BUSINESS RESEARCH METHODS

UNIT 4

DR.G.SUDHAKAR

Hypothesis Testing in Statistics

Hypothesis testing is a tool for making statistical inferences about the population data. It is an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy. Hypothesis testing provides a way to verify whether the results of an experiment are valid.

A null hypothesis and an alternative hypothesis are set up before performing the hypothesis testing. This helps to arrive at a conclusion regarding the sample obtained from the population. Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

Two types of hypothesis testing in statistics

Null Hypothesis and Alternate Hypothesis

- The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected. H0 is the symbol for it, and it is pronounced H-naught.
- The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.

A sanitizer manufacturer claims that its product kills 95 percent of germs on average.

To put this company's claim to the test, create a null and alternate hypothesis.

- H0 (Null Hypothesis): Average = 95%.
- Alternative Hypothesis (H1): The average is less than 95%.

Another straightforward example to understand this concept is determining whether or not a coin is fair and balanced. The null hypothesis states that the probability of a show of heads is equal to the likelihood of a show of tails. In contrast, the alternate theory states that the probability of a show of heads and tails would be very different.

Simple and Composite Hypothesis Testing

Depending on the population distribution, you can classify the statistical hypothesis into two types.

- Simple Hypothesis: A simple hypothesis specifies an exact value for the parameter.
- Composite Hypothesis: A composite hypothesis specifies a range of values.

Example:

A company is claiming that their average sales for this quarter are 1000 units. This is an example of a simple hypothesis. Suppose the company claims that the sales are in the range of 900 to 1000 units. Then this is a case of a composite hypothesis.

One-Tailed and Two-Tailed Hypothesis Testing

- The One-Tailed test, also called a directional test, considers a critical region of data that would result in the null hypothesis being rejected if the test sample falls into it, inevitably meaning the acceptance of the alternate hypothesis. In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.
- In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided. If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

Example:

Suppose H0: mean = 50 and H1: mean not equal to 50

According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.

In a similar manner, if H0: mean >=50, then H1: mean <50

Here the mean is less than 50. It is called a One-tailed test.

TYPE 1 AND TYPE 2 ERROR

A hypothesis test can result in two types of errors.

- Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.
- Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Example:

Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.

H0: Student has passed

H1: Student has failed

- Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks [H0 was true].
- Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing marks [H1 is true].

Level of Significance

The alpha value is a criterion for determining whether a test statistic is statistically significant. In a statistical test, Alpha represents an acceptable probability of a Type I error. Because alpha is a probability, it can be anywhere between 0 and 1. In practice, the most commonly used alpha values are 0.01, 0.05, and 0.1, which represent a 1%, 5%, and 10% chance of a Type I error, respectively (i.e. rejecting the null hypothesis when it is in fact correct).

P-Value

A p-value is a metric that expresses the likelihood that an observed difference could have occurred by chance. As the p-value decreases the statistical significance of the observed difference increases. If the p-value is too low, you reject the null hypothesis. Here you have taken an example in which you are trying to test whether the new advertising campaign has increased the product's sales. The p-value is the likelihood that the null hypothesis, which states that there is no change in the sales due to the new advertising campaign, is true. If the p-value is .30, then there is a 30% chance that there is no increase or decrease in the product's sales. If the p-value is 0.03, then there is a 3% probability that there is no increase or decrease in the sales value due to the new advertising campaign. As you can see, the lower the p-value, the chances of the alternate hypothesis being true increases, which means that the new advertising campaign causes an increase or decrease in sales.

Large sample theory

The sample size n is greater than 30 (n≥30) it is known as large sample. For large samples the sampling distributions of statistic are normal (Z test). A study of sampling distribution of statistic for large sample is known as large sample theory.

Small sample theory

If the sample size n is less than 30 (n<30), it is known as small sample. For small samples the sampling distributions are t, F and χ2 distribution. A study of sampling distributions for small samples is known as small sample theory.

Types of Hypothesis Testing

Selecting the correct test for performing hypothesis testing can be confusing. These tests are used to determine a test statistic on the basis of which the null hypothesis can either be rejected or not rejected. Some of the important tests used for hypothesis testing are given below.

