Statistics :

Statistics is a branch of mathematics that deals with the collection , analysis , interpretation , presentation , and organization of data. In applying statistics to a scientific , industrial , or social problem , it is conventional to begin with a statistical population or a statistical model to be studied.

OR

The life cycle of data are statistics.

Pillars of data :

- **Research :** How to create / collect data and it is mostly comes from our observation.
- **Collection :** Sampling.
- **Analysis :** Data Analysis.
- **Interpretation :** Graphs / Plotting.
- **Presentation :** Present plot to the general audience.
- **Organisation :** Organize the data e.g. It is used by big companies like how to google organize the big data.

Topics / Content Of Statistics :

1. Statistics Introduction.

2. What is data (5 pillars are on the above).

3. Data Types.

4. Level / Scales of measurement.

5. Data Collection Methods (Sampling).

6. Data Visualization.

7. EDA (Exploratory Data Analysis) (M.Important).

8. Descriptive Statistics (Central tendency (Mean, Median, Mode) , Variaility (Range, Variance, Standard Deviation) , Data Distribution (Skewness / kuttosis , normal distribution)).

9. Probability Theory (M.Important For ML & AI).

10. Inferential Statistics (hypothesis testing , population vs sample , sampling distribution , confidence intervals , types of errors (type 1 , type 2) , common statistical

tests (Uni variate analysis , Bi variate analysis , Multi variate analysis) , P - Values and significance).

11. Machine learning (regression , time series , etc).

1. Collect Data / Sampling :

Statistics are also helpful in data collections. They tell us how to collect data.

- Sampling from a population: When you want to study a larger population, you need to sample a smaller group. This is called sampling.

For Example :

I want to know how many people are in `IT` field so , i survay in cities and villages and ask some people about their field where they currently are working. Suppose i ask `500` people about their field and i get data that `250` are male in `IT` field , `100` are female in `IT` field , and `150` are `Non-IT` people , and analyze this data and get output. But if we think is this output is `True` / `Right` e.g. if i get the output that `50 %` people are in `IT` field , and honestly the `answer` is `wrong` / `False` because i survay only in one city and collect that data , it means `data collection` is also very important in statistics. Before the data collection there are some important step before the data collection which are given below :

2. How To Measure Data / Scales or Levels Of Measurement In Statistics :

There are 4 scales of data measurement / Scales of measurement :

1. Nominal Scale : The data which consists of `alphabetical letters` (e.g. `nouns` , `lables` or `tags` , `eye color` , `qualities`) are called `Nominal Scale` . We cannot write `numbers` , `order (ranking)` e.g. we cannot say that `Male` > `Female` OR `Female` > `Male` etc.

- **Statistics Anlaysis Apply On Nominal Scale :** `Mode` , `Frequency` , `Percentage` , `Bar Chart` , `Pie Chart` , `Histogram` , OR `Chi - Square` etc.

2. Ordinal Scale : `Ordinal Scale` is also consists of `Nominal Scale` e.g. `Nominal Scale` techniques , `alphabetical letters` etc. But it is also contains of `Order (Ranking)` e.g. my question is `is it usually very hot in Pakistan ?` and our usually answers is `I agree.` , `I strongly agree.` , and `I don't agree.` so , it is called `Rank Order` OR `Ranking (Order)` and second of `Ordinal Scale` is if i ask a question to my friend that `What is your position at school ?` and his answer is `First` , `Second` , `Third` , OR `Forth` and so on ... if he got `First` position i tick the `First` coloum , or if he got `Second` position then i tick the `Second` Position and so on ... so , its means it is based on `alphabetical lettes` data but we can `Rank` this data.

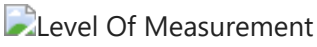
- **Statistics Analysis Apply On Ordinal Scale :** `Mode` , `Median` etc.

3. Interval Scale : It consists of `Numbers` , `Equal difference b/w numbers` , `Not True Zero` e.g. if i say `How much money i have ?` so we can't say that `I have Zero (0) money` it is called `Not True Zero` , another example is `dates` in a calender , and `Temperature` . It contains of `whole numbers` .

- **Statistics Apply On Interval Scale :** `Mean` , `Standard Deviation` , `Anova` , and `Regression` .

4. Ratio Scale : It consists of `Numbers` , `True Zero` Means `Zero` , `Arithmetic Operators` in meaningful ways e.g `Height` , `Weight` , `Income` , and `Distance` . It is also contains `Decimal Points Numbers` .

