B. Tech Computer Science & Engineering, CSE

(Semester-V)

MATH2361PROBABITY AND STATISTCS (3 credits)

UNIT-I: Data Science and Probability

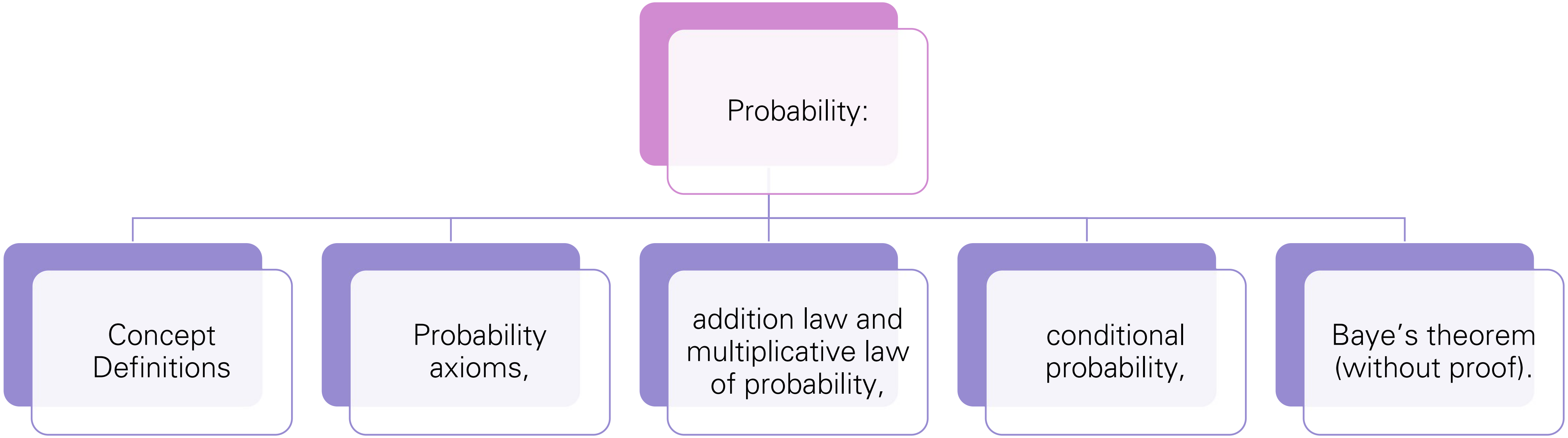
Dr Malikarjuna Reddy Doodipala
MSc, M.Phil, PGDCA, Ph D,
Associate Professor
Department of Mathematics
GITAM Hyderabad

UNIT-I: Data Science and Probability

Data Science:

- Introduction to Statistics
- Population Vs Sample
- Collection of Data-primary and Secondary Data,
- Types of Variables: Dependent, Independent, Categorical and Continuous Variables
- Data Visualization
- Measures of Central Tendency,
- Measures of Dispersion (Variance)

UNIT-I: Data Science and Probability


Probability:
Concept Definitions
Probability axioms,
addition law and multiplicative law of probability,
conditional probability,
Baye’s theorem (without proof).

UNIT-I: Outcomes

After completing this unit, the student will be able to

- Summarize the basic concepts of data science and its importance in engineering (L3).
- Analyse the data quantitatively or categorically and measure averages and variability (L4).
- Define the terms trial, events, sample space, probability and laws of probability (L3).
- Make use of probabilities of events in finite sample spaces from experiments (L3).
- Apply Baye’s theorem to practical problems (L3).

Data science

Data science is the study of data to extract meaningful insights for business.

It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.

Introduction to Statistics

- Today, there have been advancements in all sectors like **Commerce, Economics, Mathematics**, and other disciplinaries such as Technology etc.
- Not only that, but our life has also been going through a lot of development in various zones.
- Some of them are **defense, banking, and hospitality.** However, all of these depend largely on “statistics”.

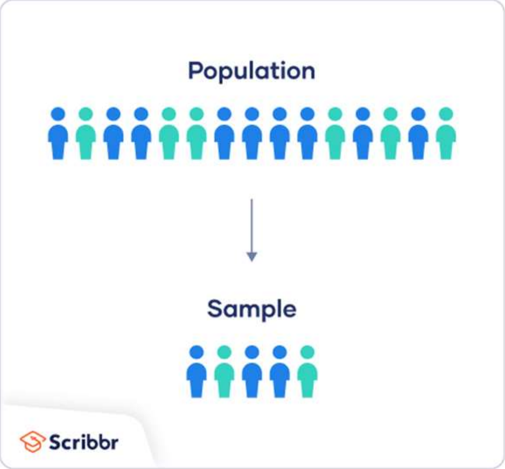
What is Statistics & Data Science?

