Module-3

Statistical Fundamental-II

Inferential Statistics

- Inferential statistics is technique that allow us to use these samples to make generalizations about the populations from which the samples were drawn. It is, therefore, important that the sample accurately represents the population. The process of achieving this is called sampling

- Often, however, you do not have access to the whole population you are interested in investigating, but only a limited number of data instead.

- For example, you might be interested in the exam marks of all students in the UK. It is not feasible to measure all exam marks of all students in the whole of the UK so you have to measure a smaller sample of students (e.g., 100 students), which are used to represent the larger population of all UK students.

Difference between Descriptive statistics and Inferential Statistics

DESCRIPTIVE STATISTICS

- Descriptive statistics Work with smaller data. There is no need for sampling and the entire population data is available.
- Process is simpler to do
- Descriptive statistics are likely to be 100% accurate because there is no assumption.
- Find result are shown in form of charts, table, graphs.
- Tool-Measure of central tendency (mean, median, mode) spread of data (range standard deviation)
- Organize, analyze and present data in meaningful manner.

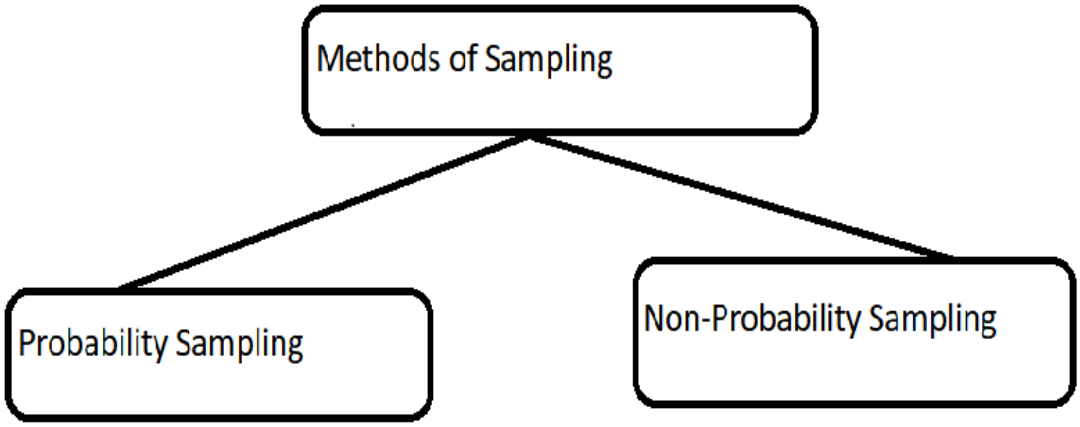
INFRENTIAL STATISTICS

- Inferential statistics work with large data set. Analyzing entire population based on sample parameter is a strength.
- Process is complex as we have to decide best sampling technique.

- This is not 100% accurate inferential statistics always make inference about larger population based on sample.
- Find result in probability score.
- Tool-hypothesis test, analysis of variance.
- Compress, test and predicts future outcome.

Method of Sampling

- There are different sampling techniques available, and they can be subdivided into two groups.



Probability Sampling

- Probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying.

Non-Probability sampling

- Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying.
- These 10,000 students are our population (N). Each of the 10,000 students is known as a unit (although sometimes other terms are used to describe a unit; see Sampling: The basics). In order to select a sample (n) of students from this population of 10,000 students.

Achieving a representative sample

- A critical component of probability sampling is the need to create a sample that is representative of the population. The more representative the sample is of the population, the more confident we can be when making statistical inferences (i.e., generalizations) from the sample to the population of interest.

- If all units within the population were identical in all respects there would be no need to sample at all.

- Under this scenario of perfect homogeneity of units, we could simply study a single unit since this would reflect the population perfectly.

Probability Sampling

Simple Random Sampling

- With simple random sampling, there is an equal chance (probability) that each of the 10,000 students could be selected for inclusion in our sample.
- If our desired sample size was around 200 students, were would select 200 students at random, probably using random number tables.

Systematic Sampling

- Systematic random sample is a variation on the simple random sample. Like simple random sampling, there is an equal chance (probability) that each of the 10,000 students could be selected for inclusion in our sample.
- Whilst you typically use random number tables to select the first unit for inclusion in your sample, the remaining units are selected in an ordered way (e.g., every 9th student).

Stratified Sampling

- Unlike the simple random sample and the systematic random sample, sometimes we are interested in particular strata (meaning groups) within the population (e.g., males vs. females; houses vs. apartments, etc.).

- With the stratified random sample, there is an equal chance (probability) of selecting each unit from within a particular stratum (group) of the population when creating the sample.

Clustered Sampling

- It is a method where we divide the entire population into sections or clusters that represent a population.
- Clusters are identified and included in a sample on the basis of defining demographic parameter such as age, location, sex etc.

Non-probability sampling

Quota sampling

- With proportional quota sampling, the aim is to end up with a sample where the strata (groups) being studied (e.g., males vs. females students) are proportional to the population being studied.
- If we were to examine the differences in male and female students, for example, the number of students from each group that we would include in the sample would be based on the proportion of male and female students amongst the 10,000 university students.