- **Hypothesis Testing Z Test:** A z test is a way of hypothesis testing that is used for a large sample size (n ≥ 30). It is used to determine whether there is a difference between the population mean and the sample mean when the population standard deviation is known. It can also be used to compare the mean of two samples. It is used to compute the z test statistic.
- **Hypothesis Testing t Test:** The t test is another method of hypothesis testing that is used for a small sample size (n < 30). It is also used to compare the sample mean and population mean. However, the population standard deviation is not known. Instead, the sample standard deviation is known. The mean of two samples can also be compared using the t test.
- **Hypothesis Testing Chi Square:** The Chi square test is a hypothesis testing method that is used to check whether the variables in a population are independent or not. It is used when the test statistic is chi-squared distributed.

TYPE I AND TYPE II ERRORS

Type I and Type II errors are subjected to the result of the null hypothesis. In case of type I or type-1 error, the null hypothesis is rejected though it is true whereas type II or type-2 error, the null hypothesis is not rejected even when the alternative hypothesis is true. Both the error type-i and type-ii are also known as “false negative”. A lot of statistical theory rotates around the reduction of one or both of these errors, still, the total elimination of both is explained as a statistical impossibility.

Type I Error

A type I error appears when the null hypothesis (H0) of an experiment is true, but still, it is rejected. It is stating something which is not present or a false hit. A type I error is often called a false positive (an event that shows that a given condition is present when it is absent). In words of community tales, a person may see the bear when there is none (raising a false alarm) where the null hypothesis (H0) contains the statement: “There is no bear”. The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true. It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type i error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.

Type II Error

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. It is losing to state what is present and a miss. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false. A type II error is assigned when a true alternative hypothesis is not acknowledged. In other words, an examiner may miss discovering the bear when in fact a bear is present (hence fails in raising the alarm). Again, H0, the null hypothesis, consists of the statement that, “There is no bear”, wherein, if a wolf is indeed present, is a type II error on the part of the investigator. Here, the bear either exists or does not exist within given circumstances, the question arises here is if it is correctly identified or not, either missing detecting it when it is present, or identifying it when it is not present. The rate level of the type II error is represented by the Greek letter β (beta) and linked to the power of a test (which equals 1−β).

PARAMETRIC AND NON-PARAMETRIC TEST

Parametric is a test in which parameters are assumed and the population distribution is always known. To calculate the central tendency, a mean value is used. These tests are common, and this makes performing research pretty straightforward without consuming much time. No assumptions are made in the Non-parametric test and it measures with the help of the median value. A few instances of Non-parametric tests are Kruskal-Wallis, Mann-Whitney, and so forth. In this article, you will be learning what is parametric and non-parametric tests, the advantages and disadvantages of parametric and nan-parametric tests, parametric and non-parametric statistics and the difference between parametric and non-parametric tests.

Parametric Test Definition

In Statistics, a parametric test is a kind of the hypothesis test which gives generalizations for generating records regarding the mean of the primary/original population. The t-test is carried out based on the students t-statistic, which is often used in that value. The t-statistic test holds on the underlying hypothesis which includes the normal distribution of a variable. In this case, the mean is known, or it is considered to be known. For finding the sample from the population, population variance is identified. It is hypothesized that the variables of concern in the population are estimated on an interval scale.

Non-Parametric Test Definition

The non-parametric test does not require any population distribution, which is meant by distinct parameters. It is also a kind of hypothesis test, which is not based on the underlying hypothesis. In the case of the non-parametric test, the test is based on the differences in the median. So, this kind of test is also called a distribution-free test. The test variables are determined on the nominal or ordinal level. If the independent variables are non-metric, the non-parametric test is usually performed.

Differences between the Parametric Test and the Non-Parametric Test

Properties	**Parametric Test**	**Non-Parametric Test**
Assumptions	Yes, assumptions are made	No, assumptions are not made
Value for central tendency	The mean value is the central tendency	The median value is the central tendency
Correlation	Pearson Correlation	Spearman Correlation
Probabilistic Distribution	Normal probabilistic distribution	Arbitrary probabilistic distribution
Population Knowledge	Population knowledge is required	Population knowledge is not required
Used for	Used for finding interval data	Used for finding nominal data

Application	Applicable to variables	Applicable to variables and attributes
Examples	T-test, z-test	Mann-Whitney, Kruskal-Wallis

