STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
   a) Total Variation = Residual Variation – Regression Variation
   b) Total Variation = Residual Variation + Regression Variation
   c) Total Variation = Residual Variation * Regression Variation
   d) All of the mentioned

Ans: b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called______outcomes.
   a) random
   b) direct
   c) binomial
   d) none of the mentioned

Ans: c) binomial

3. How many outcomes are possible with Bernoulli trial?
   a) 2
   b) 3
   c) 4
   d) None of the mentioned

Ans: a) 2

4. If Ho is true and we reject it is called
   a) Type-I error
   b) Type-II error
   c) Standard error
   d) Sampling error

Ans: a) Type- I error

5. Level of significance is also called:
   a) Power of the test
   b) Size of the test
   c) Level of confidence
   d) Confidence coefficient

Ans: a) Power of the test

6. The chance of rejecting a true hypothesis decreases when sample size is: a) Decrease
   b) Increase
   c) Both of them
   d) None

Ans: a) Decrease

7. Which of the following testing is concerned with making decisions using data? a) Probability
   b) Hypothesis
   c) Causal

d) None of the mentioned

Ans: b) Hypothesis

8. What is the purpose of multiple testing in statistical inference?
   a) Minimize errors
   b) Minimize false positives
   c) Minimize false negatives
   d) All of the mentioned

Ans: d) All of the mentioned

9. Normalized data are centred at_and have units equal to standard deviations of the original data a) 0
   b) 5
   c) 1
   d) 10

Ans: a) 0

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

Ans: **Bayes' theorem** describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different color balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

If A and B are two events, then the **formula for the Bayes theorem** is given by:

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

where P(B)≠0

Where P(A|B) is the probability of condition when event A is occurring while event B has already occurred.

Bayes Theorem Applications

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. Bayesian inference has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc. For example, we can use Bayes' theorem to define the accuracy of medical test results by considering how likely any given person is to have a disease and the test's overall accuracy. Bayes' theorem relies on consolidating prior probability distributions to generate posterior probabilities. In Bayesian statistical inference, prior probability is the probability of an event before new data is collected.

11. What is z-score?

Ans: **Z-score is also known as standard score** gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).

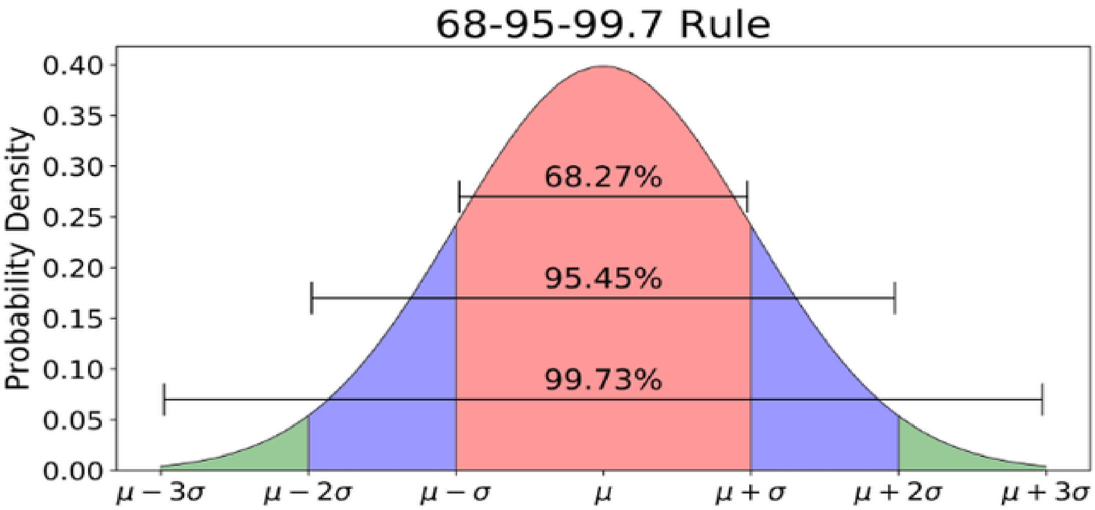
A z-score can be calculated using the following formula.

$$\mathbf{z = (X - \mu) / \sigma}$$

where,
z = Z-Score,
X = The value of the element,
μ = The population mean, and
σ = The population standard deviation