- Statistics is a mathematically-based field that seeks to collect and interpret quantitative and quantitative data
- Data science is a multidisciplinary field that uses scientific methods, processes, and systems to extract knowledge from data in a range of forms. Data scientists use methods from many disciplines, including statistics.
- While data science focuses on comparing many methods to create the best machine learning model, statistics instead improves a single, simple model to best suit the data.

Introduction to Statistics

- Statistics play a very vital role in any domain. It helps in collecting data, be it in any field.
- Along with that, it also helps in analyzing data using statistical techniques. Speaking of the present time, it has a lot of importance and application. Furthermore, if we talk about the examples, here they are.
- The government uses statistics to conduct planning in the economic sector.
- A businessman looks forward to expanding his growth in the business world by taking into the account the data and feedback.
- Similarly, politics makes use of statistics to show their accomplishments.
- For showing research papers, scholars use statistics.
- Therefore, the list and applications of statistics are endless.

PopulationVs Sample



Population is an aggregate of objects.

A population is the entire group that you want to draw conclusions about.

A sample is the specific group that you will collect data from.

The sample size is always less than the total size of the population.

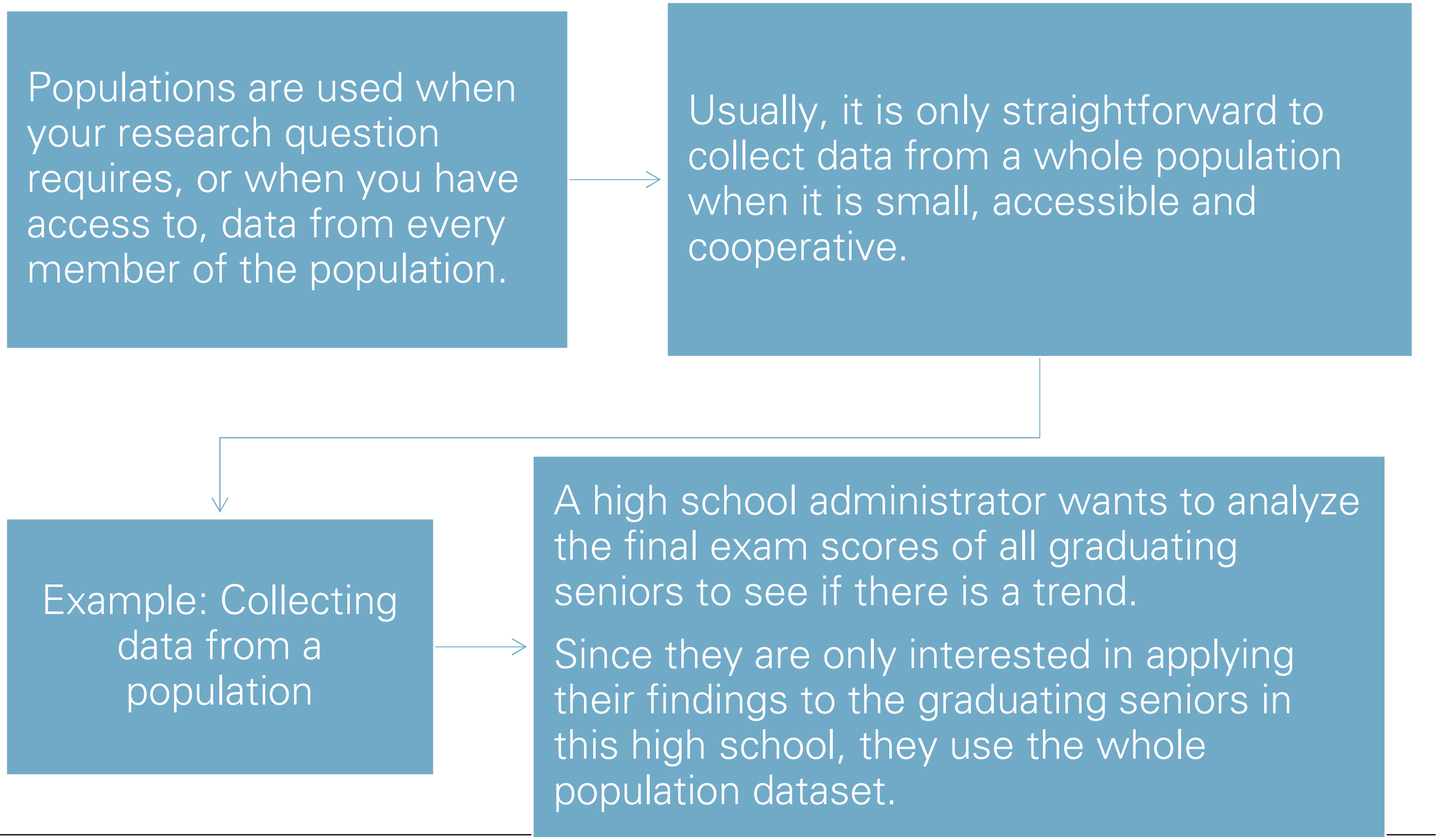
In research, a population doesn't always refer to people.

