Inferential Statistics

Confidence Interval, Hypothesis Testing, ANOVA Optimization

Statistics

- Statistics has a significant part in the field of data science.
- It helps us in the
  - collection,
  - analysis and
  - representation of data

  either by visualization or by numbers.
- Generally, we divide statistics
  - Descriptive Statistics
  - Inferential Statistics.

Population and Sample

- **Population** contains all the data points from a set of data.
- While a **sample** consists of some observations selected from the population.
- The sample from the population should be selected such that it **has all the characteristics** that a population has.
- ***Population's*** measurable characteristics such as **mean, standard deviation** etc. are called as parameters while ***Sample's*** measurable characteristic is known as a **statistic**.

Descriptive Statistics

- Descriptive statistics describe the important characteristics of data by using mean, median, mode, variance etc. It summarises the data through numbers and graphs.

What is Inferential Statistics?

- An inference from a sample about the population.
- The main aim is to draw ***some conclusions*** from the sample and ***generalise*** them for the population data.
- E.g. we have to find the average salary of a data analyst across India.
  - The first option is to consider the data of data analysts across India and ask them their salaries and take an average.
  - The second option is to take a sample of data analysts from the major IT cities in India and take their average and consider that for across India.

Importance of Inferential Statistics

- ***Making conclusions*** from a sample about the population
- To conclude if a sample selected is ***statistically significant*** to the whole population or not
- ***Comparing two models*** to find which one is more statistically significant as compared to the other.
- In ***feature selection***, whether adding or removing a variable helps in improving the model or not.

Inferential Statistics

Probability, distributions, and hypothesis testing

Probability

- It is a measure of the chance of occurrence of a phenomenon.
  - **Random Experiment:** It is an experiment in which all the possible outcomes of the experiments are already known. The experiment can be repeated numerous times under identical or similar conditions.
  - **Sample space:** It is the collection or set of all the possible outcomes of a random experiment.
  - **Event:** A subset of sample space is called an event.

Probability (Cont..)

- **Trial:** It is a special type of experiment in which we have two types of possible outcomes: ***success or failure*** with varying Success probability.
- **Random Variable:** A value is subject to variations due to randomness is called a random variable. A random variable is of two types: ***Discrete and Continuous variable***. In a mathematical way, we can say that a real-valued function X: S -> R is called a random variable where S is probability space and R is a set of real numbers.

Conditional Probability

- Conditional probability is the probability of a particular event Y, given a certain condition which has already occurred , i.e., X. Then conditional probability, P(Y|X) is defined as,