Interpretation of Z-score

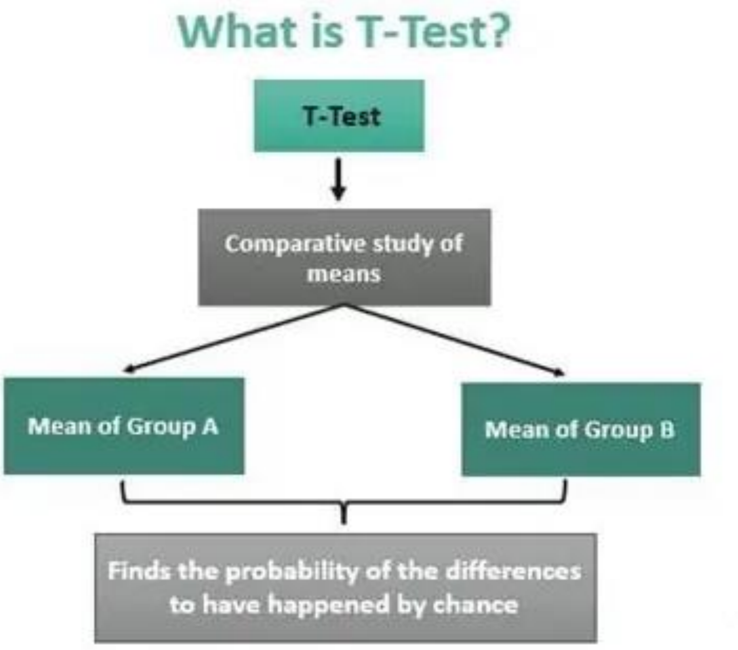
- An element having a z-score less than 0 represents that the element is less than the mean.
- An element having a z-score greater than 0 represents that the element is greater than the mean.
- An element having a z-score equal to 0 represents that the element is equal to the mean.
- An element having a z-score equal to 1 represents that the element is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean, and so on.
- An element having a z-score equal to -1 represents that the element is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean, and so on.
- If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3. This is known as the Empirical Rule or the 68-95-99.7 Rule and can be demonstrated in the image below



The 68-95-99.7 Rule for a Normal Distribution

12. What is t-test?

A T-test is the final statistical measure for determining differences between two means that may or may not be related. The testing uses randomly selected samples from the two categories or groups. It is a statistical method in which samples are chosen randomly, and there is no perfect normal distribution.


What is T-Test?
T-Test
Comparative study of means
Mean of Group A
Mean of Group B
Finds the probability of the differences to have happened by chance

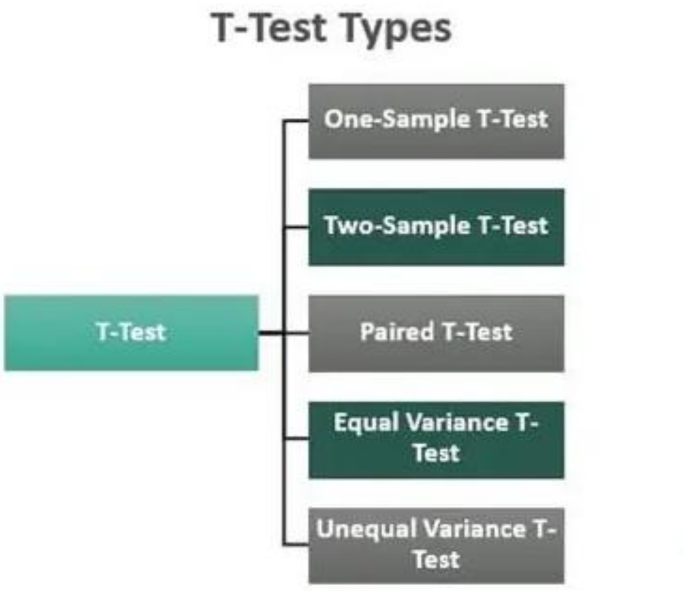
A T-test studies a set of data gathered from two similar or different groups to determine the probability of the difference in the result than what is usually obtained. The accuracy of the test depends on various factors, including the distribution patterns used and the variants influencing the collected samples. Depending on the parameters, the test is conducted, and a T-value is obtained as the statistical inference of the probability of the usual resultant being driven by chance.

