

***Module-3***

Statistical Fundamental-II

# Inferential Statistics

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

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

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

# 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.

1. **Null Hypothesis**

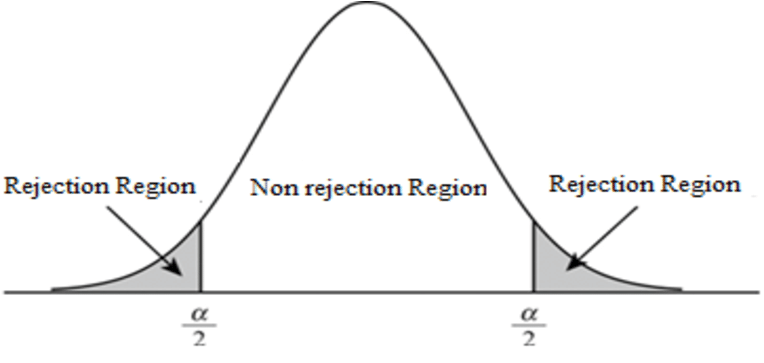
* Null Hypothesis is also called statistical hypothesis because this type of hypothesis is used for statistical testing and statically interpretation.
* The null hypothesis predicts that, there is no relationship between the independent variable and dependent variable.
* Null hypothesis is denoted by; HO: g1 = 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

1. **Alternative Hypothesis**

* Alternate Hypothesis is also called non-statistical hypothesis because this 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: μ1 ≠ μ2, which shows that there is difference between the two population means.

# Critical and Acceptance Region

* A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected.
* A confidence interval, also known as the acceptance region, is a set of values for the test statistic for which the null hypothesis is accepted.



# P value

* P-value is used to decide whether we should accept Null Hypothesis or reject it.
* If p-value is lower than the predetermined significance value (called alpha or threshold value) then we reject the null hypothesis.
* The alpha should be set before an experiment to avoid bias.
* For example, we generally consider population data to be in Normal Distribution, so we select alpha=0.05 (it means we are accepting if it lies in 95 percent of our distribution).
* This means that if our p-value is less than 0.05 we will reject the null hypothesis.

# Confidence Intervals

* A confidence interval (CI) is a range of values that is likely to contain the value of an unknown population parameter.
* 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 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.

# Parametric and Non-parametric

* A **parametric test** is considered when we have large dataset and mean is given. This 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.

|  |  |  |
| --- | --- | --- |
| **Properties** | **Parametric Test** | **Non-Parametric Test** |
| Assumptions | Yes, assumptions are made | No, assumptions are not made |
| central tendency | mean | median |
| Correlation | Pearson Correlation | Spearman Correlation |
| Distribution | Normal distribution | Arbitrary 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 | Chi-square test |

**Applications of Parametric Tests:**

* This test is used when data is quantitative and continuous.
* The data is of normal distribution.
* The data is estimated on the approximate ratio or interval scales of measurement.

**Applications of Non-Parametric Tests:**

* The data is estimated with different kinds of measurement scales.
* These are used when the distribution of the population is unknown.

# Statistical Tests

1. **T-TEST**

The T-test is an inferential statistic that is used to determine the difference or to compare the means of two groups of samples which may be related to certain features. It is performed on continuous variables.

**Examples:**

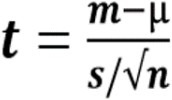
* A telecom company has two service centers in the city. The company wants to find whether the average time required to service a customer is the same in both stores.
* The company measures the average time taken by 50 random customers in each store. Store ‘A’ takes 22 minutes while store ‘B’ averages 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**

1. **One sample t-test:** One sample t-test which tells whether means of sample and population are different.

Where,

t = t-statistic

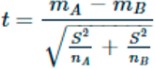
m = mean of the group

μ= theoretical value or population mean

s = standard deviation of the group

n = group size or sample size

1. **Two sample t-test:** It is known as Independent t-test. It compares the means of two independent groups and determines whether there is statistical evidence that the associated population means are significantly different.