Advantages and Disadvantages of Parametric and Nonparametric Tests

A lot of individuals accept that the choice between using parametric or nonparametric tests relies upon whether your information is normally distributed. The distribution can act as a deciding factor in case the data set is relatively small. Although, in a lot of cases, this issue isn't a critical issue because of the following reasons:

- Parametric tests help in analyzing non normal appropriations for a lot of datasets.
- Nonparametric tests when analyzed have other firm conclusions that are harder to achieve.
- The appropriate response is usually dependent upon whether the mean or median is chosen to be a better measure of central tendency for the distribution of the data.
- A parametric test is considered when you have the mean value as your central value and the size of your data set is comparatively large. This test helps in making powerful and effective decisions.
- A non-parametric test is considered regardless of the size of the data set if the median value is better when compared to the mean value.
- Ultimately, if your sample size is small, you may be compelled to use a nonparametric test. As the table shows, the example size prerequisites aren't excessively huge. On the off chance that you have a little example and need to utilize a less powerful nonparametric analysis, it doubly brings down the chances of recognizing an impact.
- The non-parametric test acts as the shadow world of the parametric test. In the table that is given below, you will understand the linked pairs involved in the statistical hypothesis tests.

Related Pairs of Parametric Test and Non-Parametric Tests

Parametric Tests for Means	**Non-Parametric Test for Medians**
1 - sample t - test	1 - sample Wilcoxon, 1 - sample sign

2 - sample t - test	Mann - Whitney Test
One - way ANOVA	Kruskal- Wallis, Mood’s median test
With a factor and a blocking variable - Factorial DOE	Friedman Test

Classification of Parametric Test and Non-Parametric Test

There are different kinds of parametric tests and non-parametric tests to check the data. Let us discuss them one by one.

Types of Parametric Test

- Student's T-Test:- This test is used when the samples are small and population variances are unknown. The test is used to do a comparison between two means and proportions of small independent samples and between the population mean and sample mean.
- 1 Sample T-Test:- Through this test, the comparison between the specified value and meaning of a single group of observations is done.
- Unpaired 2 Sample T-Test:- The test is performed to compare the two means of two independent samples. These samples came from the normal populations having the same or unknown variances.
- Paired 2 Sample T-Test:- In the case of paired data of observations from a single sample, the paired 2 sample t-test is used.
- ANOVA:- Analysis of variance is used when the difference in the mean values of more than two groups is given.
- One Way ANOVA:- This test is useful when different testing groups differ by only one factor.
- Two Way ANOVA:- When various testing groups differ by two or more factors, then a two way ANOVA test is used.
- Pearson's Correlation Coefficient:- This coefficient is the estimation of the strength between two variables. The test is used in finding the relationship between two continuous and quantitative variables.
- Z - Test:- The test helps measure the difference between two means.
- Z - Proportionality Test:- It is used in calculating the difference between two proportions.

Types of Non-Parametric Test

- 1 Sample Sign Test:- In this test, the median of a population is calculated and is compared to the target value or reference value.
- 1 Sample Wilcoxon Signed Rank Test:- Through this test also, the population median is calculated and compared with the target value but the data used is extracted from the symmetric distribution.
- Friedman Test:- The difference of the groups having ordinal dependent variables is calculated. This test is used for continuous data.
- Goodman Kruska's Gamma:- It is a group test used for ranked variables.
- Kruskal-Wallis Test:- This test is used when two or more medians are different. For the calculations in this test, ranks of the data points are used.
- The Mann-Kendall Trend Test:- The test helps in finding the trends in time-series data.
- Mann-Whitney Test:- To compare differences between two independent groups, this test is used. The condition used in this test is that the dependent values must be continuous or ordinal.
- Mood's Median Test:- This test is used when there are two independent samples.
- Spearman Rank Correlation:- This technique is used to estimate the relation between two sets of data.

Applications of Parametric Tests

- This test is used when the given data is quantitative and continuous.
- When the data is of normal distribution then this test is used.
- The parametric tests are helpful when the data is estimated on the approximate ratio or interval scales of measurement.

Applications of Non-Parametric Tests

These tests are used in the case of solid mixing to study the sampling results.

The tests are helpful when the data is estimated with different kinds of measurement scales.

The non-parametric tests are used when the distribution of the population is unknown.

T-TEST

1. It is a parametric test of hypothesis testing based on Student’s T distribution.

2. It is essentially, testing the significance of the difference of the mean values when the sample size is small (i.e, less than 30) and when the population standard deviation is not available.

