

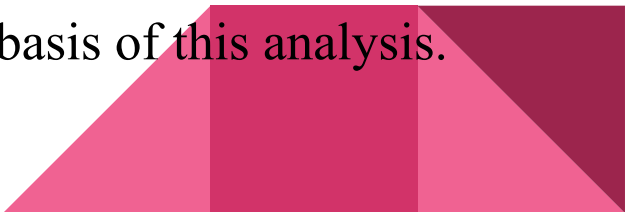
UNIT I

Descriptive Statistics and Introduction to Probability

- The basis of Data Science form by Probability and Statistics.
- For making prediction the probability theory is very useful.
- Estimates and predictions form an important part of Data science.
- With the help of statistical methods, we make estimates for the further analysis.
- Thus, statistical methods are largely dependent on the theory of probability.
- And all of probability and statistics is dependent on Data.



What is Statistics?

- Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions - Ross, Sheldon M. Introductory statistics. Academic Press, 2017.
 - Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis.- Croxton and Cowden
 - Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis.
- 

Major branches of statistics

1. Descriptive


Definition : The part of statistics concerned with the description and summarization of data is called descriptive statistics.

Descriptive statistical analysis helps us to understand our data and is very important part of Machine Learning.

2. Inference

Definition : The part of statistics concerned with the drawing of conclusions from data is called inferential statistics.

To be able to draw a conclusion from the data, we must take into account the possibility of chance- introduction to probability.



Population and sample

1. Population :


The total collection of all the elements that we are interested in is called a population.

OR

In a statistical enquiry, all the items, which fall within the purview of enquiry, are known as Population or Universe.

In other words, the population is a complete set of all possible observations of the type which is to be investigated.

For example : Total number of students studying in a school or college, total number of books in a library, total number of houses in a village or town are some examples of population.



2. Sample :

A subgroup of the population that will be studied in detail is called a sample
OR

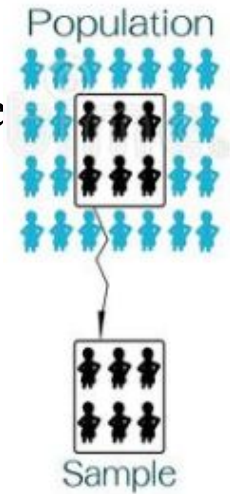
Statisticians use the word sample to describe a portion chosen from the population.

OR


A finite subset of statistical individuals defined in a population is called a sample.

The number of units in a sample is called the **sample size**.

For example : Suppose we want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income(\bar{x}) is \$34,500 (statistic). We conclude that the population mean income (μ) is likely to be close to \$34,500 as well.



Purpose of statistical analysis

- If the purpose of the analysis is to examine and explore information for its own intrinsic interest only, the study is descriptive.
 - If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.
 - A descriptive study may be performed either on a sample or on a population.
 - When an inference is made about the population, based on information obtained from the sample, does the study become inferential.
- 

Understanding the data

What is Data ?

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

OR

Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

OR

Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process



Why does Data Matter?

- Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
- Helps in predicting the future or forecast based on the previous data.
- Helps in determining patterns that may exist between data.
- Helps in detecting fraud by uncovering anomalies in the data.



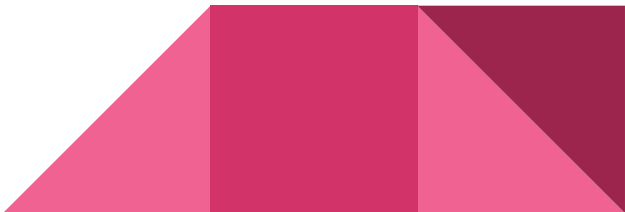
Why do we collect Data?

To study the characteristics of some group or groups of people, places, things, or events.

Example:

1. To know about temperatures in a particular month in Chennai, India.
2. To know about the marks obtained by students in their Class 12.
3. To know how many people like a new song/product/video-collected through comments.