For example, if one wishes to figure out if the mean of the length of petals of a flower belonging to two different species is the same, a T-test can be done. The user can select petals randomly from two other species of that flower and come to a standard conclusion. The final **T-test interpretation** could be obtained in either of the two ways:

- A **null hypothesis** signifies that the difference between the means is zero and where both the means are shown as equal.
- An alternate hypothesis implies the difference between the means is different from zero. This hypothesis rejects the null hypothesis, indicating that the data set is quite accurate and not by chance.

This T-test, however, is only valid and should be done when the mean or average of only two categories or groups needs to be compared. As soon as the number of comparisons to be made is more than two, conducting this is not recommended.

Some of the widely used **T-test types** are as follows:



#1 – One-Sample T-Test

While performing this test, the mean or average of one group is compared against the set average, which is either the theoretical value or means of the population. For example, a teacher wishes to figure out the average height of the students of class 5 and compare the same against a set value of more than 45 kgs.

The teacher first randomly selects a group of students and records individual weights to achieve this. Next, she finds out the mean weight for that group and checks if it meets the standard set value of 45+. The formula used to obtain one-sample t-test results is:

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

Where,

- T = t-statistic
- m = mean of the group
- µ = theoretical mean value of the population
- s = standard deviation of the group
- n = sample size

#2 – Independent Two-Sample T-Test

This is the test conducted when samples from two different groups, species, or populations are studied and compared. It is also known as an independent T-test. For example, if a teacher wants to compare the height of male students and female students in class 5, she would use the independent two-sample test.

The **T-test formula** used to calculate this is:

$$t = \frac{m_A - m_B}{\frac{\sqrt{s^2}}{\sqrt{n_A}} + \frac{\sqrt{s^2}}{\sqrt{n_B}}}$$

Where,

- mA – mB = means of samples from two different groups or populations
- nA – nB = respective sample sizes
- s2 = standard deviation or common variance of two samples

#3 – Paired Sample T-Test

This hypothesis testing is conducted when two groups belong to the same population or group. The groups are studied either at two different times or under two varied conditions. The formula used to obtain the t-value is:

$$t = \frac{m}{s/\sqrt{n}}$$

Where,

- T = t-statistic
- m = mean of the group
- s = standard deviation of the group
- n = sample size

#4 – Equal Variance T-Test

This test is conducted when the sample size in each group or population is the same or the variance of the two data sets is similar. It is also referred to as pooled T-test. The formula applied here is as follows:

$$T - value = \frac{mean1 - mean2}{(n1-1)\ x\ var1^2 + \ (n2-1)\ x\ var2^2 \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

Where,

- Mean1 and mean2 = average value of each set of samples
- var1 and var2 = variance of each set of samples
- n1 and n2 = number of records in each set

#5 – Unequal Variance T-Test

The unequal variance testing is used when the variance and the number of samples in each group are different. It is often referred to as Welch’s test, and the formula is:

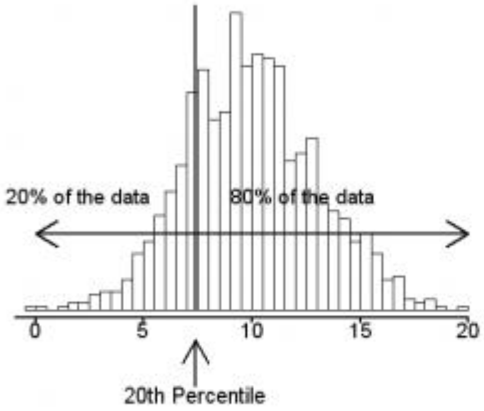
$$T - value = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}}$$

Where,

- mean1 and mean2 = Average value of each set of samples
- var1 and var2 = Variance of each set of samples
- n1 and n2 = number of records in each set

13. What is percentile?

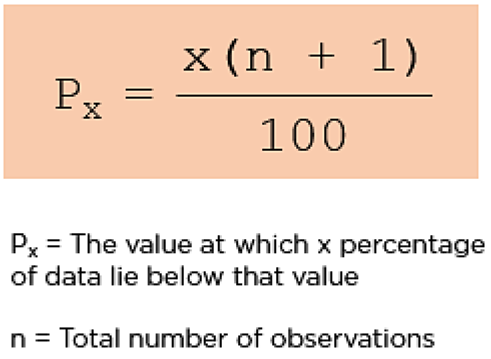
In statistics, a percentile is a term that describes how a score compares to other scores from the same set. The most common definition of a percentile is a number where **a certain percentage of scores fall below that number.** You might know that you scored 67 out of 90 on a test. But that figure has no real meaning unless you know what percentile you fall into. If you know that your score is in the 90th percentile, that means you scored better than 90% of people who took the test.



