**Statistics**

The two types of statistics are: - **Descriptive** and **inferential.**

**Descriptive statistics** involves methods of summarizing and describing data through measures such as central tendency (e.g., mean, median, mode) and variability (e.g., range, variance, standard deviation). Descriptive statistics provide a way to understand and communicate the characteristics of a dataset.

**Inferential statistics** involves making generalizations about a population based on data collected from a sample. It involves using statistical tests and procedures to test hypotheses, estimate parameters, and make predictions about the population. Inferential statistics provide a way to draw conclusions about a larger group based on information gathered from a smaller group.

In statistics, a **population** is the entire group of individuals, objects, or events that we are interested in studying, and from which we want to draw conclusions. It is often too large or too costly to collect data from the entire population, so we usually take a smaller subset of the population, called a sample.

A **sample** is a portion or subset of the population that is selected for study. The sample should be representative of the population, meaning that the characteristics of the sample should be similar to those of the population. The goal of selecting a sample is to obtain a smaller group of individuals that can be studied and analyzed to draw conclusions about the population as a whole.

Population is denoted by = **Population(N)**

Sample is denoted by = **Sample(n)**

Sampling techniques refer to the methods used to select a sample from a population.

**Simple random sampling**: This involves selecting individuals from the **population at random**, where each individual has an equal chance of being selected. This can be done using a random number generator or by assigning numbers to individuals and selecting them using a random number table.

**Stratified sampling**: This involves dividing the population into subgroups, or strata, based on some characteristic (e.g., age, gender, income), and then selecting a sample from each stratum in proportion to its size. **( Jiska jiska opinion alag hoga un log ka alag group banaoo and collect data from all types of groups proportionately )**

**Cluster sampling**: This involves dividing the population into clusters, such as geographic regions or schools, and then randomly selecting some clusters to include in the sample. All individuals within the selected clusters are then included in the sample.

**Systematic sampling**: This involves selecting individuals from the population at regular intervals, such as every 10th person, starting at a randomly selected individual. **( har dusra admi ko pucho )**

**Convenience sampling**: This involves selecting individuals who are readily available or easily accessible, such as volunteers or individuals who happen to be in a certain place at a certain time.

**( Jo bhe ass pass hai )**

1. **Numerical variables (Quantitative):** These variables take on numeric values, and can be further classified as:

* **Discrete variables**: These variables take on whole number values only, such as the number of children in a family. **(Can never be in decimal or float)**
* **Continuous variables**: These variables can take on any value within a range, such as height, weight, or temperature. **(Can be in decimal or float)**

1. **Categorical variables (Qualitative):** These variables take on values that are categories or labels, such as gender (male or female), color (red, blue, green), or type of music (rock, country, jazz).

* **Ordinal variables**: These variables take on values that can be ranked or ordered, but the differences between the values may not be equal. For example, a Likert scale rating of agreement (strongly agree, agree, neutral, disagree, strongly disagree).
* **Binary variables**: These variables are a type of categorical variable that take on only two possible values, such as yes or no, true or false, or 0 or 1.

1. Nominal scale: This scale classifies variables into categories without any numerical value or order. Examples include gender, race, or marital status. Nominal scale variables can be analyzed using frequency tables and mode.
2. Ordinal scale: This scale classifies variables into ordered categories, but the distance between each category is not equal. Examples include education level, income bracket, or rating scales. Ordinal scale variables can be analyzed using median and quartiles.
3. Interval scale: This scale measures variables with equal intervals between each value, but there is no absolute zero point. Examples include temperature measured in Celsius or Fahrenheit, or dates measured in years. Interval scale variables can be analyzed using mean, standard deviation, and correlation analysis.
4. Ratio scale: This scale measures variables with equal intervals between each value and an absolute zero point. Examples include height, weight, or income. Ratio scale variables can be analyzed using mean, standard deviation, correlation analysis, and regression analysis.

**Frequency distribution & cumulative Frequency distribution**

**Bar graph Vs Histogram**

**Bar graph is for Discrete variables.**

**Histogram is for Continuous variables.**

**Measure of central tendency Vs Measure of Dispersion**

**central tendency** – refers to the measures used to determine the center of the distribution of data.

**Measure of central tendency includes** – Mean, Median, Mode. All 3 are used to find center point of any distribution. [ Basically used to handle missing values ]

Mean – doesn’t works well with outliers. [ Takes Average ]

Median – works well with outliers. [ takes middle value after sorting ]

Mode – Works well with Categorical Variable [ Takes most frequently occurred element ]

( missing values will be replace with most frequent occurring elements ) - **Categorical Variable**

**Measure of Dispersion** – used to describe how **spread** out the values in a dataset are.

**Measure of Dispersion includes** – Range, **Variance**, **Standard deviation**, Interquartile range (IQR), Coefficient of variation (CV).

**Range:** The range is the difference between the highest and lowest values in a dataset. It is a simple measure of dispersion that is affected by outliers.