Data collection

- **Data available:** published data.
 - **Data not available:** need to collect, generate data.
- 

Unstructured and structured data

- To use the information from the database we must know the context of the numbers and text.
- When they are scattered about with no structure, the information is of very little use.
- Hence, we need to organize data

Dataset

- It is structured collection of data could be numbers, name, roll numbers.
- In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.
- <https://docs.google.com/spreadsheets/d/15nJvZ-xBZDGb0oii-NCvSIY4fETotXcJdm5pV1Fq2aI/edit#gid=0>

Variables and cases

1. **Case (observation):** A unit from which data are collected
2. **Variable:** A characteristic or attribute that varies across all units.

For Example

In school data set:

Case: each student

Variable: Name, marks obtained, Board etc.

Rows represent cases: for each case, same attribute is recorded

Columns represent variables: For each variables, same type of value for each case is recorded.



Categories of data:

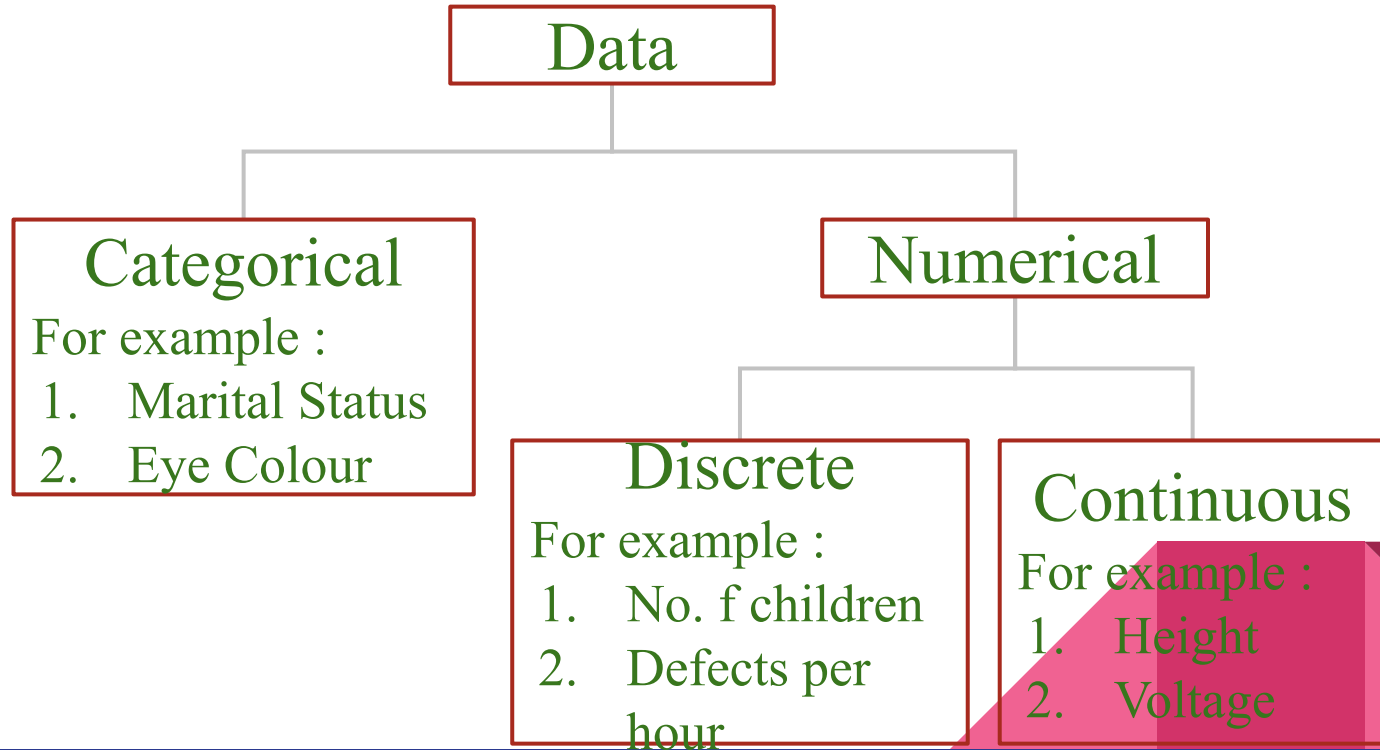
Any statistical data can be classified under two categories depending upon the sources utilized. These categories are,

1. **Primary data** : Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by survey conducted by individuals or research institution or any organisation.
2. **Secondary data** : Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency.