Convenience sampling

A convenience sample is simply one where the units that are selected for inclusion in the sample are the easiest to access. In our example of the 10,000 university students, if we were only interested in achieving a sample size of say 100 students, we may simply stand at one of the main entrances to campus, where it would be easy to invite the many students that pass by to take part in the research.

Snowball Sampling

Snowball sampling is particularly appropriate when the population you are interested in is hidden and/or hard-to-reach. These include populations such as drug addicts, homeless people, individuals with AIDS/HIV, prostitutes, and so forth.

Judgment Sampling

- Also known as selective, or subjective, sampling, this technique relies on the judgment of the researcher when choosing who to ask to participate.
- Researchers may implicitly thus choose a "representative" sample to suit their needs, or specifically approach individuals with certain characteristics.
- This approach is often used by the media when canvassing the public for opinions and in qualitative research.
- Judgment sampling has an advantage of being time-and cost-effective to perform whilst resulting in the range of responses (particularly useful in qualitative research).
- It is also prone to errors of judgment by the researcher and the findings, whilst being potentially board, will not necessarily be representative.

Bias in Sampling

- There are five important potential sources of bias that should be considered when selecting a sample, irrespective of the method used. Sampling bias may be introduced when.
- Any pre-agreed sampling rules are deviated from
- People in hard-to-reach groups are omitted
- Selected individuals are replaced with others, for example if they are difficult to contact There are low response rates
- An out-of-date list is used as the sample frame (for example, if it excludes people who have recently moved to an area)

Hypothesis

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

For example:

- A new medicine you think might work.
- A way of teaching you think might be better.

- A performance of work improved after training.

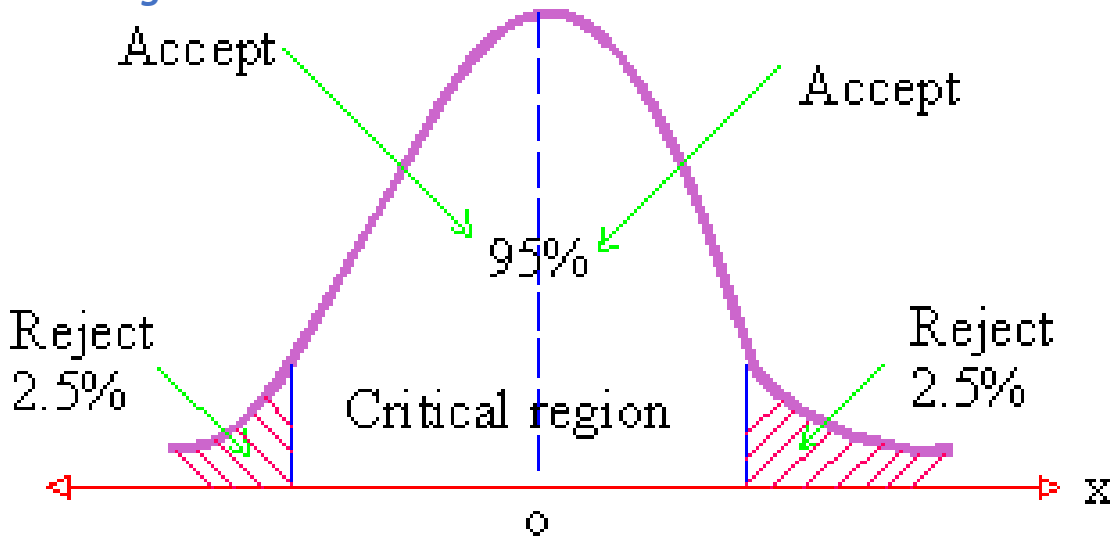
Null Hypothesis

- Null Hypothesis is also called statistical hypothesis because this type of hypothesis is used for statistical testing and statically interpretation.
- The null hypothesis predicts that, there is no relationship between the independent variable and dependent variable.
- Null hypothesis is denoted by; HO: gl = g2, which shows that there is no difference between the two population means.
- A statement in which no difference or effect is expected. If the null hypothesis is not rejected, no changes will be made

Alternative Hypothesis

- Alternate Hypothesis is also called non-statistical hypothesis because The alternate hypothesis is just an alternative to the null.
- The Alternate hypothesis predicts that, there is relationship between the independent variable and dependent variable.
- Alternate hypothesis is denoted by; Ha: $\mu 1 \neq \mu 2$, which shows that there is difference between the two population means.

Critical Region



- We need to Consider the following two facts. One is the significance level is the probability of rejecting a null hypothesis that is correct.

- The sampling distribution for a test statistic assumes that the null hypothesis is correct.

Statistical Test Interpretation

- The results of a statistical hypothesis test must be interpreted for us to start making claims.
- There is one common form that a result from a statistical hypothesis test may take, and they must be interpreted in different ways. The most common is the p-value.

