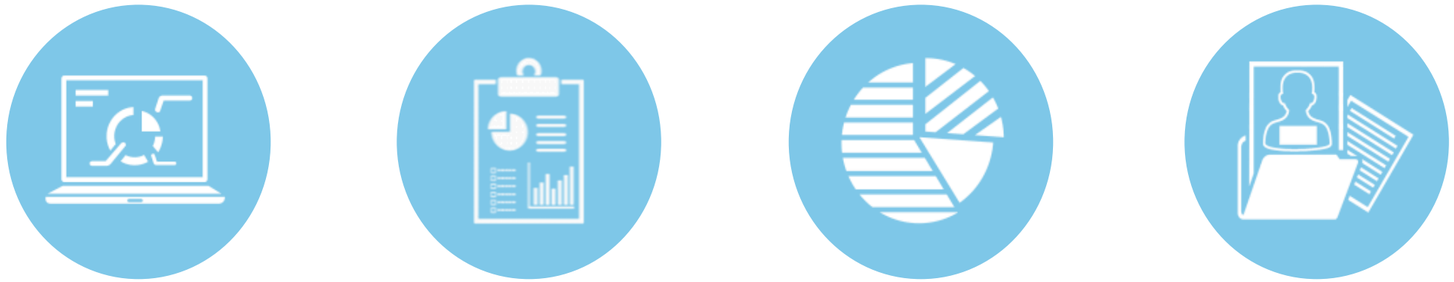
Data Science with R

Lesson 5— Statistics for Data Science – I



Learning Objectives

- Define hypothesis
- Explain data sampling
- Discuss confidence and significance levels

Statistics for Data Science – I

Topic 1—Hypothesis

Introduction

Consider a scenario where a marketing manager must decide whether to launch a new product or not.

On analysis, the manager could arrive at the following decision:

The product will be launched if the company gets a market share of 15% or more.

The product will not be launched if the company gets a market share of less than 15%.

Businesses analyze data to make optimal decisions that maximize profit at minimum risk.
Prediction of such outcomes depends on the acceptance or rejection of a **hypothesis.**

What Is a Hypothesis?

Hypothesis literally means Assumption. Assumption is a subjective term. A hypothesis is an assertion or a statement about the state of nature and the true value of an unknown population parameter.

Hypothesis: Example

Consider the following statements:

- Eating more vegetables leads to weight loss
- Brushing teeth everyday reduces cavities.

These statements have no supporting data and are hence considered hypotheses. A hypothesis needs analysis to be proved.

In statistics, most hypotheses are written as "if...then" statements. For example, If I eat more vegetables, then I will lose weight faster.

Types of Hypothesis

1. Simple Hypothesis
2. Complex Hypothesis
3. Null Hypothesis
4. Alternative Hypothesis
5. Statistical Hypothesis

Types of Hypothesis

- Simple Hypothesis
- Complex Hypothesis
- Null Hypothesis
- Alternate Hypothesis
- Statistical Hypothesis

In a simple hypothesis, there exists a relationship between two variables; one is called an independent variable or cause and the other is called a dependent variable or effect.

Example:

Given total Population = 100
Total No. of Male = 50
Total No. of Female = 50
Ho: μ=50

Types of Hypothesis

- Simple Hypothesis
- Complex Hypothesis
- Null Hypothesis
- Alternate Hypothesis
- Statistical Hypothesis

A complex hypothesis refers to the prediction of relationship between two or more independent variables or two or more dependent variables.

Example:

Total Population(μ_1) = 100
No. of Male (μ_2) = 50
No. of Female(μ) is H_O: $\mu = \mu_1 - \mu_2 = 50$

Types of Hypothesis

- Simple Hypothesis
- Complex Hypothesis
- **Null Hypothesis**
- Alternate Hypothesis
- Statistical Hypothesis

A Null Hypothesis is usually a hypothesis of “no difference.” It is denoted as H_0.

Null Hypothesis is performed for a possible rejection under a true assumption and always refers to a specified value of the population parameter, such as μ.

