

For example: We have data on two variables and we want to test if there is any correlation between these two variables, say X and Y.

If is ρ the population correlation (Pearson's correlation) then we want to test $H_o: \rho = 0$ vs. $H_1: \rho \neq 0$

We calculate the sample correlation coefficient, r, and define the test statistic as below:

$$t^r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The above statistic follows t distribution with degrees of freedom$(n-2)$, where n is the sample size.

We know that, t distribution becomes same as standard normal distribution for degrees of freedom more than 30. Hence, if sample size is high, we can use standard normal distribution as the distribution also.

For example: A financial analyst wants to find if the share price of a manufacturing company is correlated to the industrial growth index for the country or not. The quarterly data on last 10 years found the sample correlation, *r*, to be 0.32.

Hence, t^r = 2.08 which has a p-value 0.044

Chi-Square Test

Till now, we have only discussed testing about numeric variables. How to test non-numeric variable, e.g. categorical or ordinal variables? One of the most popular tests is Chi-Square test. Let's, learn this through an example.

Goodness of Fit Test

Consider the parking lot of an apartment, we want to test if the residents have any color bias about the color of their cars. We found that there are only 3 types of broad colors of cars, viz. Dark (like Blue, Steel, Black, etc.), Attractive (e.g. Red, Golden, Green, etc.), and Light (like White, Cream, Yellow, etc.).

From previous analysis, we know colors of 30% cars on road are of Dark color, 50% of Light color and remaining 20% are of Attractive color. We want to test if color preference of the residents of the apartment is different from what is seen on the road in general (Refer to Table 7).

Here, the null hypothesis is H_0: There is not color bias compared to alternative hypothesis
H_0: There is a color bias

Table 7:

Color	Dark	Attractive	Light	Total
Count of Cars	13	15	22	50

There are total 50 cars and if null hypothesis is true, then 50% of them (i.e. 25) would be of Light color, 30% (i.e. 15) would have been of Dark color and 20% (i.e. 10) would have been of Attractive color. This is the expected frequency under null hypothesis.

An analysis of deviation of observed frequency, i.e. actual count of cars, from expected counts would tell us if there is any bias or not.

Table 8:

Color	Dark	Attractive	Light	Total
Count of Cars (O)	13	15	22	50
Expected Count (E)	15	10	25	50

We measure this deviation into the statistics as $X = \sum(O - E)^2/E$ which will follow a Chi-Square distribution with degrees of freedom two (2), one less that number of categories.

The observed value of X here is 3.12, which has p-value of 0.209 (we always do right tailed test for Chi-Square test).

The above test is known as Goodness of Fit Test, as we are trying to compare one distribution with some expected distribution. This test has huge applicability when we build a model and compare its accuracy with expected values.

Chi-Square Test of Independence

The above concept of test is extended to a two-way table to test correlation or association between two categorical or ordinal variables.

Suppose, a survey was taken on viewers on a recently released movie where each individual was asked to rate the movie into either of "Excellent", "Very Good", "Good", "Not-So-Good", and "Bad/ Very bad" (Refer to Table 9).

Table 9:

Rating	Males	Females	Total
Outstanding	34	32	66
Very Good	23	25	48
Good	21	23	44
Not-So-Good	17	17	34
Bad / Very Bad	10	8	18
Total	105	105	210

Now we want to test if the rating differs across gender of the audience. So, our null hypothesis is H_0: There is not difference and the alternate is H_0: There is significant difference

Now as before, let's try to understand what should be the number of male viewers giving rating "Outstanding"? There are 50% male and 50% female in the sample. Hence, out of all 66 "Outstanding" ratings 50% would have come from males. Hence, 33 males should have given "Outstanding" rating. This is the "Expected Frequency" for the cell.

In general, expected frequency of any cell is given by,

$$E = \frac{Row\ Total\ X\ Column\ Total}{Grand\ Total}$$

Hence, the table of expected frequency is as below (Table 10):

Rating	Males	Females	Total
Outstanding	33	33	66
Very Good	24	24	48
Good	22	22	44
Not-So-Good	17	17	34
Bad / Very Bad	9	9	18
Total	105	105	210

As before, we calculate the test statistic as: $X = \sum\sum(O - E)^2/E$ summation is made over all cells.

The test statistic follows a Chi-Square distribution with $(k - 1)(l - 1)$ degrees of freedom, where k and l are the number of levels for two variables.

In the above example, k=2, l=5. The observed value of statistics turns out to be 0.457, which follows Chi-Square distribution with 4 degrees of freedom. The p-value of the test is 0.022

Parametric vs. Non-parametric Test

Note that in Z-test or t-Tests, or ANOVA we assumed that the data follows a normal distribution. But in Chi-square test we do not need to make any such assumption. This makes all statistical tests to be classified as either of “Parametric Test” and “Non-Parametric Test”

Parametric Test

In parametric test, we make distributional assumptions about the data. Most of the cases, it is assumed to follow normal distribution.

Some example of parametric tests are:

- Z Test
- t-test
- ANOVA

These tests are more popular that non-parametric tests as they are easy to execute.

Non-parametric Tests

In some tests, we do not (or we cannot) make any assumption about the distribution of the data. Then, we use non-parametric test. These tests, though more robust, are difficult to execute.

Some example of non-parametric tests are:

- Chi-square test
- Mood’s median test
- Kolmogorov Smirnov Test
- Chi-square test of independence

Write to us at

support@analyttica.com

USA Address

Analyttica Datalab Inc.
1007 N. Orange St, Floor-4, Wilmington, Delaware - 19801
Tel: +1 917 300 3289/3325

India Address

Analyttica Datalab Pvt. Ltd.
702, Brigade IRV Centre,2nd Main Rd, Nallurhalli,
Whitefield, Bengaluru - 560066.
Tel : +91 80 4650 7300