P value

- A statistical hypothesis test may return a value called p or the p-value. This is a quantity that we can use to interpret or quantify the result of the test and either reject or fail to reject the null hypothesis. This is done by comparing the p-value to a threshold value chosen beforehand called the significance level.
- The significance level is often referred to by the Greek lower-case letter alpha.
- A common value used for alpha is 5% or 0.05. A smaller alpha value suggests a more robust interpretation of the null hypothesis, such as 1% or 0.1%.The p-value is compared to the pre-chosen alpha value. A result is statistically significant when the p-value is less than alpha. This signifies a change was detected: that the default hypothesis can be rejected.
- If p-value > alpha: Fail to reject the null hypothesis (i.e. not significant result).
- If p-value <= alpha: Reject the null hypothesis (i.e. significant result).
- For example, if we were performing a test of whether a data sample was normal and we calculated a p-value of .07, we could state something like:
- The test found that the data was normal, failing to reject the null hypothesis at a 95% confidence level.

Confidence Intervals

- In mathematics, a confidence interval corresponds to the likelihood that for a certain proportion of instances, a population parameter falls between a set of values. Confidence intervals in a sampling system

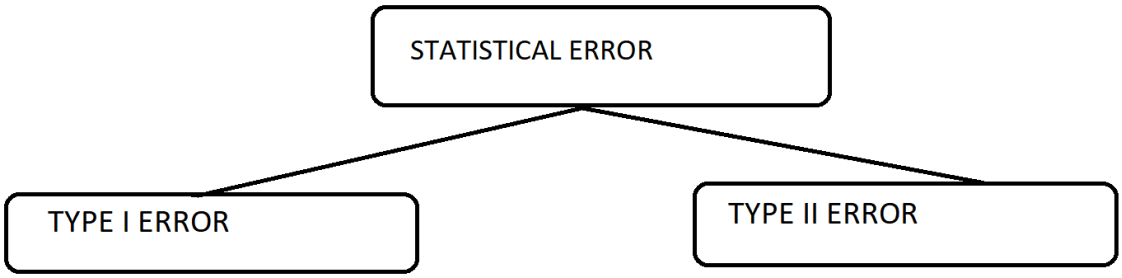
calculate the degree of ambiguity or certainty. They can take any number of probability limits, with a 95 percent to 99 percent confidence threshold being the most common one.

Error in Statistical Test

- The interpretation of a statistical hypothesis test is probabilistic.
- That means that the evidence of the test may suggest an outcome and be mistaken.
- For example, if alpha was 5%, it suggests that (at most) 1 time in 20 that the null hypothesis would be mistakenly rejected or failed to be rejected because of the statistical noise in the data sample.
- Given a small p-value (reject the null hypothesis) either means that the null hypothesis false (we got it right) or it is true and some rare and unlikely event has been observed (we made a mistake).
- If this type of error is made, it is called a false positive. We falsely believe the rejection of the null hypothesis.

Error in Statistical Test

- Alternatively, given a large p-value (fail to reject the null hypothesis), it may mean that the null hypothesis is true (we got it right) or that the null hypothesis is false and some unlikely event occurred (we made a mistake).
- If this type of error is made, it is called a false negative. We falsely believe the null hypothesis or assumption of the statistical test.



Type I and Type II Error

- Type I error: When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha.
- In hypothesis testing, the normal curve that shows the critical region is called the alpha region.
- Type II errors: When we accept the null hypothesis but it is false. Type II errors are denoted by beta.
- In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.
- Ideally, we want to choose a significance level that minimizes the likelihood of one of these errors.

WHAT IS T TEST

WHEN SHOULD WE PERFROM T TEST

Problem Statement I

- Considered a telecom company that has two service centers in the city. The company wants to find whether the average time required to service a customer is the same in both stores
- The company measures the average time taken by 50 random customers in each store. Store A takes 22 minutes while store B averages 25 minutes. Can we say that store A is more efficient than Store B in terms of customer service?
- It does seem that way, doesn't it? However, we have only looked at 50 random customers out of the many people who visit the stores. Simply looking at the average sample time might not be representative of all the customers who visit both the stores.

Assumption for Performing T test

- There are certain assumptions we need to heed before performing a t-test:
- The data should follow a continuous or ordinal scale (the IQ test scores of students, for example)
- The observations in the data should be randomly selected.

- The data should resemble a bell-shaped curve when we plot it, i.e., it should be normally distributed. You can refer to this article to get a better understanding of the normal distribution.
- Large sample size should be taken for the data to approach a normal distribution (although t-test is essential for small samples as their distributions are non-normal)

Types of T-Test

There are three types of T test.

- One sample T test
- Independent Two sample T test
- Paired sample t test

One Sample Test

- In a one-Sample t-Test, we compare the average (or mean) of one group against the set average (or mean).
- This set average can be any theoretical value (or it can be the population mean)
- A research scholar wants to determine if the average drinking time for a (standard size) coca cola differs from a set value. Let's say this value is 5 minutes. How do you think the research scholar can go about determining this?
-
- How we can perform a one-sample t-test. Here's the formula to calculate this:

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

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

Independent Two sample T Test

- The two-sample t-test is used to compare the means of two different samples.
- We want to compare the average drinking time for a (standard size) coca cola by male and female. Of course, the number of males and females should be equal for this comparison.
- Here's the formula to calculate the t-static for a two-sample t-test.

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

- Where,
  mA and mB are the means of two different samples
  nA and nB are the sample sizes
  S2 is an estimator of the common variance of the samples, such as:

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

  Here, the degree of freedom is nA + nB - 2

- We will follow the same logic we saw in a one-sample t-test to check if the average of one group is significantly different from another group. That's right – we will compare the calculated t-statistic with the t-critical value.