- **Statistics Apply On Ratio Scale :** `Mean` , `Median` , `Mode` , `Standard Deviation` , `Geometric Mean` , and `Harmonic Mean` etc.



Note : `Levels Of Measurements` helps us to understand which `statistical analysis` should we apply on the data.

Types Of Levels Of Measurement :

1. Qualitative : In this type we see the `Quanlity` of the data. It is also called `Categorical Data` .It has only `categorical data` . It's means no `numbers` involve in them. e.g. `Nominal Scale` and `Ordinal Scale` .

2. Quantitative : In this type we see the `Quantity` of the data. It is also called `Numerical Data` . It has only `numerical data` . It's means only `numbers` involve in them. e.g. `Interval Scale` and `Ratio Scale` .

Statistical Analysis Apply On Quantitative :

- We can apply wide varity of `Arithmetic Operators` e.g. + , - , * , / etc. But it depends on the data is on which scale e.g. is the data is on `Interval Scale` of `Ratio Scale`

Note : Most of the data in the world are `Quantitative` data.

Data Encoding :

Encoding is the process of converting categorical or textual data into a numerical format so that it can be efficiently used in statistical analysis or machine learning algorithms. This transformation enables algorithms to interpret and process non-numeric data by assigning unique numerical values or vectors to each category or label.

For Example :

We have the data of `Gender` and we encode this data in `0` and `1` form it is called encoding the data. `0` means `Female` and `1` means `Male`. It is also called data collection method.

Note :

The example on the above are `Qualitative Data`. But it is collected in `Numerical Form`. But basically it is `Qualitative Data`.

Benefits Of Encoding The Data :

1. It make easy to understand to our computer.
2. Other softwares can easily read `numerical` data.
3. It boosts the processing speed.

Types Of Quantitative Data :

3. Discreat Data :

The data that are `countable` with `integers` and have equal distance are called `discreat data`.

For Example :

1. Number of cars in parked area ? , suppose it is `50` then it is discreat data and have equal distance.
2. Number of children in your family ? , suppose it is `8` then it is also discreat data and have equal distance.

4. Continuous Data :

The data that are based on `Range` are called continuous data and it is `measureable` data e.g. Height , Weight , Time , and Temperature etc. The `decimal point` data are also called `continuous data`. In `python` programming language the `decimal point` are also called `float data`.

For Example :

`1 - 2` are continuous data , the data between this range i.e. `1` , `1.1` , `1.2` , `1.3` ... till `2` are range data.

5. Binary Data :

The data that are based on only `two values` are called `binary data`.

For Example :

The data that are based on `Boolen` means `Ture` and `False` , `0` and `1` , and `Yes` or `No` etc are `binary data` .

Note :

`Binary` data are also called `Qualitative` data but , it is measured in `0` and `1` , `Boolen` , or `Yes` or `No` means `encoding` the data. But this data is called `Binary Data` .

Time Series Data :

Time series data is a sequence of data points collected or recorded at successive, evenly spaced points in time. It captures how a variable changes over time, allowing for the analysis of trends, patterns, seasonality, and forecasting future values based on historical observations. Examples include daily stock prices, monthly sales figures, monthly rainfall , and hourly temperature readings etc.

Spatial Data :

`Spatial data` , also known as `geospatial data` or `geographic information` , refers to data that represents the location , shape , and relationship of objects or features on the Earth's surface. It includes information about the position (coordinates) , geometry (points , lines , polygons) , and attributes (such as name , type , or value) of natural or man-made features. Examples include maps of cities , locations of weather stations , satellite imagery , and boundaries of countries or land parcels. The analysis of spatial data often involves geographic information systems (GIS) and spatial analysis techniques. The analysis of spatial is also called `geospatial analysis` .

Categorical Data :

The categorical data is the data that is classified into categories or groups. It is also known as `Qualitative Data` . We can group our data into categories like `Male` , `Female` , `Yes` , `No` , `High` , `Low` , `Blood Group AB` , and `Blood Group A` , etc.These examples scales are `Nominal` scales because it has no order e.g. i can't say that `Male` is greater than `Female` or vice versa etc.

Ordinal Categorical Data :

The categorical data is the data that is classified into categories or groups and also has some order. It is also known as `Ordinal Data` . We can group our data into categories like `Low` , `Medium` , `High` , `Yes` , `No` , and `Education Level` etc.