$$\mathbf{P(Y|X) = P(X \cap Y) / P(X); P(X) > 0}$$

Probability Distribution and Distribution function

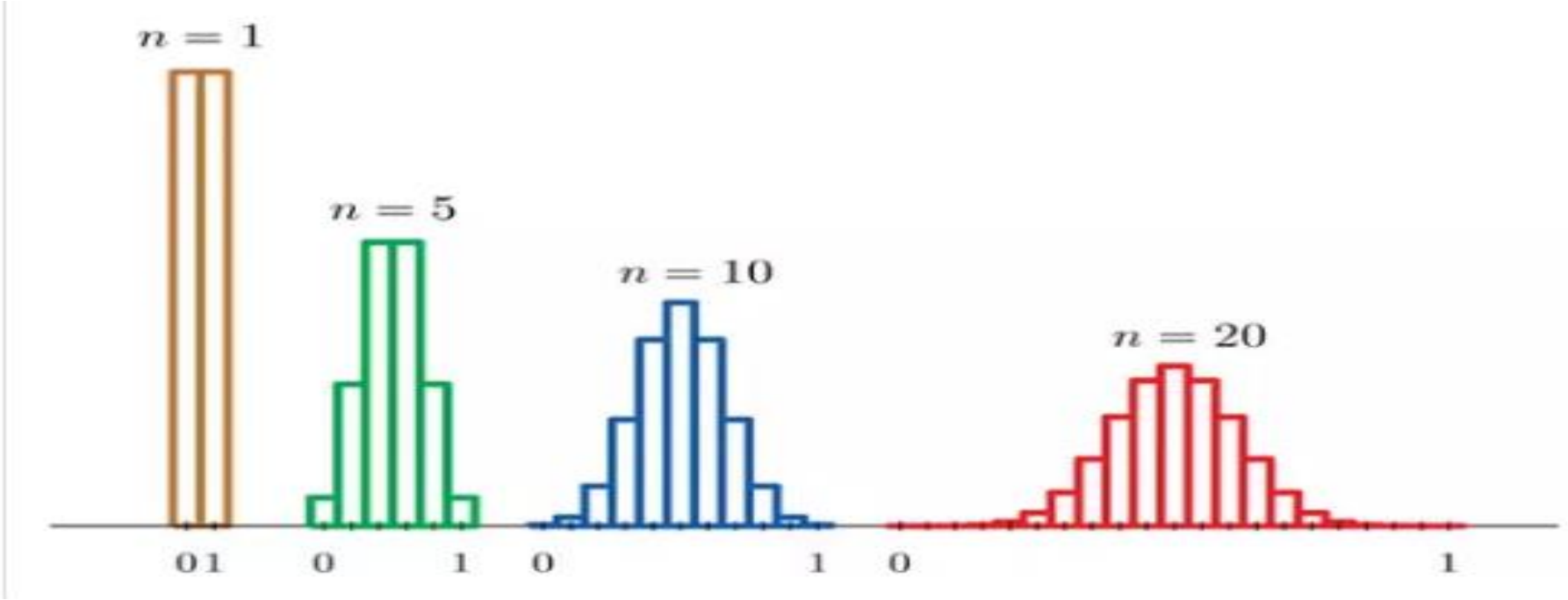
- The mathematical function describing the randomness of a random variable is called probability distribution.
- It is a depiction of all possible outcomes of a random variable and their associated probabilities

Sampling Distribution

- Probability distribution of statistics of a large number of samples selected from the population is called sampling distribution.
- When we increase the size of sample, sample mean becomes more normally distributed around population mean.
- The variability of the sample decreases as we increase sample size.

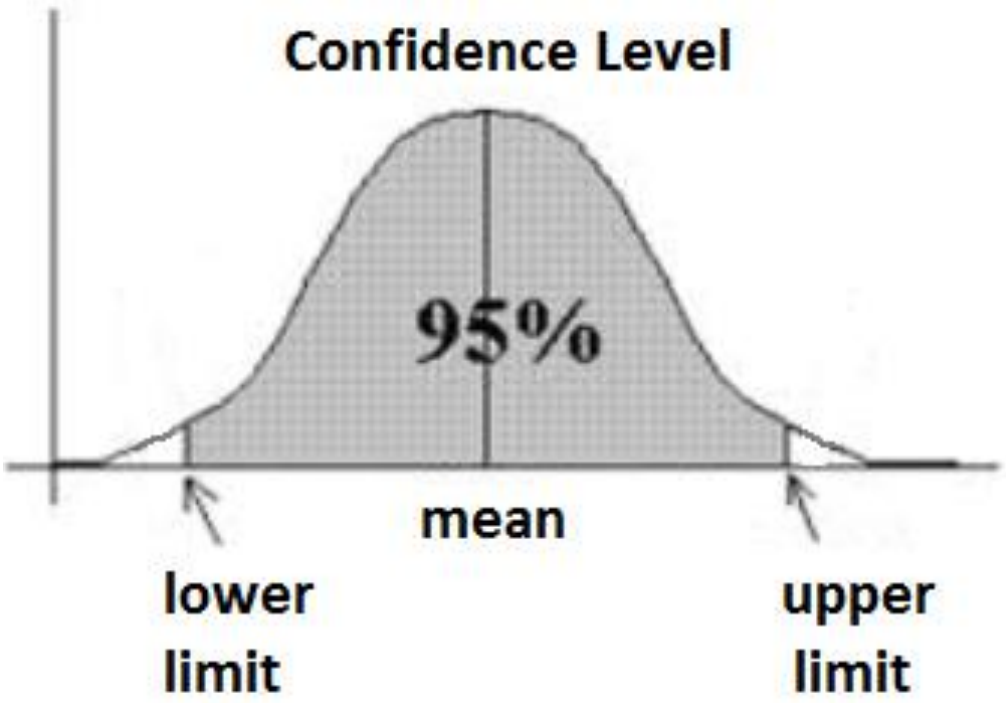
Central Limit Theorem

- CLT tells that when we increase the sample size, the distribution of sample means becomes normally distributed as the sample, whatever be the population distribution shape.