Business Problem

- Suppose, an HR Manager wants to find out whether male employees earn more than female employees.

Business Benefit

- Once the test is completed, p-value is generated which indicates whether there is statistical difference between income of two groups.
- Based on this value, a manager can easily conclude that whether average income earned by female employees is statistically different from male employees and if the different is statistically significant then which gender earns higher or lower.

Paired Sample t-test

- A certain manager realized that the productivity level of his employees was trending significantly downwards. This manager decided to conduct a training program for all his employees with the aim of increasing their productivity levels.
- How will the manager measure if the productivity levels increased? It's simple – just compare the productivity level of the employees before versus after the training program.

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

One-Sample Z-Test

- For example, you might be asked to test the hypothesis that the mean weight gain of pregnant women was more than 30 pounds. Your null hypothesis would be: H0: μ = 30 and your alternate hypothesis would be H,sub>1: μ > 30.

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- x̄ stands for the "sample mean"
  μ0 /population mean/hypothesized mean (in other words, the mean you are testing the hypothesis for)
  (σ) standard deviation
  (n) number of items in the sample

- Example: 500 women followed the Atkin's diet for a month. A random sample of 29 women gained an average of 6.7 pounds. Test the hypothesis that the average weight gain per woman for the month was over 5 pounds. The standard deviation for all women in the group was 7.1.

$Z = 6.7 - 5 / (7.1/\sqrt{29}) = 1.289.$

Two sample Z-tests

- This tests for a difference in proportions. A two-proportion z-test allows you to compare two proportions to see if they are the same.

- The null hypothesis (H0) for the test is that the proportions are the same.
- The alternate hypothesis (H1) is that the proportions are not the same.

Problem

- Vaccine A works on 351 people out of a sample of 605. Are the two drugs comparable? Use a 5% alpha level.

In order to solve this problem, we need to follow these steps

Step-1

- Find the two proportions
- P1 = 41/195 = 0.21 (that's 21%)
- P2 = 351/605 = 0.58 (that's 58%)

Step-2

- Find the overall sample proportion. The numerator will be the total number of "positive" results for the two samples and the denominator is the total number of people in the two samples.
- P = (41 + 351)/(195 + 605) = 0.49

Step – 3

- **Insert the numbers from step 1 and step 2 into the test statistic formula**

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad Z = \frac{(.58 - .21) - 0}{\sqrt{.49\ (1 - .49)\left(\frac{1}{195} + \frac{1}{605}\right)}}$$

- Solving the formula, we get: Z = 8.99. we need to find out if the z-score falls into the "rejection region".

Step - 4

CONFIDENCE LEVEL	ALPHA	ALPHA/2	Z ALPHA/2
90%	10%	5.0%	1.645
95%	5%	2.5%	1.96
98%	2%	1.0%	2.326
99%	1%	0.5%	2.576

- Find the z-score associated with α/2. I'll use the following table of known values:
- The z-score associated with a 5% alpha level / 2 is 1.96.

Chi Square

- The Chi Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.
- It is used to determine whether there is a statistically significant association between the categorical variables.
- Thus it finds out if the relationship exists between any two business parameters that are of categorical data type
- Examples:
- We could use chi-square test for independence to determine whether gender is related to a voting preference.
- We could determine if region has any influence on product category purchased.

Product Preferences	Cosmetics	Clothes	Sports	Perfumes
Gender				
Female	112	309	7	90
Male	162	267	0	53

- Here is a formula to calculate chi-square analysis

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- O – Observed frequency
- E – Expected frequency
- For finding out the expected frequency for each value of observed frequency:
- Expected frequency for Row1 and Column! = (Row1 total) *(Col1 Total)/Grand Total Calculations are as follows for given contingency table:

Business Problem

- Let's conduction the One-way Anova test on following two variables, one is a dimension and the other is a measure

Output

Dimension containing gender of a purchaser

Dimension containing product category purchased

Gender	Product Category
M	Footwear
F	Clothing
F	Clothing
F	Cosmetics
M	Accessories
M	Footwear

It is used to determine whether there is a statistically significant difference among more than two group means.

Example Anova-

- We could use One-way Anova test to determine if out of three or more rivers, at least two of them differ significantly from each other in terms of pH, TDS etc.
- We could determine if at least two regions differ significantly in terms of average sales of a particular product category

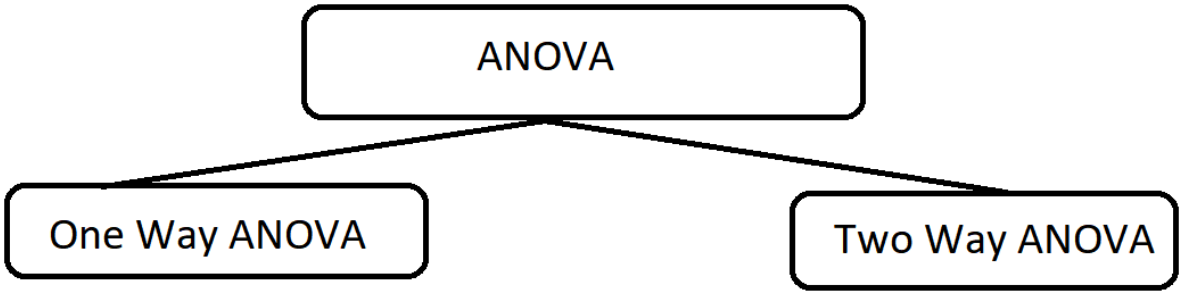
Business Problem

Let's conduct the one-way Anova test on following two variables, one is a dimension and the other is a measure.