It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.

Population Vs Sample

Population(N)	Sample(n)
Advertisements for IT jobs in the in India	The top 50 search results for advertisements for IT jobs in the India on April 1, 2021
Songs from the Song Contest	Winning songs from the Song Contest that were performed in Hindi
Undergraduate students in the in India	300 undergraduate students from three deemed universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2010

Collecting data from a population



Populations are used when your research question requires, or when you have access to, data from every member of the population.

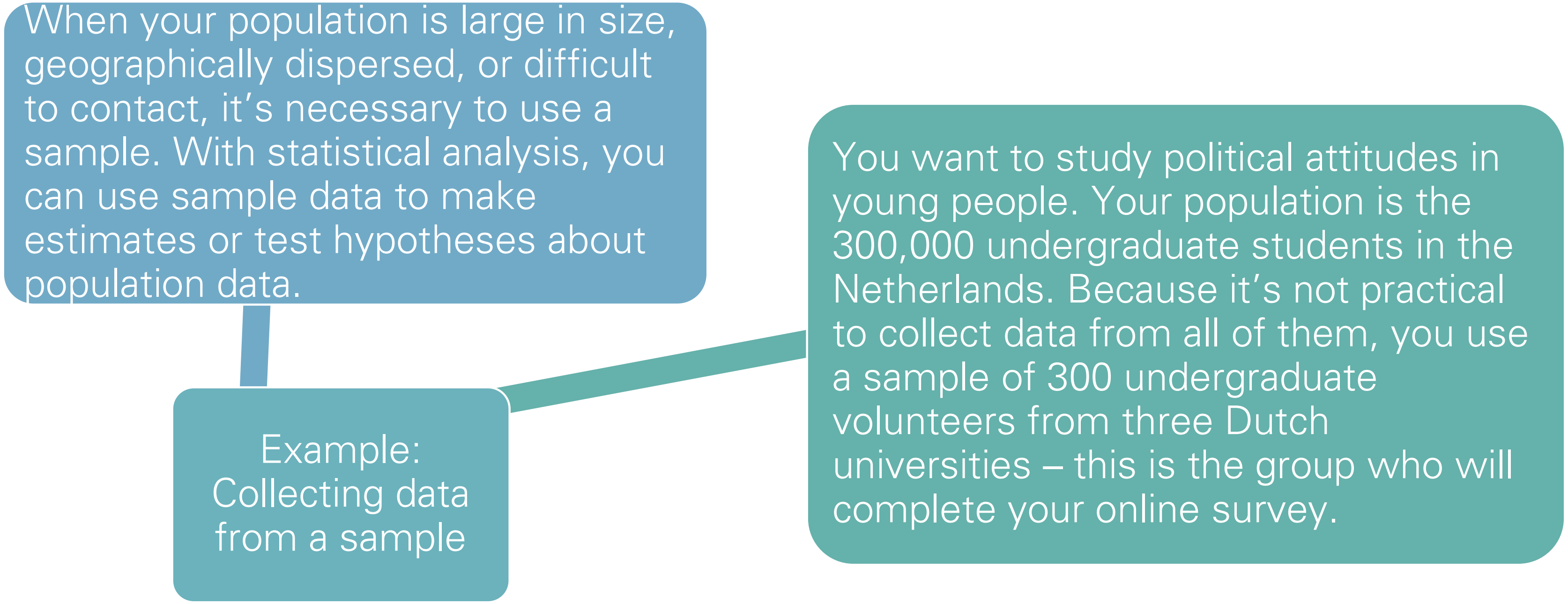
Usually, it is only straightforward to collect data from a whole population when it is small, accessible and cooperative.

Example: Collecting data from a population

A high school administrator wants to analyze the final exam scores of all graduating seniors to see if there is a trend.

Since they are only interested in applying their findings to the graduating seniors in this high school, they use the whole population dataset.

Collecting data from a sample



When your population is large in size, geographically dispersed, or difficult to contact, it’s necessary to use a sample. With statistical analysis, you can use sample data to make estimates or test hypotheses about population data.