Confidence Interval

- Ex: Average weights of Cats in the world

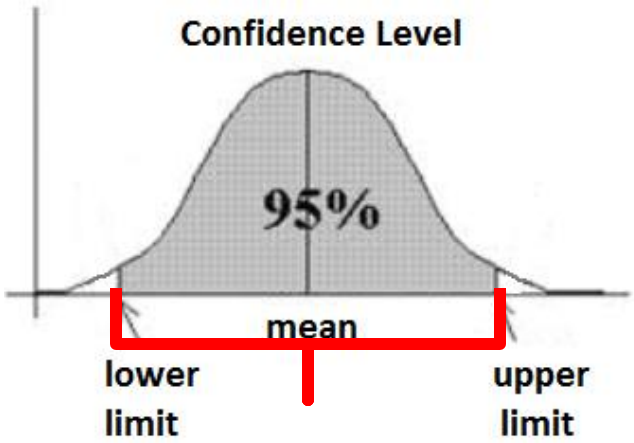


- Mean – true population mean
- Lower limit to upper limit – confidence interval

Calculate Confidence Interval for Population mean(μ)

- Example:

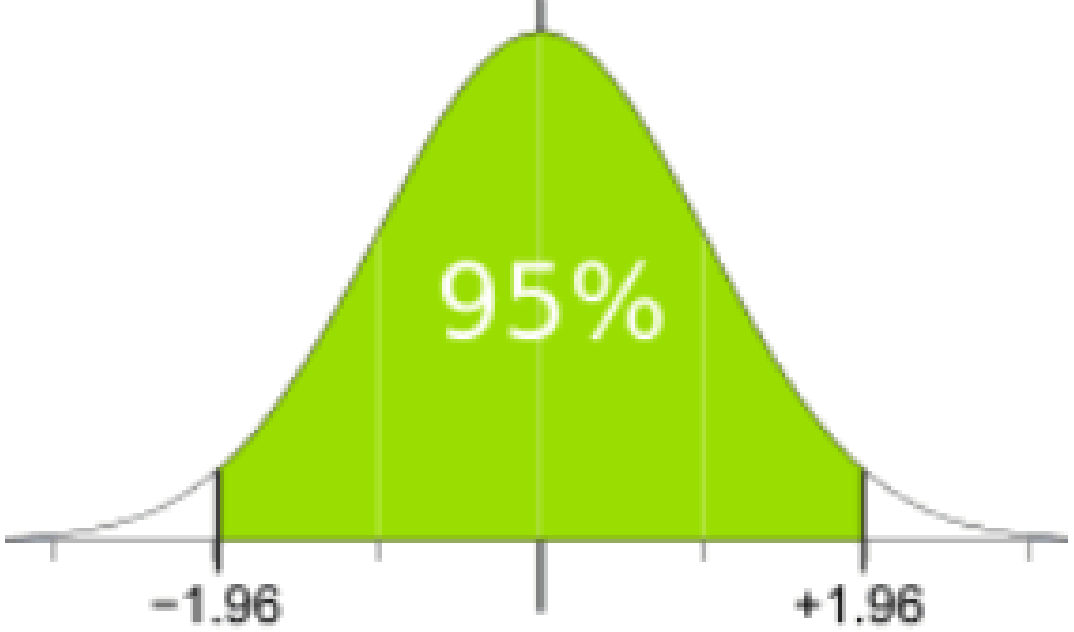
  From a normally distributed population, we took an Simple Random Sample(SRS) of 500 students with a mean score of 461 on the math section. Suppose the standard deviation of the population is 100, what is the estimated true population mean for the 95% confidence interval



- **Step 1- Organize the data**
  - Sample size n = 500
  - Sample mean X bar = 461
  - Confidence level C = 95%
  - Standard deviation σ = 100
- **Step 2 - Should satisfy below conditions**
  - Population should be normally distributed
  - Sample should be randomly selected
- **Step 3 - Calculate z value based on confidence level**