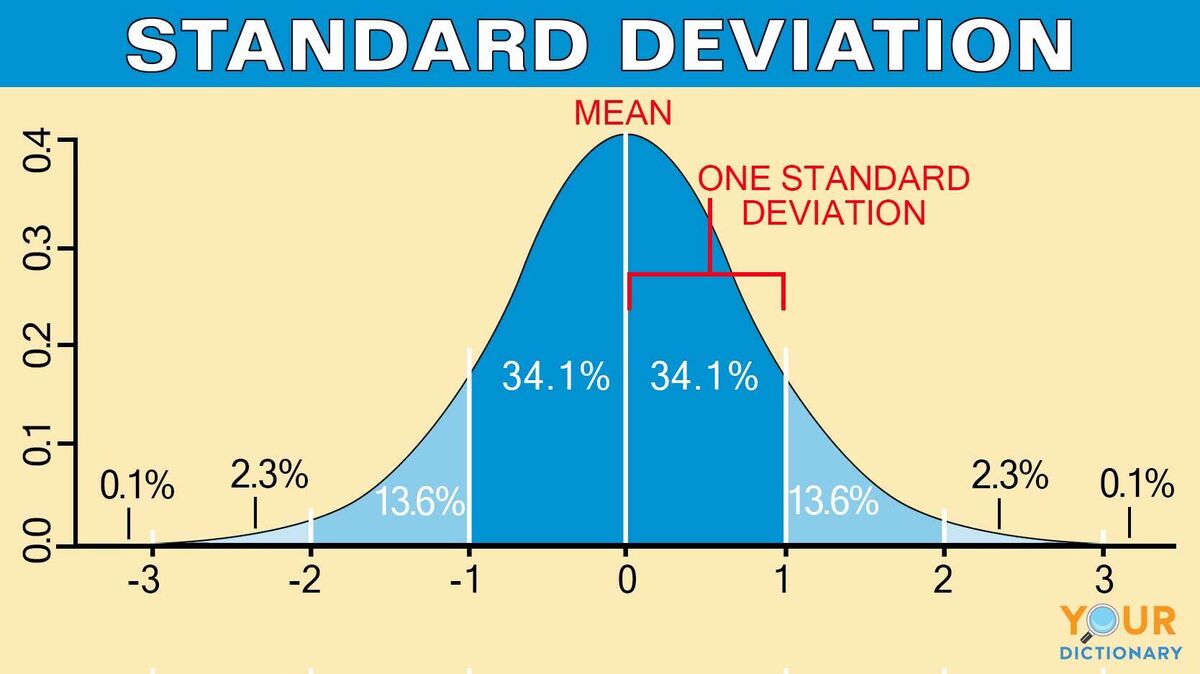
**Variance:** The variance is a measure of how far the values in a dataset are from the mean. It is calculated by finding the average of the squared differences between each value and the mean. A larger variance indicates a greater spread of values in the dataset.

**Standard deviation:** The standard deviation is the square root of the variance. It is a more commonly used measure of dispersion than variance because it is expressed in the same units as the data. A larger standard deviation indicates a greater spread of values in the dataset.

**Interquartile range (IQR):** The IQR is the range of the middle 50% of the values in a dataset, and is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). It is less sensitive to outliers than the range.

**[ Lower = Q1 – 1.5(IQR) Upper = Q3 + 1.5(IQR) ] - Anything beyond this range ia an Outlier.**

**Coefficient of variation (CV):** The CV is the ratio of the standard deviation to the mean, expressed as a percentage. It is used to compare the variability of datasets with different units or scales.



The Y-axis in the normal distribution represents the "**density of probability**." Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis.

The X-axis is “**standard deviation”.**

**Types of Distributions**

* Bernoulli Distribution.
* Uniform Distribution.
* Binomial Distribution.
* Normal Distribution.
* Poisson Distribution.
* Exponential Distribution.

**Standardization**

**Z-score** indicates how much a given value differs from the standard deviation.

**Any normal distribution can be standardized by converting its values into z scores.**

Standard Normal Distribution = Mean = 0 and SD(**σ**) = 1

**Normalization**

Normalization is **a scaling technique method in which data points are shifted and rescaled so that they end up in a range of 0 to 1**. It is also known as min-max scaling.