Example: Collecting data from a sample

You want to study political attitudes in young people. Your population is the 300,000 undergraduate students in the Netherlands. Because it’s not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers from three Dutch universities – this is the group who will complete your online survey.

Collecting data from a sample

- Ideally, a sample should be randomly selected and representative of the population.
- Using probability sampling methods (such as simple random or stratified sampling) reduces the risk of sampling bias and enhances internal and external validity.
- For practical reasons, researchers often use non-probability sampling methods.
- Non-probability samples are chosen for specific criteria; they may be more convenient or cheaper to access. Because of non-random selection methods, any statistical inferences about the broader population will be weaker than with a probability sample.

Reasons for Sampling

Necessity: Sometimes it’s simply not possible to study the whole population due to its size or inaccessibility.

Practicality: It’s easier and more efficient to collect data from a sample.

Cost-effectiveness: There are fewer participant, laboratory, equipment, and researcher costs involved.

Manageability: Storing and running statistical analyses on smaller datasets is easier and reliable.

Population parameter vs sample statistic

When you collect data from a population or a sample, you can calculate various measurements and numbers from the data.

A parameter is a measure that describes the whole population.

A statistic is a measure that describes the sample.

You can use estimation or hypothesis testing to estimate how likely a sample statistic is to differ from the population parameter.

Origin and History

- Statistics is a very well-known term in the history, be it ancient or medieval. However, there are still a few unanswered questions.
- One such question is – “origin of the word ‘statistics’.”
- There are several views related to the same.
- One such view is that it has a Latin origin and the word that it comes from is ‘status.’
- On the contrary, another view speaks of its Italian origin and that it comes from ‘statista.’
- According to scholars, the origin is German and the word it comes from is ‘statistik.’ Similarly,
- according to more suggestion, the origin is traced back to a French word called ‘statistique.’
- Each of the term means political state.
- In the past, statistics was all about “collection” of data. Also, the goal was to maintain the data for the welfare of the everyone in the area.
- According to various calculations, there were several predictions that led to one or the other answer.

What is Statistics?

Statistics can come forward in two ways: singular and plural.

In plural form, statistics is quantitative as well as qualitative.

In the plural sense, data is generally taken into account keeping in mind the statistical analysis.

Singularly, it is more like a scientific method that helps in presenting, collecting, as well as analyzing data.

All of this brings some major characteristics into the limelight.

Definition of Statistics

- Statistics has been defined differently by different Authors as a statistical and statistical method
- According to Webster "Statistics are the classified facts representing the conditions of the people in a state. Especially those facts which can be stated in numbers or any tabular or classified arrangement."
- According to Bowley statistics are "Numerical statements of facts in any department of inquiry placed in relation to each other."
- According to Yule and Kendall, statistics means quantitative data affected to a marked extent by multiplicity of causes.
- More broad definition of statistics was given by Horace Secrist.
- According to him, statistics means aggregate of facts affected to marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.

Definition of Statistics (Cont’d)

This definition points out some essential characteristics that numerical facts must possess so that they may be called statistics.

These characteristics are:

They are enumerated or estimated according to a reasonable standard of accuracy

They are affected by multiplicity of factors.

They must be numerically expressed.

They must be aggregate of facts.

Definition of Statistics (Cont'd)

Modern Definition

"Statistics are the numerical statement of facts capable of analysis and interpretation and the science of statistics is the study of the principles and the methods applied in collecting, presenting, analysis and interpreting the numerical data in any field of inquiry."

Limitations

Limitations come a lot before directly applying the statistical methods.

It is necessary to be aware of it in order to move ahead. Some of the primary limitations of statistics are:

Statistics is all about “aggregates.” Be it an individual or a statistician, they are all a part of the aggregate.

It also deals with quantitative data. However, it is not a very difficult task to do a conversion from qualitative to quantitative.

All that is needed is the numerics and description related to the qualitative data.

Limitations

In order to propose specific projections, i.e. sales, price, quantity and so on, there is a requirement of a set of conditions.

So, if, by any chance, these conditions turn out to be wrong or are violated, there is a chance that the projections and its outcome will be inaccurate.

Statistical inferences make use of random sampling options.

Hence, not following the rules for sampling would be a very bad idea as it can lead to wrong results.

The conclusions coming off would have errors. So, the idea here is to consult the experts before hopping into the sampling scheme, directly.

Question:
What are the different branches of statistics?

- Basically, there are two branches of statistics.
- They are – Descriptive and Inferential.
- Descriptive:
- This branch deals with the basic and major aspects related to numerics.
- The numerics and data contain graphs, tables, and many more quantities.
- These quantities help with serving information.