Confidence level	Critical (z) value to be used in confidence interval calculation
50%	0.67449
75%	1.15035
90%	1.64485
95%	1.95996
97%	2.17009
99%	2.57583
99.9%	3.29053

z(95)-value for 95% confidence level = 1.960



Confidence Interval Formula $\bar{x} \pm z * (\frac{\sigma}{\sqrt{n}})$

$461 \pm 1.96\ (100/\sqrt{500}\ )$
$461 \pm 8.765,$

here 8.765 is called **margin of error**

- To find the range of values you just have to add and subtract 8.765 from 461 and you will get (452.23, 469.77)

Hypothesis Testing

- Hypothesis testing is a part of statistics in which we make assumptions about the population parameter.
- Hypothesis testing mentions a proper procedure by analyzing a random sample of the population to ***accept or reject the assumption.***

Types of Hypothesis Testing

- **Null hypothesis:** Null hypothesis is a type of hypothesis in which we assume that the sample observations are purely by chance. It is denoted by **H0**.
- **Alternate hypothesis**: The alternate hypothesis is a hypothesis in which we assume that sample observations are not by chance. They are affected by some non-random situation. An alternate hypothesis is denoted by **H1 or Ha**.

Steps of Hypothesis Testing

- To determine whether to reject a null hypothesis or to fail to reject the null hypothesis, based on sample data.
  1. Define the null and alternate hypothesis
  2. Define an ***analysis plan*** to find how to use sample data to estimate the null hypothesis
  3. Do some analysis on the sample data to create a ***single number*** called **'test statistic'**
  4. Understand the result by ***applying the decision rule*** to check whether the Null hypothesis is true or not

If the value of t-stat is less than the ***significance level*** we will reject the null hypothesis, otherwise, we will fail to reject the null hypothesis.

Terms in Hypothesis testing

- **Significance level**

The significance level is defined as the ***probability*** of the case when ***we reject the null hypothesis*** but in actual it is true.

E.g., a 0.05 significance level indicates that there is 5% risk in assuming that there is some difference when in actual there is no difference. It is denoted by alpha (α).

- **P-value**

The p-value is defined as the probability of seeing a t-statistic as extreme as the calculated value if the null hypothesis value is true. *Low enough p-value is ground for rejecting the null hypothesis.*

We reject the null hypothesis if the p-value is less than the significance level.

Errors in hypothesis testing

- **Type-1 error**: When we reject the null hypothesis but in actual it was true. The probability of having a Type-1 error is called significance level alpha (α).
- **Type-2 error**: When we fail to reject the null hypothesis but actually it is false. The probability of having a type-2 error is called beta(β).

- **α = P (Null hypothesis rejected | Null hypothesis is true)**
- **β= P (Null hypothesis accepted | Null hypothesis is false)**

Z-test

A Z-test is mainly used when the data is normally distributed. We find the Z-statistic of the sample means and calculate the z-score.

$$\textbf{Z-score} = (x - \mu) / \sigma$$

T-test

The t-test is similar to z-test. The only difference is that it is used when we have sample standard deviation but don’t have population standard, or have a small sample size ($n<30$).

Different types of T-test

- **One Sample T-test**
- **Two sample T-test**
  - Independent Sample T-test
  - Paired T-test

Chi-square test

- Chi-square test is used in the case when we have to compare categorical data.
- Chi-square test is of two types. Both use chi-square statistics and distribution for different purposes.
  - ***Goodness of fit***: It determines if sample data of categorical variables matches with population or not.
  - ***Test of Independence***: It compares two categorical variables to find whether they are related with each other or not.
- Chi-square statistic is given by:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

ANOVA (Analysis of variance)

- It is a way to find out if an experiment results are significant or not.
- It is generally used when there are more than 2 groups and we have to test the hypothesis that the mean of multiple populations and variances of multiple populations are equal.

E.g. Students from different colleges take the same exam. We want to see if one college outperforms others.

Types of ANOVA

- One-way ANOVA
- Two-way ANOVA
- The test statistic in Anova is given by:

F-stat = MSB / MSW

MSB = Mean Square between

MSW = Mean square within