3. Assumptions of this test:

- Population distribution is normal, and
- Samples are random and independent
- The sample size is small.
- Population standard deviation is not known.

4. Mann-Whitney ‘U’ test is a non-parametric counterpart of the T-test.

A T-test can be a:

One Sample T-test: To compare a sample mean with that of the population mean.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Where,

- x̄ is the sample mean
- s is the sample standard deviation
- n is the sample size
- µ is the population mean

Two-Sample T-test: To compare the means of two different samples.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1{}^2}{N_1} + \frac{S_2{}^2}{N_2}}}$$

Where,

- x̄1 is the sample mean of the first group
- x̄2 is the sample mean of the second group
- S1 is the sample-1 standard deviation
- S2 is the sample-2 standard deviation
- n is the sample size

Conclusion:

- If the value of the test statistic is greater than the table value -> rejects the null hypothesis.

- If the value of the test statistic is less than the table value -> do not reject the null hypothesis.

Z-Test

1. It is a parametric test of hypothesis testing.

2. It is used to determine whether the means are different when the population variance is known and the sample size is large (i.e, greater than 30).

3. Assumptions of this test:

- Population distribution is normal
- Samples are random and independent.
- The sample size is large.
- Population standard deviation is known.

A Z-test can be:

One Sample Z-test: To compare a sample mean with that of the population mean.

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

$\bar{x}$	= sample mean
μ	= population mean
σ	=population standard deviation
n	=sample size

Two Sample Z-test: To compare the means of two different samples.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Where,

- x̄1 is the sample mean of 1st group
- x̄2 is the sample mean of 2nd group
- σ1 is the population-1 standard deviation
- σ2 is the population-2 standard deviation
- n is the sample size

F-Test

1. It is a parametric test of hypothesis testing based on Snedecor F-distribution.

2. It is a test for the null hypothesis that two normal populations have the same variance.

3. An F-test is regarded as a comparison of equality of sample variances.

4. F-statistic is simply a ratio of two variances.

5. It is calculated as:

F = s12/s22

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

6. By changing the variance in the ratio, F-test has become a very flexible test. It can then be used to:

- Test the overall significance for a regression model.
- To compare the fits of different models and
- To test the equality of means.

7. Assumptions of this test:

- Population distribution is normal, and
- Samples are drawn randomly and independently.

ANOVA

1. Also called as Analysis of variance, it is a parametric test of hypothesis testing.

2. It is an extension of the T-Test and Z-test.

3. It is used to test the significance of the differences in the mean values among more than two sample groups.

4. It uses F-test to statistically test the equality of means and the relative variance between them.

5. Assumptions of this test:

- Population distribution is normal, and
- Samples are random and independent.
- Homogeneity of sample variance.

6. One-way ANOVA and Two-way ANOVA are is types.

7. F-statistic = variance between the sample means/variance within the sample

CHI-SQUARE TEST

1. It is a non-parametric test of hypothesis testing.

2. As a non-parametric test, chi-square can be used:

- test of goodness of fit.
- as a test of independence of two variables.

3. It helps in assessing the goodness of fit between a set of observed and those expected theoretically.

4. It makes a comparison between the expected frequencies and the observed frequencies.

5. Greater the difference, the greater is the value of chi-square.

6. If there is no difference between the expected and observed frequencies, then the value of chi-square is equal to zero.

7. It is also known as the "Goodness of fit test" which determines whether a particular distribution fits the observed data or not.