Example:

The population mean is 100
Or
H_0:μ = 100

Types of Hypothesis

- Simple Hypothesis
- Complex Hypothesis
- Null Hypothesis
- **Alternate Hypothesis**
- Statistical Hypothesis

An alternate hypothesis is complementary to the null hypothesis. It is denoted by H_1.

Alternate hypothesis is used to decide whether to employ a one-tailed test or two-tailed test.

Example:

For H_0:$\mu = 100$, the alternative hypothesis could be:

- H_1:$\mu \neq 100$
- H_1:$\mu > 100$
- H_1:$\mu < 100$

You will learn about one-tailed and two-tailed tests in the upcoming topics.

Types of Hypothesis

- Simple Hypothesis
- Complex Hypothesis
- Null Hypothesis
- Alternate Hypothesis
- Statistical Hypothesis

A statistical hypothesis is a method of statistical inference performed using data from a scientific study.

Example:

Given, total no of cities = 10
Mean population(μ) = 75
$H_0 : \mu = 75$

Statistics for Data Science – I

Topic 2—Data Sampling

What Is Data Sampling?

Data sampling is a statistical hypothesis technique used to select, manipulate, and analyze a subset of data points to discover hidden patterns and trends in the larger data set.

The sampling theory draws valid inferences about the population parameters on the basis of sample results.

Chances of Errors in Sampling

Consider the following scenarios:

- A quality inspector accepts or rejects hardware components supplied by a vendor, generally on the basis of test results of a random sample.

- A bank accepts or rejects a loan on the basis of a random sample of test results of loan payback with Interest and tenure.

In such cases, statistical decisions are taken on the basis of evidence and provide complete confidence to reduce the chances of error.

Types of Errors

The errors in statistical decisions are of two types:

Type I Error

- Reject H_0 when it is true
- Probability is denoted by α

Type II Error

- Accept H_0 when it is wrong or H_1 is true
- Probability is denoted by β

In practice, a Type I error means rejecting a lot when it is good (producer's risk) and Type II error means accepting a lot when it is bad (consumer's risk).

Statistics for Data Science – I

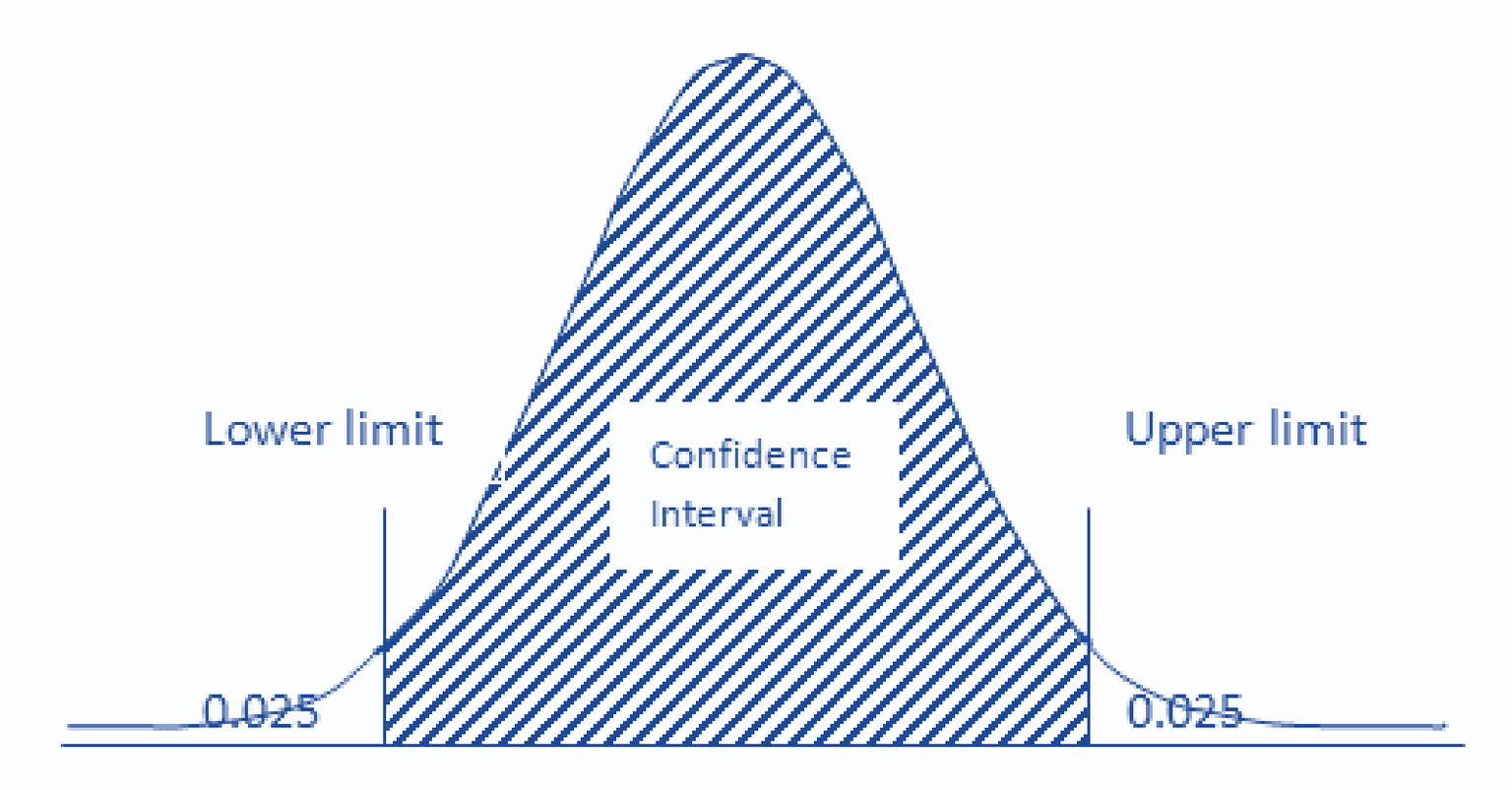
Topic 3—Confidence and Significance Levels