Percentiles show how a given value compares to others. The general rule is that if a value is in the kth percentile, it is greater than K per cent of the total values.

Raw test scores are frequently uninformative. When you get a score on the SAT, CAT, or GRE, the units themselves are meaningless. A total CAT score of 120 isn't necessarily significant. Instead, you're more interested in knowing what percentage of test-takers you outperformed. A total score of 120 on the CAT is roughly in the 90th percentile. Congratulations, you scored better than 90% of the other test-takers. Only 10% scored better than you.

You can calculate percentiles in statistics using the following formula:

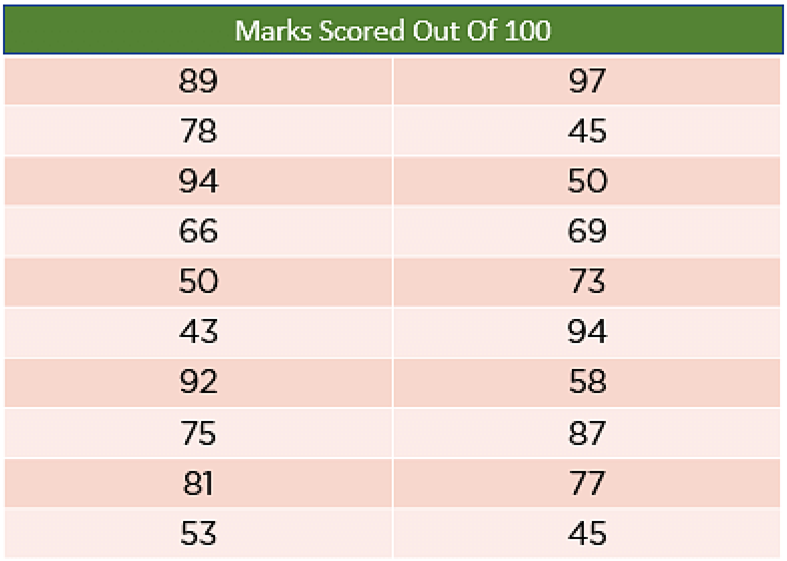


$$P_x = \frac{x(n + 1)}{100}$$

P_x = The value at which x percentage of data lie below that value

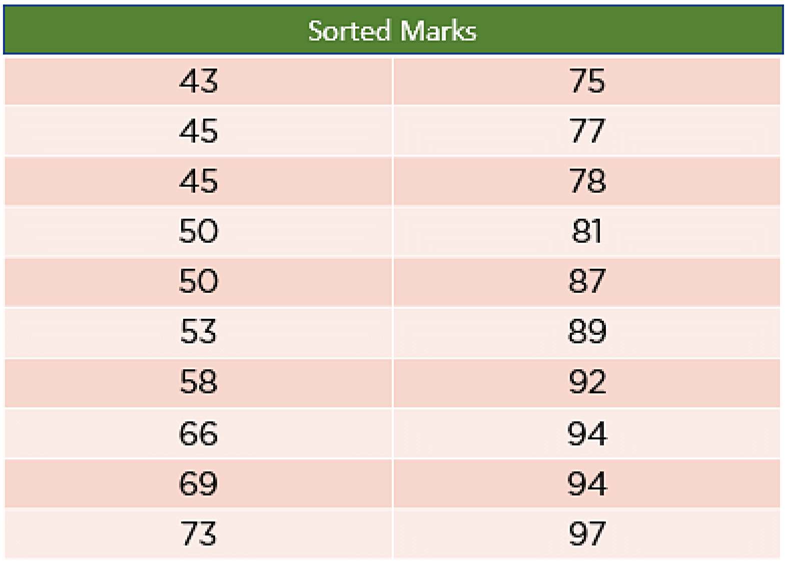
n = Total number of observations

For example:
Imagine you have the marks of 20 students. Now, try to calculate the 90th percentile.



Marks Scored Out Of 100	
89	97
78	45
94	50
66	69
50	73
43	94
92	58
75	87
81	77
53	45

Step 1: Arrange the score in ascending order.



Sorted Marks	
43	75
45	77
45	78
50	81
50	87
53	89
58	92
66	94
69	94
73	97

Step 2: Plug the values in the formula to find n.