Where,

mA and mB are the means of two different samples

nA and nB are the sample sizes

1. **Paired t-test:** When we want to compare means of different samples from same group or means from the same group at different times.
2. **Chi Square**

Chi-Square test is used when we perform hypothesis testing on two categorical variables from a single population or we can say that to compare categorical variables from a single population. By this we find is there any significant association between the two categorical variables.

**Examples:**

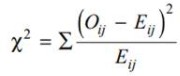
* A shop owner claims that an equal number of customers come into his shop each weekday.
* A policy maker wants to know whether or not gender is associated with political party preference.
* When we want to know whether or not marital status is associated with education level.

**Assumptions for performing Chi-square test:**

There are certain assumptions before performing a Chi-square test:

* Both variables are categorical.
* All observations are independent.
* Cells in the contingency table are mutually exclusive.
* Expected value of cells should be 5 or greater in at least 80% of cells.

**Formula to calculate chi-square analysis**



Where,

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

1. **ANOVA**

ANOVA stands for Analysis of Variance and is used to compare multiple (three or more) samples with a single test. It is used when the categorical feature has more than two categories.

The hypothesis is:

* Null: All pairs of samples are same i.e. all sample means are equal
* Alternate: At least one pair of samples is significantly different

**Examples:**

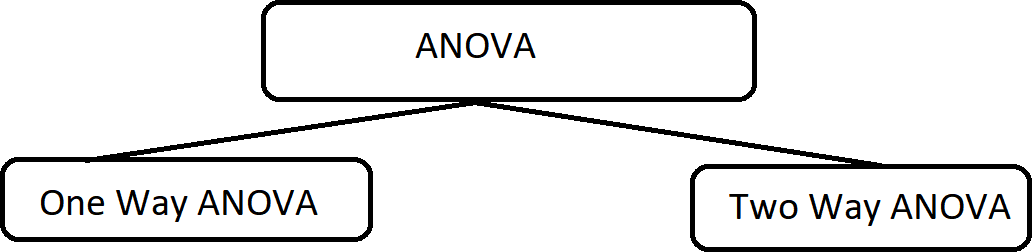
* There is a group of patients who are suffering from fever and are given three different medicines to cure fever. To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.
* A teacher wants to know how good he is in teaching with the students. So, he can split the students of the class into different groups and assign different projects related to the topics taught to them.

**Assumptions for performing ANOVA test:**

There are certain assumptions we need to heed before performing ANOVA test:

* Each group sample is drawn from a normally distributed population
* All samples have a common variance
* All samples are drawn independently of each other
* Within each sample, the observations are sampled randomly and independently of each other

**Types of ANOVA:**

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

1. **One Way ANOVA**

* A one-way ANOVA has one independent variable.
* Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

1. **Two Way ANOVA**

* A two-way ANOVA has two independent variables.
* Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master’s), and race finishing times in a marathon.

**Note:** All ANOVAs are designed to test differences among three or more groups. If we want to test for a difference between two groups, we can use t-test instead.

# Tails in Hypothesis Testing

1. **ONE TAILED TEST**

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:**

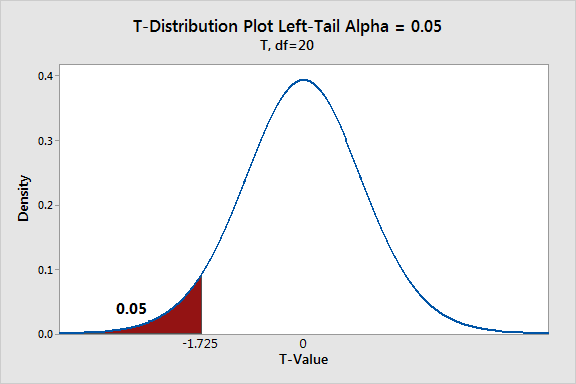
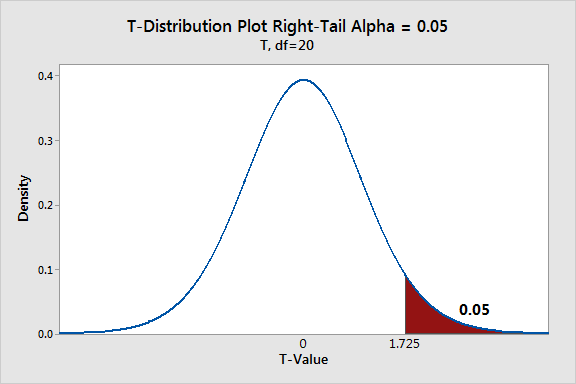
Null: The effect is less than or equal to zero.