Classification of data

I. Categorical and numerical



Categorical and numerical variables


1. Categorical data

- It is also called qualitative variables.
- Identify group membership

2. Numerical data

- It is also called quantitative variables.
- Describe numerical properties of observations
- It has measurement units

Measurement units: Scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimeters, etc.



II. Cross-sectional and time-series data

1. **Time series** - data recorded over time
2. **Timeplot** – graph of a time series showing values in chronological order
3. **Cross-sectional** - data observed at the same time

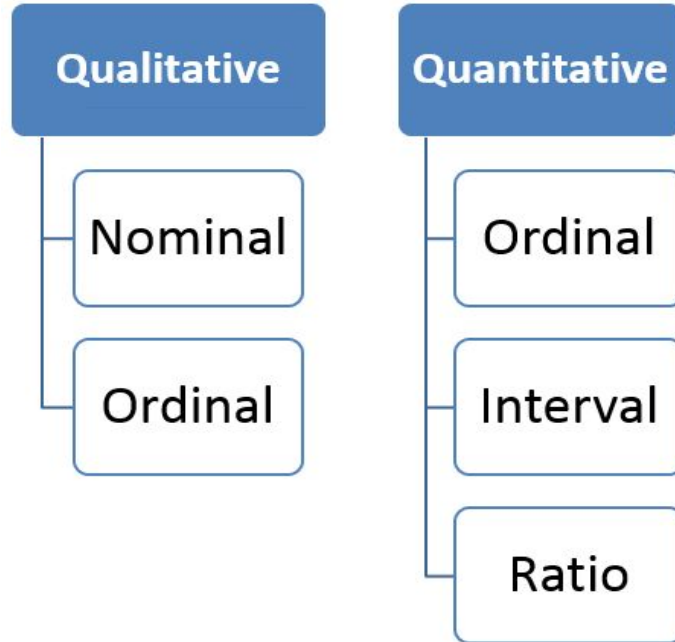


Time-series data- Example

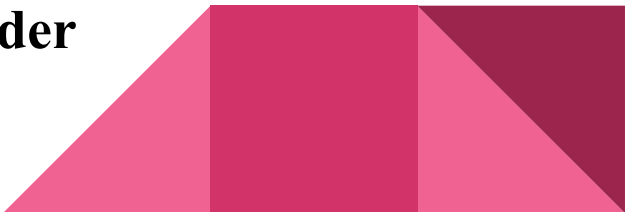
<i>Date</i>	<i>Potato</i>		
	<i>Qty(KG)</i>	<i>cost (Rs.)</i>	<i>Selling price(Rs.)</i>
01-Mar	0	21	24
02-Mar	1350	20.05	24
03-Mar	675	20.5	24
04-Mar	0	NA	NA
05-Mar	675	20.8	24
06-Mar	675	21.25	24
08-Mar	20	20.5	24
09-Mar	900	20.5	24
10-Mar	900	20.5	24
11-Mar	0	NA	NA
12-Mar	900	20.3	24
13-Mar	1125	19.4	22
15-Mar	1125	18.8	22

Scales of measurement

Data collection requires one of the following scales of measurement:
nominal, ordinal, interval, or ratio.



1. Nominal scale of measurement

- When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale.
 - Data at this level is categorized using names, labels or qualities.
Examples: Name, Board, Gender, Blood group, Brand name, Zipcode etc.
 - Sometimes nominal variables might be numerically coded.
For example: We might code Men as 1 and Women as 2. Or Code Men as 3 and Women as 1. Both codes are valid.
 - There is no ordering in the variable.
 - **Nominal: name categories without implying order**
- 

2. Ordinal scale of measurement

- Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered a ordinal scale.
- Data at this level can be arranged in order or ranked and can be compared. **eg:** Grades, Star Reviews, Position in Race, Date
- Each customer who visits a restaurant provides a service rating of excellent, good, or poor.
 - The data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data.
 - In addition, the data can be ranked, or ordered, with respect to the service quality.
- **Ordinal – name categories that can be ordered**