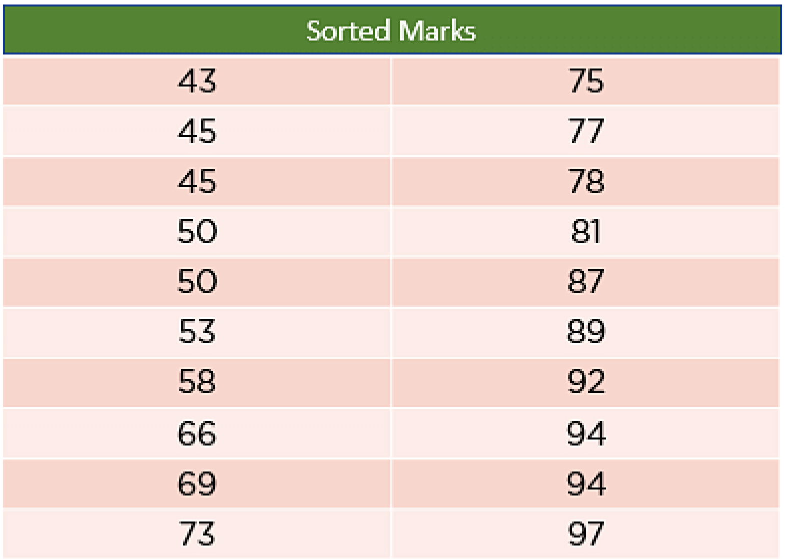
$$P_{90} = \frac{90(20 + 1)}{100}$$

$$P_{90} = \frac{1890}{100}$$

$$P_{90} = 18.9 \sim 19$$

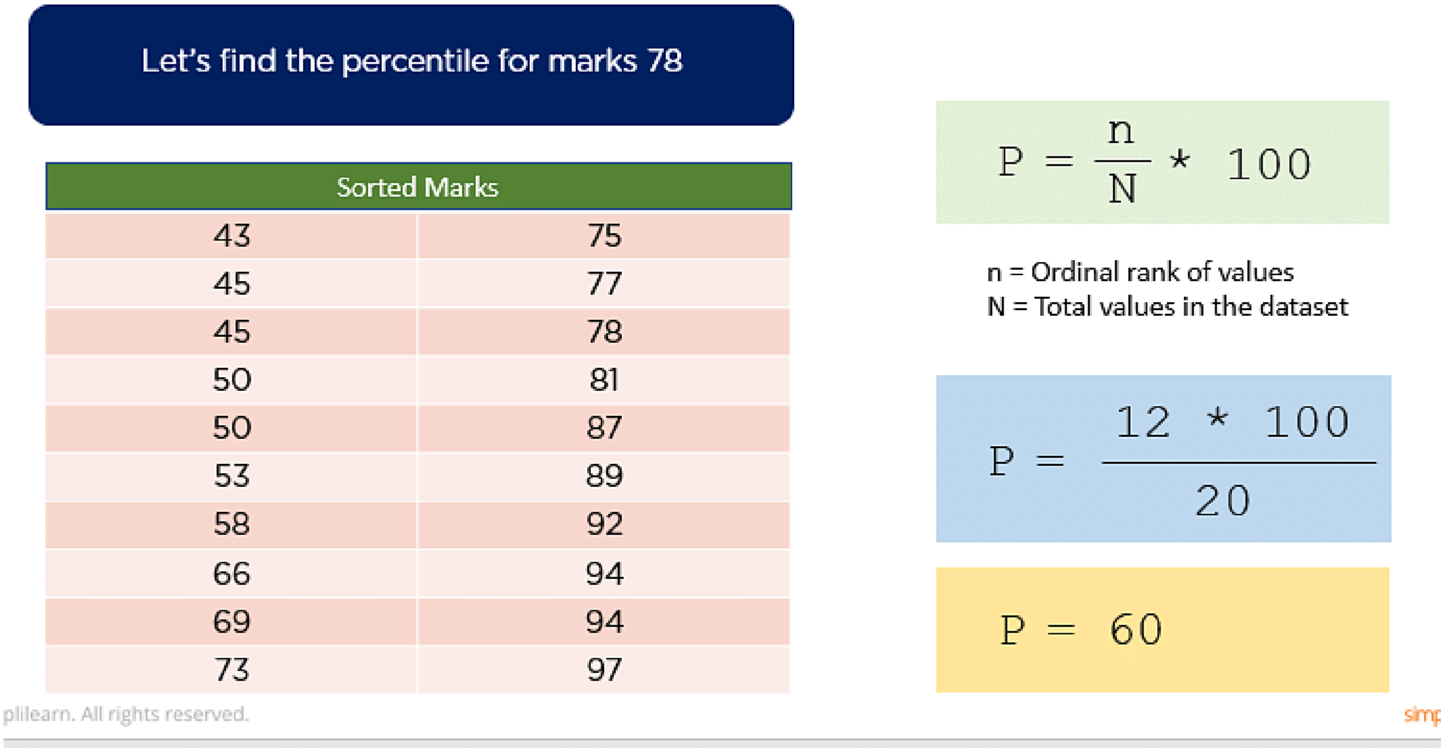
$$P_{90} = 94$$

P90 = 94 means that 90% of students got less than 94 and 10% of students got more than 94
Let’s look at another way how you can find the percentile in statistics.
Suppose you want to find the percentile mark of 78 marks in the data set.
Step 1: Sort the marks in ascending order.



Sorted Marks	
43	75
45	77
45	78
50	81
50	87
53	89
58	92
66	94
69	94
73	97

Step 2: Substitute the value in the formula.



Let’s find the percentile for marks 78

Sorted Marks	
43	75
45	77
45	78
50	81
50	87
53	89
58	92
66	94
69	94
73	97

$$P = \frac{n}{N} * 100$$

n = Ordinal rank of values
N = Total values in the dataset

$$P = \frac{12 * 100}{20}$$

$$P = 60$$

simp

P = 60 means that 78 marks point to the 60th percentile in the dataset.

14. What is ANOVA?