Date	River	pH level
1/7/17	Narmada	5.6
1/7/17	Sabarmati	6.8
1/7/17	Yamuna	6.8
1/8/17	Sabarmati	6.7
1/8/17	Yamuna	6.8
1/8/17	Narmada	5.6

Types of ANOVA

- Analysis of Variance (ANOVA) is a statistical technique, commonly used to study differences between two or more groups



One Way ANOVA

- H0: The mean of one group is different
- This test is similar to the t-test, although ANOVA test is recommended in situation with more than 2 groups. Except that, the t-test and ANOVA provide similar results.

Two Way ANOVA

- A two-way ANOVA test adds another group variable to the formula. It is identical to the one-way ANOVA test, though the formula changes slightly.
- H0: the Means ar equal for both variables (i.e., factor variable)
- H3: the Means are different for both variables

What are Tails in Hypothesis Testing

A one-tailed test is a statistical hypothesis test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

When you perform a one-tailed test, the entire significance level percentage goes into the extreme end of one tail of the distribution.

You can choose either of the following sets of generic hypothesis:

Null: The effect is less than or equal to zero.

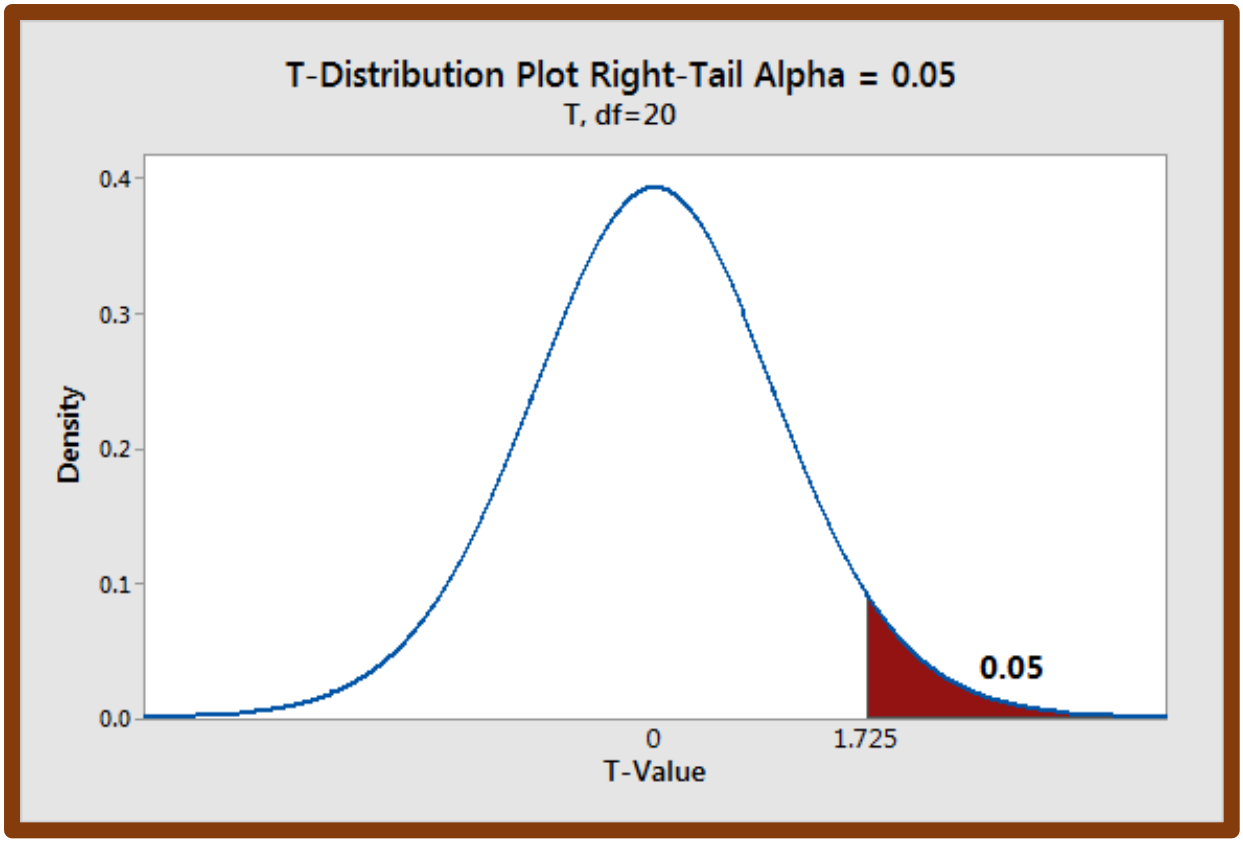
Null: The effect is greater than or equal to zero