8. It is calculated as:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = *the frequencies observed*

E = *the frequencies expected*

$\sum$ = *the 'sum of'*

9. Chi-square is also used to test the independence of two variables.

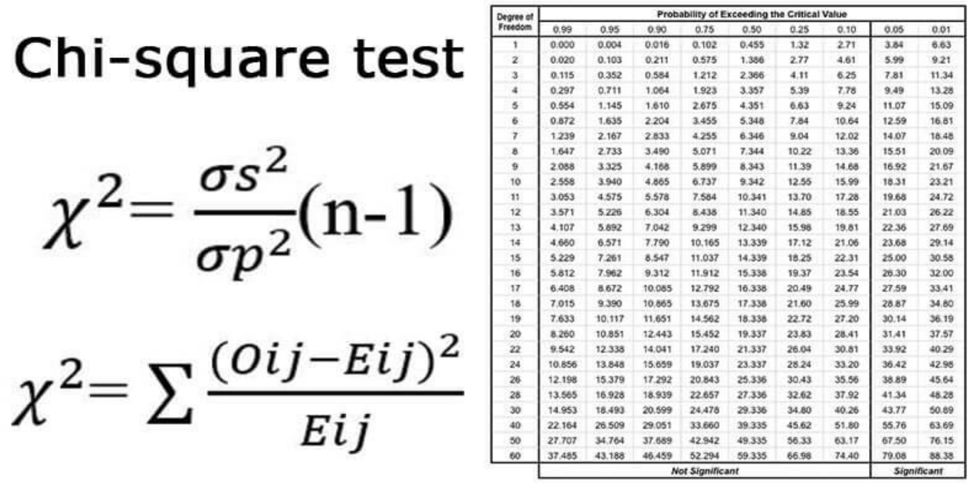
10. Conditions for chi-square test:

- Randomly collect and record the Observations.
- In the sample, all the entities must be independent.
- No one of the groups should contain very few items, say less than 10.
- The reasonably large overall number of items. Normally, it should be at least 50, however small the number of groups may be.

11. Chi-square as a parametric test is used as a test for population variance based on sample variance.

12. If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by (n-1), where n means the number of items in the sample, we get the values of chi-square.

13. It is calculated as:



Chi-square test

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

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

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
	Not Significant							*Significant*	

Mann-Whitney U-Test

1. It is a non-parametric test of hypothesis testing.

2. This test is used to investigate whether two independent samples were selected from a population having the same distribution.

3. It is a true non-parametric counterpart of the T-test and gives the most accurate estimates of significance especially when sample sizes are small and the population is not normally distributed.

4. It is based on the comparison of every observation in the first sample with every observation in the other sample.

5. The test statistic used here is “U”.

6. Maximum value of “U” is ‘n1*n2’ and the minimum value is zero.

7. It is also known as:

Mann-Whitney Wilcoxon Test.

Mann-Whitney Wilcoxon Rank Test.

8. Mathematically, U is given by:

U1 = R1 – n1 (n1+1)/2

Where n1 is the sample size for sample 1, and R1 is the sum of ranks in Sample 1.

U2 = R2 – n2 (n2+1)/2

When consulting the significance tables, the smaller values of U1 and U2 are used. The sum of two values is given by,

U1 + U2 = {R1 – n1 (n1+1)/2} + { R2 – n2 (n2+1)/2 }

Knowing that R1+R2 = N (N+1)/2 and N=n1+n2, and doing some algebra, we find that the sum is:

U1 + U2 = n1*n2

Kruskal-Wallis H-test

1. It is a non-parametric test of hypothesis testing.
2. This test is used for comparing two or more independent samples of equal or different sample sizes.
3. It extends the Mann-Whitney-U-Test which is used to comparing only two groups.
4. One-Way ANOVA is the parametric equivalent of this test. And that's why it is also known as 'One-Way ANOVA on ranks.
5. It uses ranks instead of actual data.
6. It does not assume the population to be normally distributed.
7. The test statistic used here is "H".

DIFFERENT LEVELS OF DATA ANALYSIS

When it comes to the level of analysis in statistics, there are three different analysis techniques that exist. These are –

- Univariate analysis
- Bivariate analysis
- Multivariate analysis

The selection of the data analysis technique is dependent on the number of variables, types of data and focus of the statistical inquiry. The following section describes the three different levels of data analysis –

Univariate analysis

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

Here is one example of Univariate analysis-

In a survey of a class room, the researcher may be looking to count the number of boys and girls. In this instance, the data would simply reflect the number, i.e. a single variable and its quantity as per the below table. The key objective of Univariate analysis is to simply describe the data to find patterns within the data. This is be done by looking into the mean, median, mode, dispersion, variance, range, standard deviation etc.

Univariate analysis is conducted through several ways which are mostly descriptive in nature-

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

Bivariate analysis

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Here is one simple example of bivariate analysis –

In a survey of a classroom, the researcher may be looking to analysis the ratio of students who scored above 85% corresponding to their genders. In this case, there are two variables – gender = X (independent variable) and result = Y (dependent variable). A Bivariate analysis is will measure the correlations between the two variables.

