

# Introduction to Statistics

---

Md. Aminul Islam Shazid

# Outline

- 1 Introduction
- 2 Some Basic Statistical Concepts
- 3 Variable and Measurement
- 4 Sources of Data

# Introduction

---

# What is Statistics?

- Collecting data
- Gaining insights from data
- Making decisions based on the insights gained from the data

# Definition

Statistics can be defined as the art and science of:

- collecting, cleaning and organizing data
- summarizing and analyzing data
- presenting the summary or the analysis
- interpreting the analysis results
- gaining insights through analysis of data
- and finally, drawing valid conclusions and making sound decisions through the use of data.

# Why Statistics is Necessary

-

# Some Basic Statistical Concepts

---

# Population and Sample

- **Population** is the collection/set of all items or individuals of interest in a given study
- **Sample** is a *representative* portion of the population

For example:

- A study may target all the people in Bangladesh. However, it is unfeasible to collect information of everyone in the country in a timely or cost-effective way
- Therefore, data is collected from only a small portion of people from *all over the country*, this is called sampling. The individuals in a sample are usually selected randomly



# Census and Survey

# Parameter and Statistic

- A **parameter** is a characteristic or function of every objects or individuals in a population. For a fixed population, it is a fixed (but, usually unknown) value
- A **statistic** is a characteristic or function of every objects or individuals in a sample.
- A **statistic** is used to *estimate* a **parameter**

## Parameter and Statistic (cont.)

- For a fixed population, the value of a parameter is fixed (but usually unknown)
- However, due to randomization, different samples can include different individuals from a population
- Therefore, the value of a statistic can vary across different samples

## Parameter and Statistic (cont.)

For example:

- Suppose the goal is to find the average height of the students of a class
- The population average is a fixed value and it is unknown unless data is collected from everyone in the class
- If the heights of a some students are collected as a random sample, then we can estimate the population average using the sample average
- If another sample is collected, the same individuals as the first sample may not be selected, therefore, the estimate shall be different from the first estimate

# Types of Statistics

- **Descriptive statistics:** Methods for organizing, summarizing and presenting data in an informative way. For example:
  - A hypothetical customer survey finds that 50% of the customers are satisfied with a product
- **Inferential statistic:** Methods for using sample data to make predictions, test hypotheses, and generalize conclusions about a larger population. For example:
  - A study finds association between smoking and cancer

# Variable and Measurement

---

# Variable

- Variable means something that can vary
- It is a characteristic that can vary across individuals or objects or items or cases of a phenomenon
- For example:
  - Age
  - Gender
  - Socioeconomic status
  - Temperature

# Types of Variables

Variables:

- Qualitative
- Quantitative:
  - Discreet
  - Continuous



# Qualitative vs Quantitative

- **Qualitative:** Qualitative variables describe qualities and are categorical. These are non-numerical and descriptive values that represent attributes or categories. For example:
  - Name of a person
  - Gender
  - Hair colour
- **Quantitative:** Quantitative variables measure quantities with numbers. These are numeric data that can be counted or measured, allowing for mathematical calculations. For example:
  - Height
  - Temperature
  - Number of students in a class

# Discrete vs Continuous

- **Discrete:** *Countable*, usually whole numbers. Finite number of possible values. For example:
  - Number of students in a class
- **Continuous:** *Measurable*, can have fractional values. Can take values in a given range. Infinite number of possible values in any range. For example:
  - Height

# Scales of Variables or Level of Measurement

Scales of variables, or levels of measurement, define how data is categorized and quantified/measured. It dictates which statistical analysis is appropriate for a variable.

There are four levels of measurement:

- Nominal
- Ordinal
- Interval
- Ratio

Nominal and ordinal are for categorical or qualitative data, and the last two are for quantitative data.

# Nominal

- Nominal scale is for variables with categories with no inherent order or ranking (labels/names only)
- Example:
  - Gender
  - Religion
- Allowed operations:
  - We can only **count** the occurrence or frequency of each of the categories of a nominal variable

# Ordinal

- Ordinal scale is for variables with categories that can be *ordered* or ranked, but differences between ranks are not necessarily equal or measurable
- Example:
  - Satisfaction level (low, medium, high)
  - Race position (first, second, third)
  - Socioeconomic status (lower, mid, upper)
- Allowed operations:
  - We can **count** the frequency of each of the categories of an ordinal variable as well as **order** the categories

# Interval

- It is for numerical variables. The values are ordered, with equal, meaningful *intervals* between values
- There is no true zero point
- Example:
  - Temperature (0°C does not imply the lack of heat or thermal energy)
  - Years on a Calendar
- Since there is no true zero point, taking ratio of two values of an interval scale variable is meaningless
- Allowed operations:
  - We can **add**, **subtract** different values of an interval scale variable

# Ratio

- An ratio scale variable includes all properties of interval scales, plus a true zero point, allowing for meaningful ratios
- Example:
  - Height
  - Weight
  - Income
- Allowed operations:
  - All arithmetic operations (addition, subtraction, multiplication, division, ratios)

## Sources of Data

---



# Data

- Data in statistics consists of numerical or qualitative facts collected for analysis
- They are recorded observations about individuals, objects, or events
- Data arise from measurement, counting, surveys, or experiments
- Each observation consists of one or more variables
- Usually presented in a tabular form, can also be in other forms (such as images, audio, video etc.)
- Statistics uses data to summarize information and draw conclusions about a larger context

# Sources of Data

- **Primary data:** data collected firsthand by the researcher for a specific study
- **Secondary data:** data previously collected by others for a different or a more general purpose
- **Tertiary data:** data that summarize, compile, or index primary and secondary sources

# Primary Data

Data collected firsthand by the researcher for a specific study.

- Advantages:
  - Directly relevant to the research objectives
  - Greater control over data quality and measurement methods
  - Up-to-date and specific to the population of interest
  - Clear understanding of how the data were collected
- Disadvantages:
  - Time-consuming to collect
  - Often expensive in terms of money and resources
  - Requires careful planning and technical expertise
  - Limited scope compared to large existing datasets

# Secondary Data

Data previously collected by others for a different or a more general purpose

- Advantages:
  - Quick and inexpensive to obtain
  - Often covers large populations or long time periods
  - Useful for comparisons and trend analysis
  - No need for data collection infrastructure
- Disadvantages:
  - May not exactly match the research objectives
  - Limited control over data quality and measurement methods
  - Possible issues with outdated or incomplete data
  - Documentation and variable definitions may be unclear

# Other Ways to Classify Data

- Observational data
- Experimental data
- Administrative records/registers
- Publications and reports

**Questions?**

---