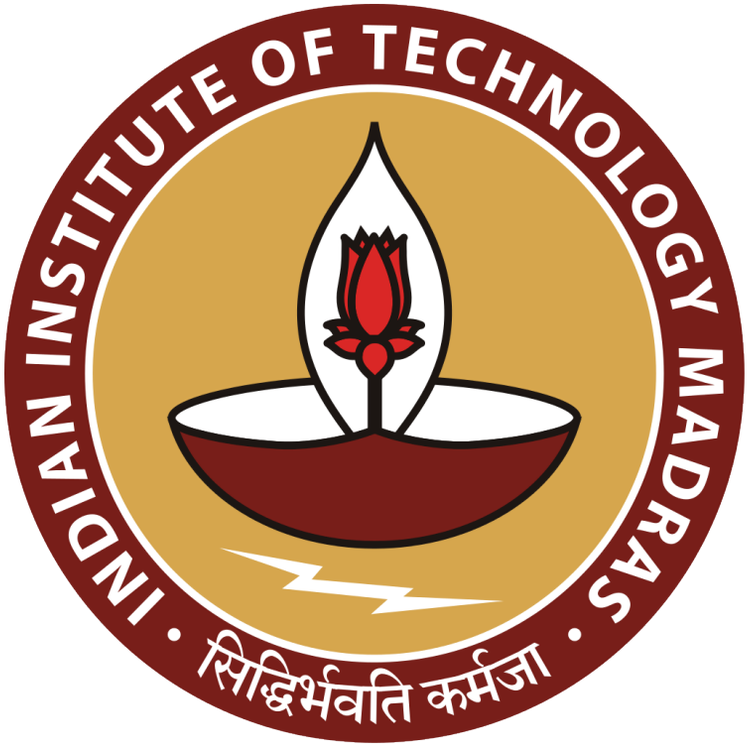

IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Introduction and types of data

Usha Mohan

Indian Institute of Technology Madras

Learning objectives

1. What is statistics?
   - Descriptive statistics, inferential statistics.
   - Distinguish between a sample and a population.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - classify data as categorical(qualitative) or numerical(quantitative) data.
   - Understand cross-sectional versus time-series data.
   - Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.

What is Data

In order to learn something, we need to collect data.

Definition

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

- Statistics relies on data, information that is around us.

Why do we collect Data

Why do we collect Data

- Interested in the characteristics of some group or groups of people, places, things, or events.

Why do we collect Data

- Interested in the characteristics of some group or groups of people, places, things, or events.
- Example: To know about temperatures in a particular month in Chennai, India.

Why do we collect Data

- Interested in the characteristics of some group or groups of people, places, things, or events.
- Example: To know about temperatures in a particular month in Chennai, India.
- Example: To know about the marks obtained by students in their Class 12.

Why do we collect Data

- Interested in the characteristics of some group or groups of people, places, things, or events.
- Example: To know about temperatures in a particular month in Chennai, India.
- Example: To know about the marks obtained by students in their Class 12.
- 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.

Data collection

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

We assume data is available and our objective is to do a statistical analysis of available data.

Unstructured and structured data

Unstructured and structured data

- For the information in a database to be useful, we must know the context of the numbers and text it holds.

Unstructured and structured data

- For the information in a database to be useful, we must know the context of the numbers and text it holds.
- When they are scattered about with no structure, the information is of very little use.

Unstructured and structured data

- For the information in a database to be useful, we must know the context of the numbers and text it holds.
- When they are scattered about with no structure, the information is of very little use.
- Hence, we need to organize data

Dataset

- A structured collection of data.
- it is a collection of values-could be numbers, names, roll numbers.
- https://docs.google.com/spreadsheets/d/15nJvZ-xBZDGb0oii-NCvSIY4fETotXcJdm5pV1Fq2aI/edit?usp=sharing
- https://docs.google.com/spreadsheets/d/1qZWmXsIpFx10srpFcmj9DPA961UMbTXkCiUr_SxBYq4/edit?usp=sharing
- https://docs.google.com/spreadsheets/d/1lrmhe-E0A2LWpTB9cBK9dm-sL2SPVXYZl0MJHI6vqhM/edit?usp=sharing

Variables and cases

Variables and cases

- Case (observation): A unit from which data are collected

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that “varies”.

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.
- In our school data set:

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.
- In our school data set:
  - Case: each student

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.
- In our school data set:
  - Case: each student
  - Variable: Name, marks obtained, Board etc.

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.
- In our school data set:
  - Case: each student
  - Variable: Name, marks obtained, Board etc.
- Rows represent cases: for each case, same attribute is recorded

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that "varies".
  - Formally: A characteristic or attribute that varies across all units.
- In our 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.

Summary

We have organized data in a spreadsheet into a table

Each variable must have its own column.

Each observation must have its own row.