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



1. **TWO TAILED TEST**

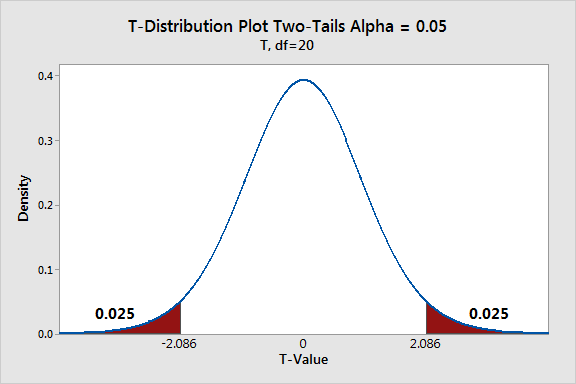
Two-tailed hypothesis tests are also known as non-directional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

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.

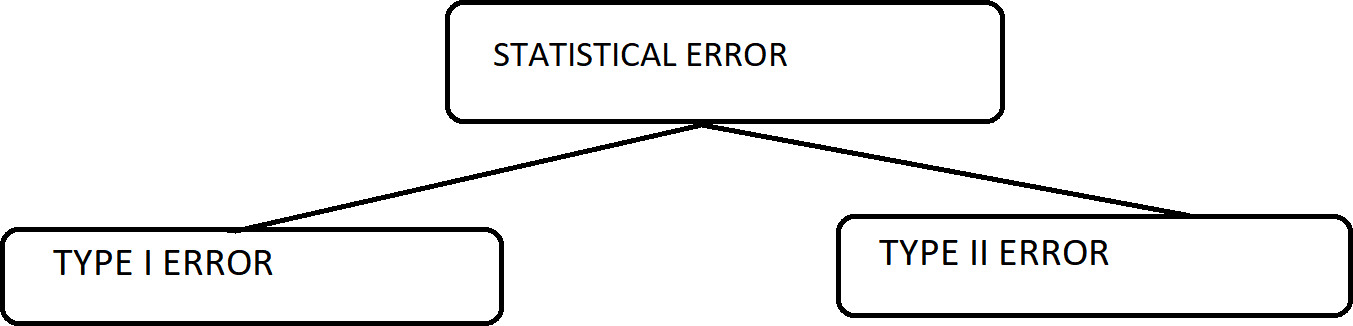
Alternative: The effect does not equal zero.

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



# Errors 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.
* 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.



1. **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.
* It is also known as false positive means test shows a positive result, but it should have been a negative (-) result.
* **Example:** the test result says you have corona-virus, but you actually don’t.

1. **Type II error:**

* 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 critical region is called the alpha region.
* It is also known as false negative means test shows a negative result, but it should have been a positive (+) result.
* **Example:** the test result says you don’t have corona-virus, but you actually do.

# Degrees of Freedom (DOF) in Statistics:

The degree of freedom represents the number of variables that have the freedom to vary in a calculation.

**Example:**

On Sunday, consider choosing 1 of the 7 cars. Drive any one of the 7 cars. On the second day, you should choose from the remaining cars. The pattern continues as follows:

Sunday: 7 cars to choose from

Monday: 6 cars to choose from

Tuesday: 5 cars to choose from

Wednesday: 4 cars to choose from

Thursday: 3 cars to choose from

Friday: 2 cars to choose from

Saturday: 1 car to choose from (no freedom to choose car on last day)

**Degrees of Freedom Formula:**

df=N−1

**Where**, N = number of values in the data set (sample size).