Confidence Levels

The confidence level is the frequency of possible confidence intervals that contain the true value of their corresponding parameter.

A confidence interval is a type of interval estimate that is computed from the observed data.

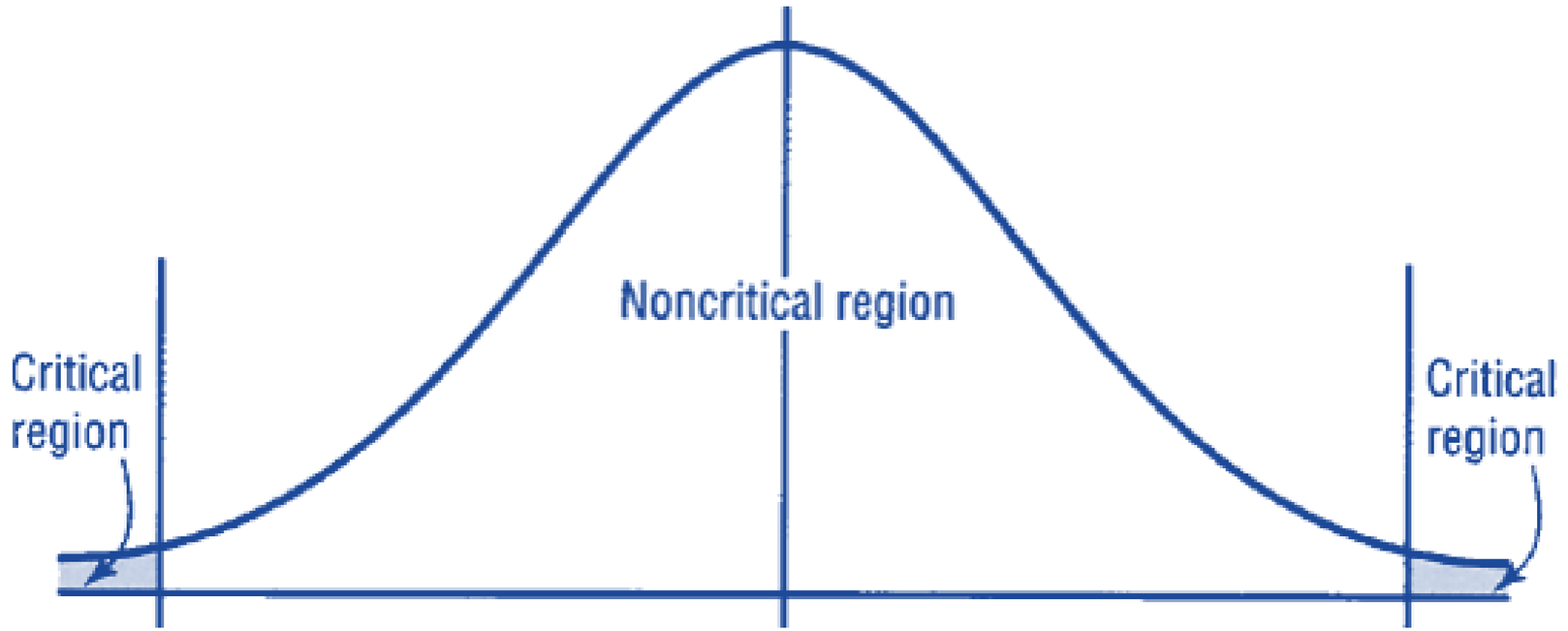
Confidence interval for a normal distribution is evaluated for continuous variables like sales in dollars, income of customers, age of customers, score in mathematics, etc.