Bivariate analysis is conducted using –

- Correlation coefficients
- Regression analysis

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

Here is an example of multivariate analysis –

A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits? In this instance, a multivariate analysis would be required to understand the relationship of each variable with each other.

Commonly used multivariate analysis technique include –

- Factor Analysis

- Cluster Analysis
- Variance Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Principal Component Analysis
- Redundancy Analysis

FACTOR ANALYSIS

Factor analysis is a way to take a mass of data and shrinking it to a smaller data set that is more manageable and more understandable. It's a way to find hidden patterns, show how those patterns overlap and show what characteristics are seen in multiple patterns. It is also used to create a set of variables for similar items in the set (these sets of variables are called dimensions). It can be a very useful tool for complex sets of data involving psychological studies, socioeconomic status and other involved concepts. A "factor" is a set of observed variables that have similar response patterns; they are associated with a hidden variable (called a confounding variable) that isn't directly measured. Factors are listed according to factor loadings, or how much variation in the data they can explain.

The two types: exploratory and confirmatory.

- Exploratory factor analysis is if you don't have any idea about what structure your data is or how many dimensions are in a set of variables.
- Confirmatory Factor Analysis is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

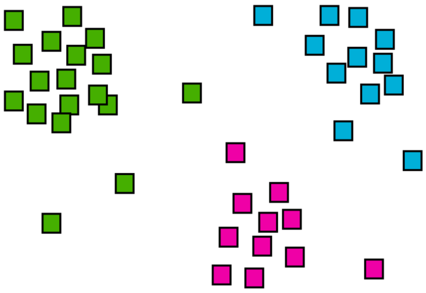
DISCRIMINANT ANALYSIS

Discriminant analysis is a versatile statistical method often used by market researchers to classify observations into two or more groups or categories. In other words, discriminant analysis is used to assign objects to one group among a number of known groups.

Discriminant analysis is most often used to help a researcher predict the group or category to which a subject belongs. For example, when individuals are interviewed for a job, managers will not know for sure how job candidates will perform on the job if hired. Suppose, however, that a human resource manager has a list of current employees who have been classified into two groups: "high performers" and "low performers." These individuals have been working for the company for some time, have been evaluated by their supervisors, and are known to fall into one of these two mutually exclusive categories. The manager also has information on the employees' backgrounds: educational attainment, prior work experience, participation in training programs, work attitude measures, personality characteristics, and so forth. This information was known at the time these employees were hired. The manager wants to be able to predict, with some confidence, which future job candidates are high performers and which are not. A researcher or consultant can use discriminant analysis, along with existing data, to help in this task.

CLUSTER ANALYSIS

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg. products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc. When we try to group a set of objects that have similar kind of characteristics, attributes these groups are called clusters. The process is called clustering. It is a very difficult task to get to know the properties of every individual object instead, it would be easy to group those similar objects and have a common structure of properties that the group follows.



CORRELATION AND REGRESSION

Correlation and regression are statistical measurements that are used to give a relationship between two variables. For example, suppose a person is driving an expensive car then it is assumed that she must be financially well. To numerically quantify this relationship, correlation and regression are used.

Correlation

Correlation can be defined as a measurement that is used to quantify the relationship between variables. If an increase (or decrease) in one variable causes a corresponding increase (or decrease) in another then the two variables are said to be directly correlated. Similarly, if an increase in one causes a decrease in another or vice versa, then the variables are said to be indirectly correlated. If a change in an independent variable does not cause a change in the dependent variable then they are uncorrelated. Thus, correlation can be positive (direct correlation), negative (indirect correlation), or zero. This relationship is given by the correlation coefficient.