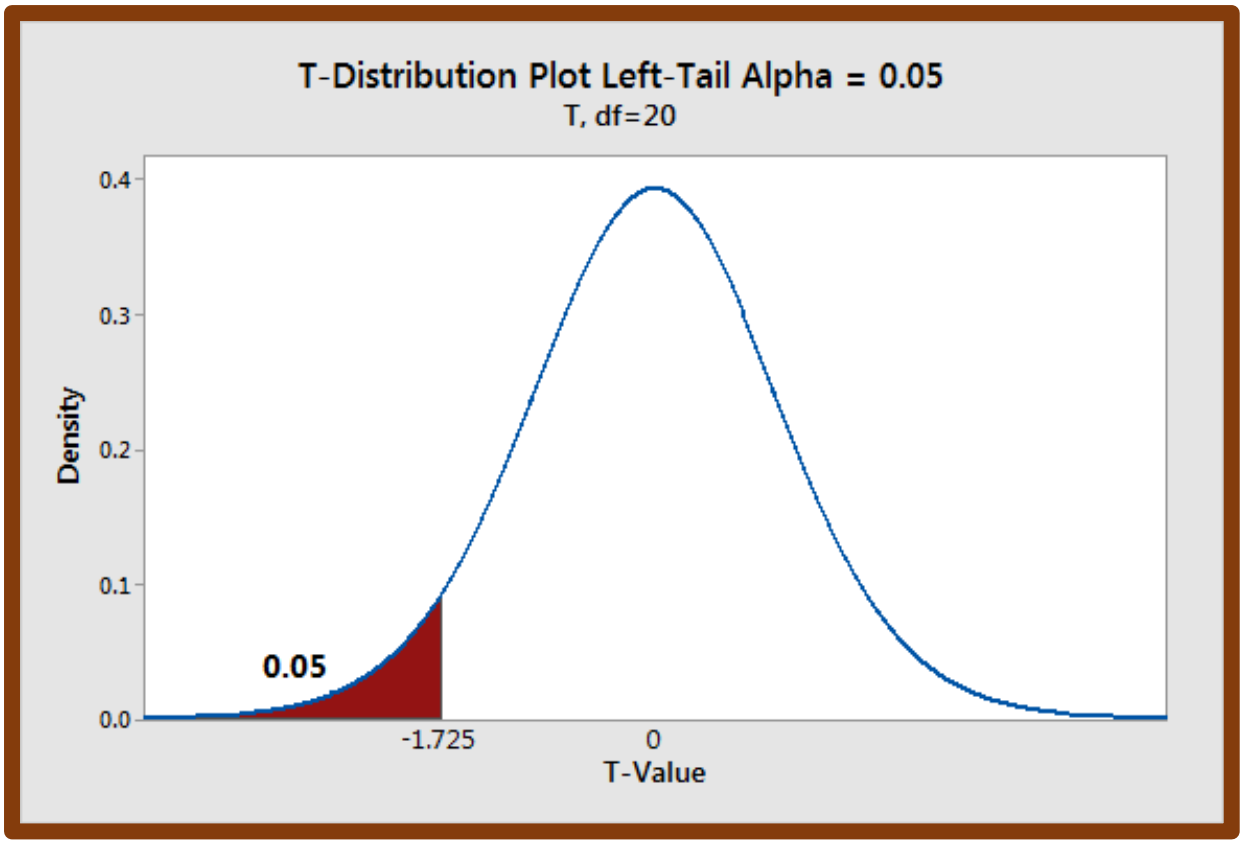
OR

Alternative: The effect is greater than zero.

Alternative: The effect is less than zero

ONE TAILS


T-Distribution Plot Right-Tail Alpha = 0.05
T, df=20
Density
0.4
0.3
0.2
0.1
0.0
0.05
0
1.725
T-Value


T-Distribution Plot Left-Tail Alpha = 0.05
T, df=20
Density
0.4
0.3
0.2
0.1
0.0
0.05
-1.725
0
T-Value