Critical Region

The sampling distribution of a test statistic has two regions—a region of rejection (critical region) and a region of acceptance.

The critical region amounts to rejection of H_0, corresponding to the test statistic **t** in the sample space **S**.



Decision Making

The critical region helps in decision making by defining the region of acceptance and region of rejection. A decision can be correct or incorrect.

Correct Decision:

If sample test statistic falls in the rejection region, reject H_0.

Incorrect Decision:

If sample test statistic falls in the rejection region, accept H_0.

In the decision making approach to hypothesis testing, it is crucial to decide the level of significance prior to the collection of the sample data.

Level of Significance

Level of significance refers to the probability of a Type I error (α), that is, a random value of statistic t belonging to the critical region. It is usually set at 5% or 1% when employed in hypothesis testing.

- If α = 0.05 and you reject H_0, then there is a 5% probability that you have rejected H_0 when it is true.
- The desired level of significance depends on the amount of risk you want to take in rejecting H_0 when it is true.

Confidence Coefficient

Confidence coefficient is the complement of the probability of a Type I error (1-α) that yields confidence level when multiplied by 100%.

It represents the probability of concluding that a specific value of parameter being tested under H_0 is possible when, in fact, it is true.

It is a measure of accuracy and repeatability of a statistical test.

Critical Region Deviation

If w = critical region and t = t (X_1, X_2,.....,X_n) (Based on a random sample of size n)

Then,

- P (t ϵ w/H_0) = α, P (t ϵ w/H_1) = **β**
  (where, complementary set of w is the acceptance region)
- W U w = S, w ∩ w = φ

β Risk

β risk is the probability of committing a Type II error and depends on the difference between the hypothesized and actual values of the population parameter.

It is inversely proportional to α.

Beta risk depends on the magnitude of the difference between sample means and is managed by increasing test sample size.

Power of Test

The value **(1- β)** is known as the “power” of a statistical test.

It is the complement of the probability of a Type II error (1-β) and refers to the probability of rejecting H_0 when it is false.

Factors Affecting the Power of Test

Population Standard Deviation	Inversely proportional
Sample Size Used	Directly proportional
Level of Significance	Directly proportional

Key Takeaways

- Null Hypothesis is performed for a possible rejection under a true assumption and always refers to a specified value of the population parameter, such as μ.
- Data sampling is a statistical hypothesis technique used to select, manipulate, and analyze a subset of data points to discover hidden patterns and trends in the larger data set.
- The confidence level is the frequency of possible confidence intervals that contain the true value of their corresponding parameters.
- Level of significance refers to the probability of a Type I error (α), that is, a random value of statistic t belonging to the critical region. It is usually set at 5% or 1% when employed in hypothesis testing.
- Power of test is the complement of the probability of a Type II error (1-β) and refers to the probability of rejecting H_0, when it is false.