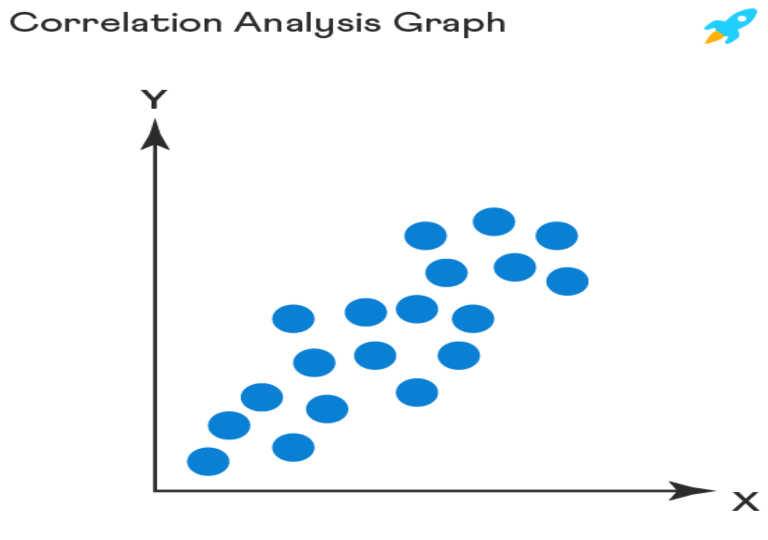
Regression

Regression can be defined as a measurement that is used to quantify how the change in one variable will affect another variable. Regression is used to find the cause and effect between two variables. Linear regression is the most commonly used type of regression because it is easier to analyze as compared to the rest. Linear regression is used to find the line that is the best fit to establish a relationship between variables.

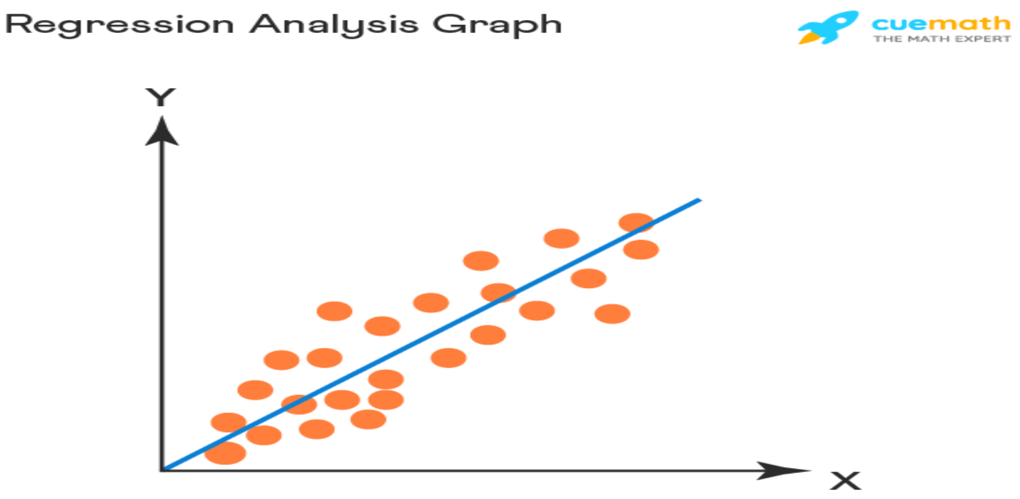
Correlation and Regression Analysis

Both correlation and regression analysis are done to quantify the strength of the relationship between two variables by using numbers. Graphically, correlation and regression analysis can be visualized using scatter plots.

Correlation analysis is done so as to determine whether there is a relationship between the variables that are being tested. Furthermore, a correlation coefficient such as Pearson's correlation coefficient is used to give a signed numeric value that depicts the strength as well as the direction of the correlation. The scatter plot gives the correlation between two variables x and y for individual data points as shown below.



Regression analysis is used to determine the relationship between two variables such that the value of the unknown variable can be estimated using the knowledge of the known variables. The goal of linear regression is to find the best-fitted line through the data points. For two variables, x, and y, the regression analysis can be visualized as follows:



MULTI-DIMENSIONAL SCALING

Multi-dimensional scaling (MDS) is a statistical technique that allows researchers to find and explore underlying themes, or dimensions, in order to explain similarities or dissimilarities (i.e. distances) between investigated datasets. You can analyse any kind of similarity or dissimilarity matrix using multi-dimensional scaling. Plotting these data sets on a multi-dimensional scale allows for easier interpretation and comparison by researchers than a linear dataset permits.

A possible example of when multi-dimensional scaling (MDS) might be used is if we have six utility companies and we want to understand how they are considered differently by respondents. We would invite consumers to complete a survey in which each of the six companies would be paired with each of the others, and the respondents would be asked in a series of scale based questions how similar they believe them to be, for a number of attributes. Examples of attributes may be: quality, service and price.