Ans: ANOVA Test is used to analyze the differences among the means of various groups using certain estimation procedures. ANOVA means analysis of variance. ANOVA test is a statistical significance test that is used to check whether the null hypothesis can be rejected or not during hypothesis testing.

An ANOVA test can be either one-way or two-way depending upon the number of independent variables.

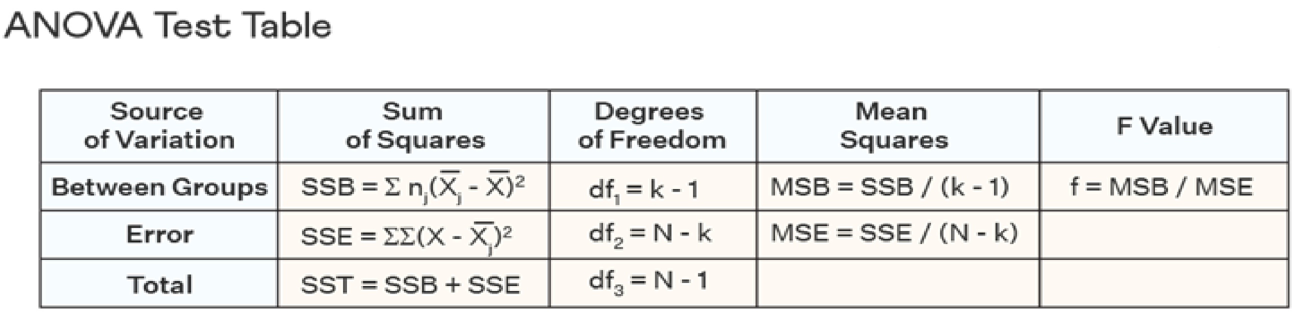
ANOVA Test

ANOVA test is used to check whether the means of three or more populations are equal or not. The ANOVA test applies when there are more than two independent groups. The goal of the ANOVA test is to check for variability within the groups as well as the variability among the groups. The ANOVA test statistic is given by the f test.

ANOVA Test Definition

ANOVA test can be defined as a type of test used in hypothesis testing to compare whether the means of two or more groups are equal or not. This test is used to check if the null hypothesis can be rejected or not depending upon the statistical significance exhibited by the parameters. The decision is made by comparing the ANOVA test statistic with the critical value.