TWO TAILED TEST

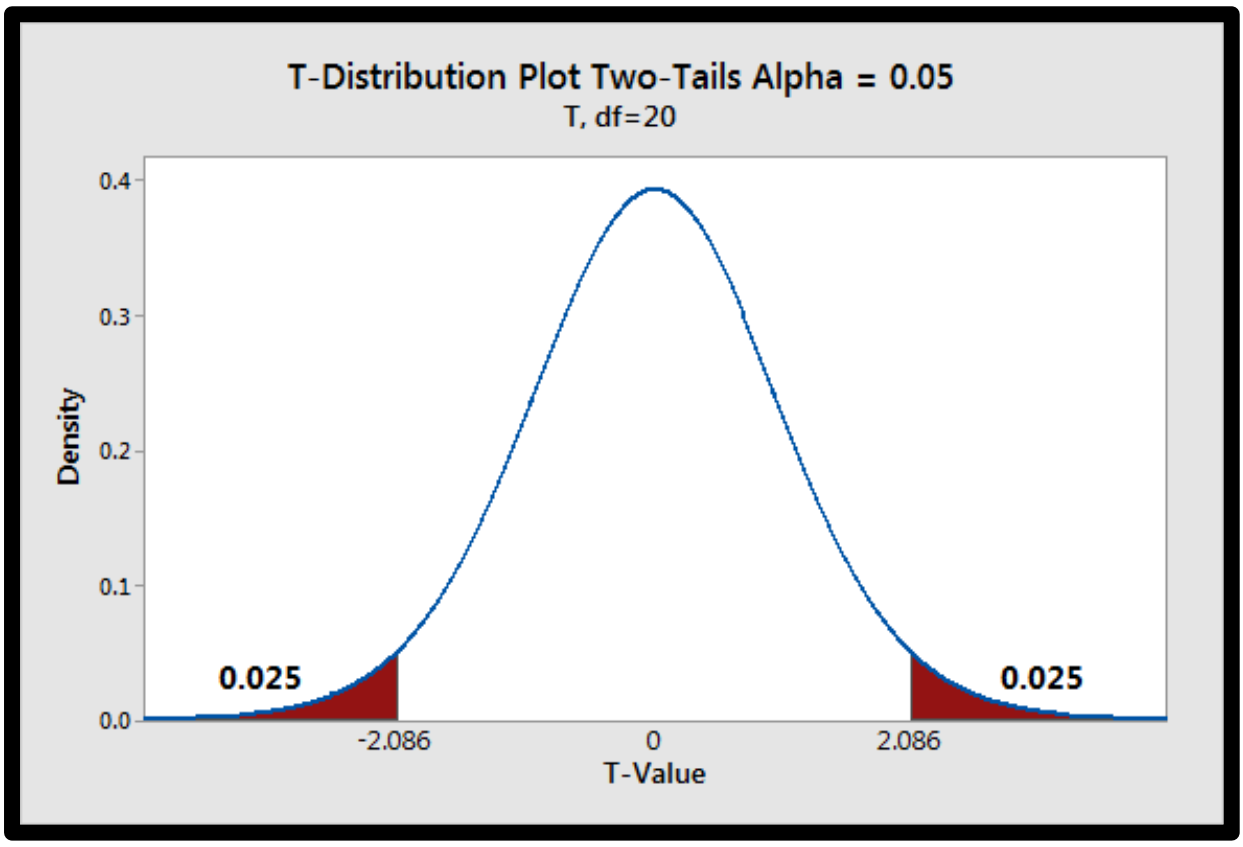
Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

When a test statistic falls in either critical region, your sample data are sufficiently incompatible with the null hypothesis that you can reject it for the population.

Null: The effect equals zero.

Alternative: The effect does not equal zero.

The specifics of the hypotheses depend on the type of test you perform because you might be assessing means, proportions, or rate.



Nonparametric and Parametric

- Nonparametric tests don't require that your data follow the normal distribution
- They're also known as distribution-free tests and can provide benefits in certain situations.
- people who perform statistical hypothesis tests are more comfortable with parametric tests than nonparametric tests.
    - ➢ Parametric analysis to test group means.
    - ➢ Nonparametric analysis to test group median

Reasons to use Parametric TEST

Reason-1

- Parametric tests can perform well with skewed and nonnormal distributions parametric analysis Test:
- 1-sample t test
- 2-sample t test
- One-way ANOVA

Reason -2

- Parametric tests can perform well when the spread of each group is different
- For nonparametric tests that compare groups, a common assumption is that the data for all groups must have the same spread (dispersion). If you groups have a different spread, the nonparametric tests might not provide valid results.

Reason-3

- Parametric tests usually have more statistical power than nonparametric tests. Thus, we are more likely to detect a significant effect when one truly exists.

Variance

- In statistics, variance is used to characterize the spread within a data set depending on its mean value. The probability-weighted average of squared deviations from the predicted value is determined by locating it.

Thus, the greater the deviation, the greater the difference between the numbers in the set and the mean. In comparison, a greater deviation indicates that the numbers in the set are similar to the mean.

Covariance

- Covariance is a measure of the relation between two random variables in mathematics and statistics. The metric measures how often the variables move together and what degree. It is basically a calculation of the difference between two variables, in other words. The metric does not however, measure the dependence between variables.

	Correlation	**Covariance**
Meaning	Correlation is an indicator of how strongly these 2 variables are related, provided other conditions are constant. The maximum value is +1, denoting a perfect dependent relationship.	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.
Relationship	Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance with standard deviation.	Correlation can be deduced from a covariance.
Values	Correlation is limited to values between the range -1 and +1.	The value of covariance lies in the range of -∞ and +∞.
Scalability	Correlation is not affected by a change in scales or multiplication by a constant.	Affects covariance
Units	Correlation is a unitless absolute number between -1 and +1, including decimal values.	Covariance has a definite unit as it is deduced by the multiplication of two numbers and their units

Univariate Analysis

The most basic type of statistical data analysis approach is Univariate Analysis. If the data contains only one variable and does not deal with relationships with causes or effects, then a Univariate analysis approach is used.

For example, in a classroom survey, the researcher might be searching for the number of boys and girls to be counted. The data will simply represent the number in this case i.e., a single variable and its quantity, as per the table below.

Variable = X	**Number = n**
Boys	87
Girls	80

The primary goal of univariate analysis to clearly define the data in order to identify patterns hidden in the data. This can be achieved by looking at the mean, median, mode, dispersion, variance, range, standard deviation, etc.