Question: What are the different branches of statistics? (Cont'd)

Inferential: This branch deals with making inferences about the large data group. The knowledge for making inferences generally comes from samples.Sample evidence brings out inferences.

Data

We do not generally associate data with mathematics. However, data is the base of all operations in statistics.

So let us learn more about data collection, primary data, secondary data, and a few other important terms.

Data can be defined as a systematic record of a particular quantity.

It is the different values of that quantity represented together in a set.

It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis.

When arranged in an organised form, can be called information. The source of data (primary data, secondary data) is also an important factor.

Types of Data

Data may be qualitative or quantitative.

Once you know the difference between them, you can know how to use them.

Qualitative Data:

- They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated.
- For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative.
- They are more exploratory than conclusive in nature.

Types of Data (cont’d)

Quantitative Data:

- These can be measured and not simply observed.
- They can be numerically represented, and calculations can be performed on them.
- For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport.
- This information is numerical and can be classified as quantitative.

Collection of Data : Primary Data

Depending on the source, it can classify as primary data or secondary data. Let us take a look at them both.

Primary Data

- These are the data collected for the first time by an investigator/Surveyor for a specific purpose.
- Primary data are 'pure' since no statistical operations have been performed on them and they are original. An example of primary data is the Census of India.

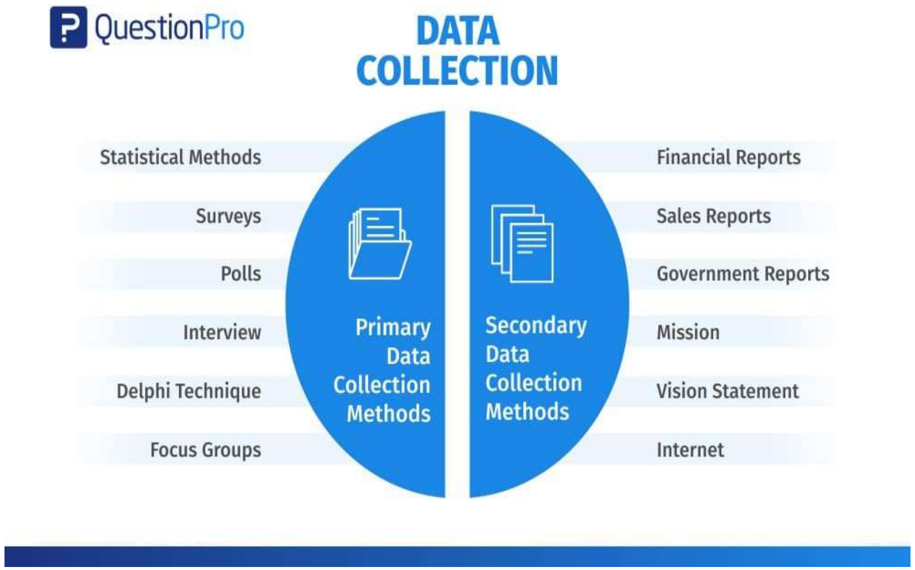
Data Collection: Secondary Data

They are the data that are sourced from some place that has originally collected it.

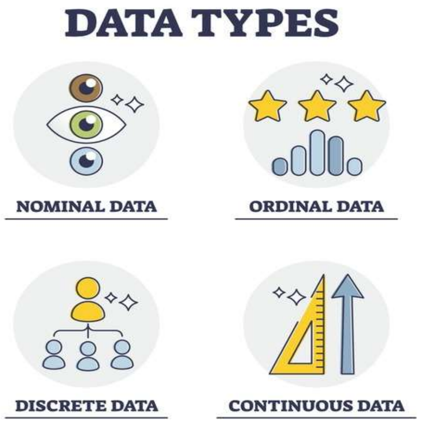
This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form.

This information is impure as statistical operations may have been performed on them already.

An example is an information available on the Government of India, the Department of Finance's website or in other repositories, books, journals, etc.



Discrete and Continuous Data



Discrete Data:

These are data that can take only certain specific values rather than a range of values.

- For example, data on the blood group of a certain population or on their genders is termed as discrete data.
- A usual way to represent this is by using bar charts.

Discrete and Continuous Data (Cont'd)

Continuous Data:

- These are data that can take values between a certain interval (range) with the highest and lowest values.
- The difference between the highest and lowest value is called the range of data.

For example, the age of persons can take values even in decimals or so is the case of the height and weights of the students of your school.

- These are classified as continuous data.
- Continuous data can be tabulated in what is called a frequency distribution.
- They can be graphically represented using histograms.

Thank You !