ANOVA Formula

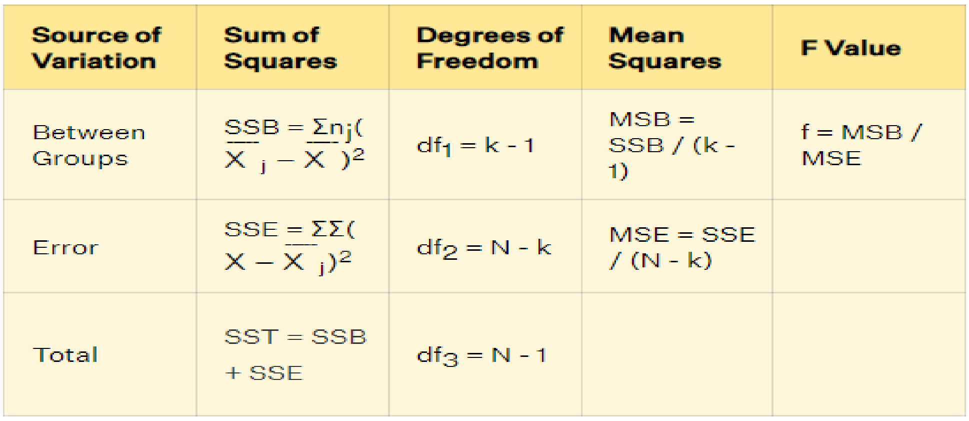


ANOVA Test Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

ANOVA Table

The ANOVA formulas can be arranged systematically in the form of a table. This ANOVA table can be summarized as follows:



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

One Way ANOVA

The one way ANOVA test is used to determine whether there is any difference between the means of three or more groups. A one way ANOVA will have only one independent variable. The hypothesis for a one way ANOVA test can be set up as follows:

Null Hypothesis, H0H0**:** μ1μ1 = μ2μ2 = μ3μ3 = ... = μkμk

Alternative Hypothesis, H1H1**:** The means are not equal

Decision Rule: If test statistic > critical value then reject the null hypothesis and conclude that the means of at least two groups are statistically significant.

The steps to perform the one way ANOVA test are given below:

- **Step 1:** Calculate the mean for each group.
- **Step 2:** Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.
- **Step 3:** Calculate the SSB.
- **Step 4:** Calculate the between groups degrees of freedom.
- **Step 5:** Calculate the SSE.
- **Step 6:** Calculate the degrees of freedom of errors.
- **Step 7:** Determine the MSB and the MSE.
- **Step 8:** Find the f test statistic.
- **Step 9:** Using the f table for the specified level of significance, αα, find the critical value. This is given by F(αα, df_1. df_2).
- **Step 10:** If f > F then reject the null hypothesis.

Limitations of One-Way ANOVA Test

The one-way ANOVA is an omnibus test statistic. This implies that the test will determine whether the means of the various groups are statistically significant or not. However, it cannot distinguish the specific groups that have a statistically significant mean. Thus, to find the specific group with a different mean, a post hoc test needs to be conducted.

Two Way ANOVA

The two-way ANOVA has two independent variables. Thus, it can be thought of as an extension of a one-way ANOVA where only one variable affects the dependent variable. A two-way ANOVA test is used to check the main effect of each independent variable and to see if there is an interaction effect between them. To examine the main effect, each factor is considered separately as done in a one-way ANOVA. Furthermore, to check the interaction effect, all factors are considered at the same time. There are certain assumptions made for a two-way ANOVA test. These are given as follows:

- The samples drawn from the population must be independent.
- The population should be approximately normally distributed.
- The groups should have the same sample size.
- The population variances are equal

Suppose in the two-way ANOVA example, as mentioned above, the income groups are low, middle, high. The gender groups are female, male, and transgender. Then there will be 9 treatment groups and the three hypotheses can be set up as follows:

H01H01: All income groups have equal mean anxiety.

H11H11: All income groups do not have equal mean anxiety.
H02H02: All gender groups have equal mean anxiety.
H12H12: All gender groups do not have equal mean anxiety.
H03H03: Interaction effect does not exist
H13H13: Interaction effect exists.

Important Notes on ANOVA Test

- ANOVA test is used to check whether the means of three or more groups are different or not by using estimation parameters such as the variance.
- An ANOVA table is used to summarize the results of an ANOVA test.
- There are two types of ANOVA tests - one way ANOVA and two-way ANOVA
- One way ANOVA has only one independent variable while a two-way ANOVA has two independent variables.

15. How can ANOVA help?

Ans:

- ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.
- The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent. When you understand how each independent variable’s mean is different from the others, you can begin to understand which of them has a connection to your dependent, and begin to learn what is driving that behavior.
- ANOVA can also be used in feature selection process of machine learning. The features can be compared by performing an ANOVA test and similar ones can be eliminated from the feature set.