**Probability**

Types of Distributions: -

1. Binomial Distribution
2. Normal Distribution
3. Poisson Distribution
4. Exponential Distribution
5. Geometric Distribution
6. Lognormal Distribution

**Discrete Continuous**

Binomial Distribution Normal Distribution

Poisson Distribution Exponential Distribution

Geometric Distribution Lognormal Distribution

**PMF+CDF PDF+CDF**

from scipy.stats import norm

from scipy.stats import poisson

from scipy.stats import expon

from scipy.stats import binom

from scipy.stats import lognorm

from scipy.stats import geom

1. **Binomial Distribution**

binom.pmf (n = No. of Trials, k = k-success, p = Probability of success)

binom.cdf (n = No. of Trials, k = k-success, p = Probability of success)

binom.expect (args=( n = No. of Trials, p = Probability of success ))

2. **Normal Distribution**

Norm.cdf (pass SD to get the fraction of data that falls before that SD)

Norm.cdf(target no. , loc = mean , scale = SD )

Norm.ppf (pass the fraction to get the SD at that point in graph, mu, SD)

Norm.pdf (pass the fraction to get the SD at that point in graph, mu, SD)

1. **Poisson Distribution**

Poisson.pmf (k = no. of times, mu = Mean)

Poisson.cdf (k = no. of times, mu = Mean)

1. **Exponential Distribution**

expon.cdf (x = no. of times, scale = Mean)

expon.pdf (x = no. of times, scale = Mean)

1. **Geometric Distribution**

geom.pmf (k = no. of times, p = Probability)

geom.cdf (k = no. of times, p = Probability)

geom.expect (args = (probability))

6. **Lognormal Distribution**

lognorm.cdf (x = no. of times, s = SD, scale = Mean)

* Binomial Distribution – N success & n Trials are given.
* Normal Distribution – no. of fixed trials is not given; we need to find probability.
* Poisson Distribution – rate per something is given. Rate / hour, time, area…….
* Exponential Distribution – graph goes bottom to up exponentially.
* Geometric Distribution – opposite of exponential dist. graph it comes down from top.
* Lognormal Distribution

**Hypothesis testing**

**P value** – Probability of observing data as extreme as the observed test statistic(**T**) under the assumption that the null hypothesis is true.

**P value – p( [data | H0 is True] )**  P Value = [ probability of 7 Heads | if coin is Fair ]

If P Value is very low, we reject Null Hypothesis. **P Value < Alfa we reject Null H0**

**Hypothesis testing -** The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample.

Null hypothesis(**H0**) – default assumption >=, <=, =

Alternate hypothesis(**H1**)– opposite of Null hypothesis >, <, <>

**Confidence interval** - A confidence interval displays the probability that a parameter will fall between a pair of values around the mean. Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

**P Value** - You can use either P values or confidence intervals to determine whether your results are statistically significant. If a hypothesis test produces both, these results will agree. The confidence level is equivalent to 1 – the alpha level. So, **if your significance level is 0.05, the corresponding confidence level is 95%**.

**Alfa = 0.05 then CI = 95%**

**P Value < Alfa we reject Null H0**

**Type – 1 error & Type – 2 error**

Type – 1 error is known as **false positive**,(innocent ko chor bna diya) i.e., when we reject the correct null hypothesis, whereas type -2 error is also known as a **false negative(Chor bach gaya)**, i.e., when we fail to reject the false null hypothesis.

**Z-Test (Z-score) ( Test for 1 catagorical features & 1 Numerical feature)**

**Test statistic(z test) = (Mean µ** **0** - **Mean µ** **1)/ (SD/√ n)**

**P value** = 1-norm.cdf (**Test statistic(z test)**)

**(1 sample test) vs (2 sample test)**

(1 sample test) – compare with benchmark.

Test statistic (z test) = (**Mean µ 0 - Mean µ 1)/ (SD/√ n**)

From statsmodels.stats.weightstats import ztest

ztest(data, value = mean , alternative="two-sided" / “smaller” / “larger”)

(2 sample test) – compare with each other.

Test statistic (z test) = (**Mean µ 1 - Mean µ 2) / √ ( (SD1/n1) + (SD2/n2) )**

From statsmodels.stats.weightstats import ztest

ztest(data1 , data2 , alternative="two-sided" / “smaller” / “larger”)

**Z Test vs T Test**

If SD of population is not given and If sample size id less than 30 use t test if greater than 30 use z test

**T\_Test ( Test for 1 catagorical features & 1 Numerical feature)**

**Test statistic = (Mean µ** **0** - **Mean µ** **1)/ (SD0/√ n0 + SD1/√ n1)**

(1 sample test) – compare with benchmark

From scipy.stats import ttest\_1samp

ttest\_1samp( Data , popmean = mean , alternative= “less” / “greater”)

**AB Test**

(2 sample test) – compare with each other. **2 sample test are generally AB test .**

From scipy.stats import ttest\_ind

ttest\_ind( Data1 , Data2 , alternative= “less” / “greater”)

**Chi2 ( Test for 2 catagorical features)**

Crosstab wala test

From scipy.stats import chi2\_contingency

chi2\_contingency(pd.crosstab( index = df[“catagory1”] , columns = df[“catagory2”] )

**ANOVA (multiple t test)**

**Assumption1 = Normality all distribution should be normal**

**Assumption2 = Variance all should have same variance**

It is used for multiple testing at once.

**Test statistic** = F – ratio = variance between groups / variance within groups

From scipy.stats import f\_oneway

F\_oneway (data1 , data2 , data 3 , …..)

Corelation