Note : `Columns` of the data frame are also called `variables` OR `attributes` .

Univariate Data :

The data that is consists of one column are called `univariate` data.

For example :

Height of students in a class.

Bivariate Data :

The data that is consists of two columns are called `Bivariate` data.

For example :

Height and Weight of students in a class.

Multivariate Data :

The data that is consists of multiple columns are called `multivariate` data.

For example :

Titanic dataset , Iris dataset , Economics indicators , or Sales dataset etc. Most of the big data consists of `multivariate` means `multiple columns`

Structured Data : (M.Important)

Structured data refers to information that is organized into a predefined format , typically in rows and columns , making it easily searchable , accessible , and analyzable. Examples include data stored in spreadsheets , relational databases , excel spreadsheets , google spreadsheets , and CSV files etc , where each field has a specific data type and meaning.

For example :

The data in spreadsheets like titanic dataset , or iris dataset etc.

Unstructured Data : (M.Important)

Unstructured data refers to information that does not have a predefined data model or is not organized in a systematic manner. It typically consists of text , images , audio , video , emails , social media posts , and other formats that are not easily searchable or analyzable using traditional databases. Unstructured data requires advanced techniques such as natural language processing , image recognition , or machine learning for extraction and analysis. It does not fit convential method of storing data means we cannot put the data in rows and columns.

For example :

Text files data , multimedia , webpages , audio , images , or video etc.

Textual Data :

The data that is consists of texts are called textual data.

For example :

The data that are stored in pdf forms , E - Books , Sotial media posts , or Comments etc.

Semi Structured Data :

The word `Semi` means `Half - Half` Or a mixture. In this case it is the mixture of `Structured` and `Unstructured` data are called semi structured data.

For example :

Emails headers are `structured` and their body are `unstructured` , and `Json` files etc.

Boolean Data : (M.Important)

The data that is consists of only two possible values are called `Boolean` data. It is similar to the `Binary` data. It is very popular data type in data science. It is mostly useful in `data collection` .

For example :

`On / Off` , `True / False` , or `Yes / No` etc.

Measurement : (M.Important)

Why we measure the data :

We measure our data to `label` our data. e.g. `Height` we need a standard unit (ft , inches , or cm etc) to tell the world what is my height ? , `Temperature` , and `Weight` etc.

Operationalization :

Definiton

Operationalization is a method of defining a concept or variable in a way that makes it measurable.

`OR`

It is the process of translating a theoretical concept into a concrete , observable , and quantifiable measure.

For example :

Sometime we got the data that has no `specific` method to `measure` them. e.g. `Happiness` , `Satisfaction` , `Stress` , `Quality of life` , and `Heartbeat` etc. We need to `operationalize` them. e.g. `Happiness` can be measured by `Smile` , `Laugh` , `Smile duration` , and `Smiliness` etc. , `Satisfaction` can be measured by `Customer satisfaction survey` , `Customer feedback` , `Customer rating` etc.

- ## Proxy Measurements : (M.Important)

  - **Proxy :** A proxy is a substitute `(Uss ki jaga)` or stand-in used to represent or estimate something else when direct measurement is not possible.
  - **Proxy Measurements :** A proxy measurement is an indirect way of measuring something when direct measurement is difficult or impossible. Instead , we use another related variable that can represent or estimate the concept we want to measure.

For example :

`A person's income` level can be difficult to measure directly, so we might use their job title or the neighborhood they live in as a proxy to estimate their income , `Tree age` can be difficult to measure directly , so we might use the diameter of the trunk as a proxy to estimate the age of the tree , or Suppose you want to measure `air pollution` in a city , but you do not have direct access to air quality sensors. Instead , you use the number of people visiting hospitals for respiratory problems as a proxy measurement. The hospital visits indirectly reflect the level of air pollution , even though you are not measuring pollution directly.

True & Error Score :

When we measure anything in this world there is high possiblity of having an error in it. So , its mean this world is based on `continuous` data. It's mean we cannot measure anything exactly there is an error in them and of course we usually ignore that error. So , we need to calculate the `True Score` and `Error Score` of our measurement.

For example :

Water in a glass suppose our measurement instrument shows it is `400` ml but , it is not a exact measurement of the water there is execption in them which is mostly are ignored , `weight` , and `height` etc.

Equation Of True & Error Score :

> X = T + E

where ,

- `X` is observed value.
- `T` is true score.
- `E` is error score.