3. Interval scale of measurement

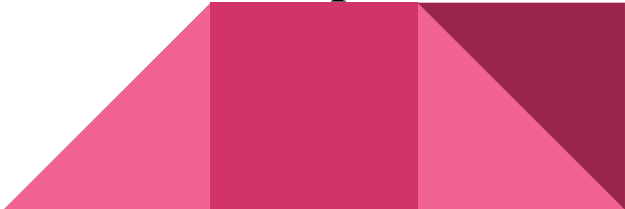
- If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale.
- Interval data are always numeric. Can find out difference between any two values.
- Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. **eg:** Temperature in Celsius, Year of Birth
- Ratios of values have no meaning here because the value of zero is arbitrary.
- **Interval: numerical values that can be added/subtracted (no absolute zero)**

Example: temperature

- Suppose the response to a question on how hot the day is comfortable and uncomfortable, then the temperature as a variable is **nominal**.
- Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is **ordinal**.
- Example: Consider a AC room where temperature is set at 20°C and the temperature outside the room is 40°C . It is correct to say that the difference in temperature is 20°C , but it is incorrect to say that the outdoors is twice as hot as indoors.
- Temperature in degrees Fahrenheit or degrees centigrade is an interval variable. No absolute zero.

	Celsius	Fahrenheit
Freezing point	0	32
Boiling point	100	212

4. Ratio scale of measurement

- If the data have all the properties of interval data and the ratio of two values is meaningful, then the scale of measurement is ratio scale.
 - Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points.
 - A ratio scale is a quantitative scale where there is a true zero and equal intervals between neighboring points. Unlike on an interval scale, a zero on a ratio scale means there is a total absence of the variable you are measuring.
 - **Example:** height, weight, age, marks, Length, area, and population etc.
 - **Ratio: numerical values that can be added, subtracted, multiplied or divided (makes ratio comparisons possible)**
- 

True zero exists-ratios possible

Ratio Scale

Age, height, weight, marks etc.

Numerical Data



No absolute zero.
Difference exists

Interval Scale

Temperature, GPA etc.

Named + ordered categories

Ordinal Scale

Ranking, rating etc.

Categorical Data

Named categories

Nominal Scale

Name, Blood group etc.

	Nominal	Ordinal	Interval	Ratio
Categories	●	●	●	●
Rank order		●	●	●
Equal spacing			●	●
True zero				●

1. <https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/>
2. <https://www.scribbr.com/statistics/ratio-data/>
3. [Introduction to Descriptive Statistics and Probability for Data Science | by Abhishek Kumar | Towards Data Science](#)
4. [Probability and Statistics for Data Science Part-1 | by Badreesh Shetty | Towards Data Science](#)
5. <https://medium.com/analytics-vidhya/statistics-101-grouped-and-ungrouped-data-lets-talk-with-data-5cbf18c2feb9>

Ungrouped and Group data

Ungrouped Data :

- The statistical data collected are generally raw data or ungrouped data.
- Ungrouped data which is also known as raw data is data that has not been placed in any group or category after collection.
- Data is categorized in numbers or characteristics therefore, the data which has not been put in any of the categories is ungrouped.
- Here the data are presented in a way that exact measurement of units are clearly indicated.
- For example, when conducting census and you want to analyze how many women above the age of 45 are in a particular area, you first need to know how many people reside in that area.
- For example, consider the following :
- Height of students:
(171,161,155,155,183,191,185,170,172,177,183,190,139,149,150,150,152,158,159,174,178,179,190,170,143,165,167,187,169,182,163,149,174,174,177,181,170,182,170,145,143) :
This is raw/ungrouped data.

Grouped Data :

- Grouped data is the type of data which is classified into groups after collection. The raw data is categorized into various groups and a table is created.
- When raw data have been grouped in different classes then it is said to be grouped data.
- The primary purpose of the table is to show the data points occurring in each group.
- For instance, when a test is done, the results are the data in this scenario and there are many ways to group this data.
- For example, the number of students that scored above each 20 mark can be recorded.

For previous example :

Height Intervals(in cms)	No of students(F)
131-140	1
141-150	7
151-160	5
161-170	9
171-180	9
181-190	10
Total	41