Bivariate Analysis

Bivariate analysis is relatively more analytical than univariate analysis. If two variables are included in the data set and researchers plan to compare the two data sets, then the correct method of analysis is bivariate analysis.

For instance, in a classroom survey, the investigator could examine the ratio of students who scored more than 85 percent corresponding to their genders. In the aforementioned scenario there are 2 different variables – gender = X (independent variable) and result = Y (dependent variable). A bivariate analysis, as shown in the table below will calculate the correlations between the two variables.

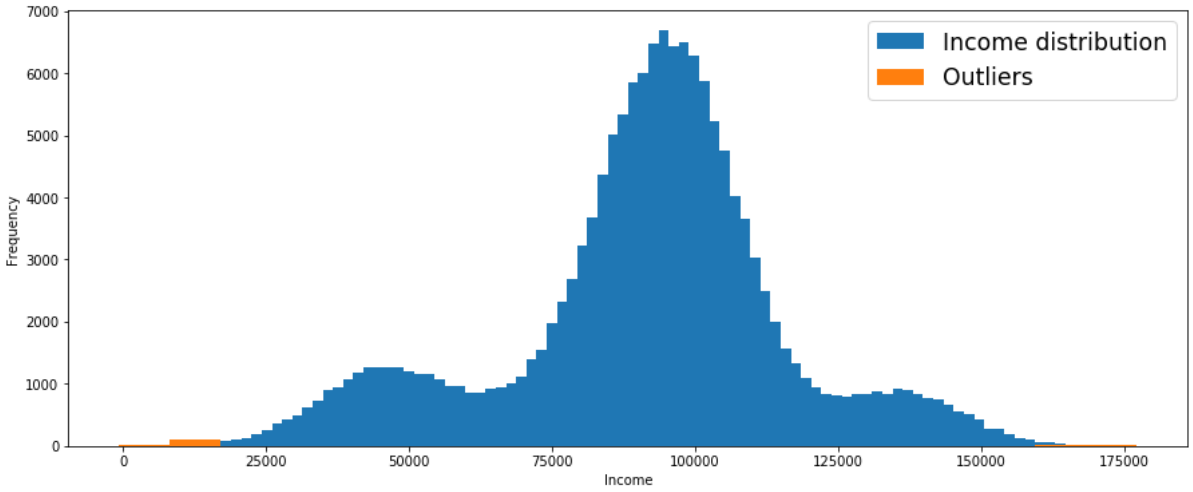
Gender = X (independent variable)	**Number = n**	**Ratio of students who scored more than 85% = Y (Dependent variable)**
Boys	55	6
Girls	50	11

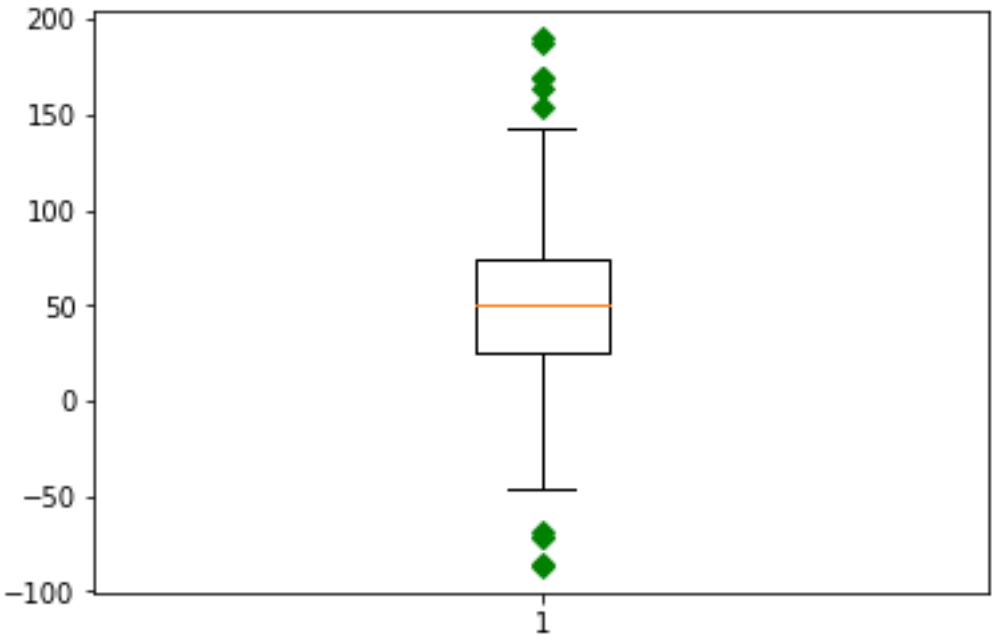
Correlation

- For measuring relationships between quantitative variables or categorical variables, correlation is used. It's an indicator of how things are connected. In other words, the study of how variables are correlated is called correlation analysis.

Outliers

- The concept of outliers is closely associated with robustness. Outliers are "abnormal" observations in the sample that seem very unlikely for the assumed distribution model or are remarkably different from the rest of sample observations. Outliers can be originated by measurement errors, exceptional circumstances, changes in the data generating process, etc.




200
150
100
50
0
−50
−100
1