This equation is mostly applied on `continuous` data.

For example :

If we measure the weight of a person as `70` kg , the true weight might be `70` kg , where `69.5` kg might be true score and `0.5` might be error score. So , the observed value is :

X = 69.5 kg + 0.5 kg
so ,
X = 70 kg

which is observed value and the error score is `0.5` kg which is ignored in our measurement.

Note :

I can control the error `E` or i cannot even control it e.g. lets say the error is `0.333 ...` and so on. So , i cannot control this error because of infinite maybe i will round off it but still it is uncontrolable and i can control the error e.g. the error is `0.5` so i can control this error.

These errors will highly effect our mechine learning algorithms and models. So , we need to control them. It's means we need to reduce them. So , we need to use `data preprocessing` techniques to reduce them. It's means it is very important to learn how to control `error` in our data.

Pro Tips For Statistics : (M.Important)

The errors are always present in our data. So , we need to control them. We need to minimize these erorrs. If we understand these errors we will understand the whole statistics analysis.

For example :

The best example to understand errors is `Chatgpt` model. This model is trained on `large` dataset. So , it is very accurate. When `Chatgpt` launched in this world version `3.5` it had many errors in them. But when there new version `4.0` is launched it very accurate then before now the question is how `Chatgpt` developers did this ? , the simple answer is they reduce errors that are in the previous version and trained the `Chatgpt` new verion `4.0` more accuratly but still it also have an errors but up coming version will solve this too.

Types Of Errors : (M.Important)

There are two types of error :

1. Random errors.
2. Systematic errors.

1. Random Errors :

Random errors are unpredictable variations that occur during measurement , caused by unknown or uncontrollable factors. They make results slightly different each time you measure , even under the same conditions.

For example :

When we throw a loodo daies all the time there output is different from previous one. This is due to random errors. It is unpredicatable and uncontrollable etc.

2. Systematic Errors :

Systematic errors are consistent , repeatable mistakes that occur due to flaws in equipment , measurement methods , or experimental design. They cause measurements to be biased in the same direction every time.

For example :

If a weighing scale is not calibrated correctly and always shows 2 kg more than the actual weight , every measurement you take will be consistently 2 kg higher than the true value. This is a systematic error. It means that the scale is not working properly and is giving wrong readings. It is a flaw in the equipment. It means the that are occur by instrument or by `system` is called systematic error. If we measure our weight again and again every time it shows me different this is called `random error` and it is also called `replicates` and for removing or minmizing random error just simply take there `mean ()` . and for removing systematic error we have to `calibrate` the instrument or fix the instrument.

Note :

Random error is more difficult to remove than systematic error.

To minimize the random error usually we take there `mean ()` but it is apply on `numerical data` .

To remove or minimize the random error usually we `replicate` the instrument.

The most important thing is accuracy of data. Data is very important for every thing if the data is accurate we can get accurate result. If the data is not accurate we can not get accurate result and we can also predict future using the data.

3. Type 1 Error :

Type 1 error is also called `alpha error` or `false positive` . It occurs when a true null hypothesis is rejected. It is a type of error that occurs when a false positive result is obtained.

It is a type of error that occurs when a true null hypothesis is rejected.

For example :

Suppose a medical test is designed to detect a disease. The null hypothesis is `the patient does not have the disease`. If the test result incorrectly indicates that the patient has the disease (when they actually do not) , this is a Type 1 error — a false positive.

Summary :

- **Null hypothesis :** Patient does not have the disease.
- **Test result :** Positive (says patient has the disease).
- **Reality :** Patient does not have the disease.
- **Error :** Type 1 error (false positive).

4. Type 2 Error :

Type 2 error is also called `beta error` or `false negative`. It occurs when a false null hypothesis is not rejected `OR` it is a type of error that occurs when a false negative result is obtained.

For example :

Suppose a medical test is designed to detect a disease. The null hypothesis is `the patient does not have the disease`. If the test result incorrectly indicates that the patient does not have the disease (when they actually do) , this is a Type 2 error — a false negative.

Summary :

- **Null hypothesis :** Patient does not have the disease.
- **Test result :** Negative (says patient does not have the disease).
- **Reality :** Patient actually has the disease.
- **Error :** Type 2 error (false negative).

Graphical Example Of Type 1 & Type 2 Error :

Why Do We Care About These Errors In Data Science ?