

UNIT 1

Introduction to Statistics

(3 Hours) — Probability and Statistics (CACS202)

1. Meaning, Scope and Limitations of Statistics

1.1 Meaning of Statistics

The word **statistics** is used in two senses. In the **plural sense** it means numerical facts or **data** — for example, the marks of students or the monthly sales of a shop. In the **singular sense** it is a **science** (method) of collecting, organizing, presenting, analysing and interpreting data so that sound decisions can be made.

Definition (Croxtan and Cowden)

Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data.

Statistics is divided into two branches:

Branch	Purpose	Examples
Descriptive statistics	Summarize and describe the data at hand	Mean, median, standard deviation, tables, graphs
Inferential statistics	Draw conclusions about a population from a sample	Estimation, hypothesis tests, ANOVA, regression

The main **stages of a statistical investigation** are: (1) collection of data, (2) organization (editing, classification, tabulation), (3) presentation (tables, diagrams, graphs), (4) analysis (averages, dispersion, correlation, etc.) and (5) interpretation and reporting.

1.2 Scope of Statistics

Statistics is used in almost every field of human activity:

- **Business and economics** — sales forecasting, quality control, market research, pricing and risk.
- **Government and planning** — census, national income, poverty measurement, budget preparation.

- **Computer science and IT** — data mining, machine learning, network traffic modelling, software reliability and performance testing.
- **Science and medicine** — design of experiments, clinical trials, epidemiology.
- **Social sciences and education** — opinion surveys, test scores, demographic studies.
- **Research** — testing hypotheses and validating results in every discipline.

1.3 Limitations of Statistics

- Statistics deals with **aggregates**, not with individual items; it cannot study a single case.
- It deals only with **quantitative** data (qualitative facts must be converted to numbers).
- Statistical results are **true on average** and hold for large groups, not for every individual.
- It can be **misused** — wrong sampling, selective data or misleading graphs give false conclusions. “Figures do not lie, but liars can figure.”
- Results depend on the **quality of the data**; wrong data gives wrong conclusions, and statistics alone cannot prove cause and effect.
- It is only one of the tools of investigation and needs subject knowledge to interpret.

2. Types and Sources of Data

2.1 Types of Data

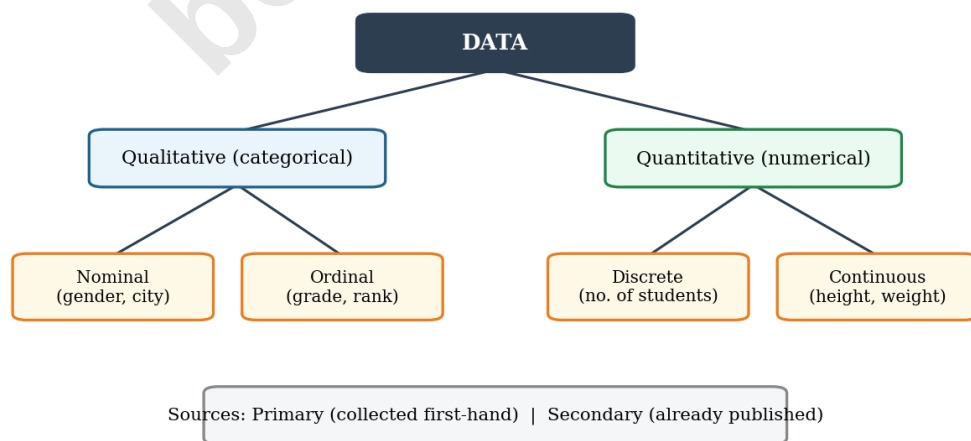


Figure 1.1 — Classification of data and its sources

Type	Description	Examples
Qualitative (categorical)	Describes a quality or category; not measured numerically	Gender, religion, colour, grade

Type	Description	Examples
Nominal	Categories with no natural order	Blood group, city, faculty
Ordinal	Categories with a meaningful order	Rank, satisfaction level (low / medium / high)
Quantitative (numerical)	Measured or counted numerically	Age, income, marks
Discrete	Countable, takes separate (integer) values	Number of students, number of defects
Continuous	Can take any value in an interval	Height, weight, time, temperature

Data are also classified by time or structure: **cross-sectional data** (many units at one time), **time-series data** (one unit over time) and **panel data** (many units over time). By **scale of measurement** there are four levels — **nominal**, **ordinal**, **interval** (equal intervals but no true zero, e.g., temperature in °C) and **ratio** (true zero, e.g., weight, income).

Example 1.1 — Identify the type of data

Classify: (a) number of cars in a parking lot, (b) time taken to complete a task, (c) student ranking in class, (d) nationality.

Solution

- (a) **Quantitative, discrete** (counted).
- (b) **Quantitative, continuous** (measured).
- (c) **Qualitative, ordinal** (ordered rank).
- (d) **Qualitative, nominal** (no order).

2.2 Sources of Data

Primary data	Secondary data
Collected first-hand by the investigator for a specific purpose	Already collected by someone else and available in published/unpublished form
Original, more reliable and up to date for the problem	May be less suitable, may be outdated or defined differently
Costly, time-consuming, needs organization	Cheap, quick and easy to obtain
Examples: own survey, experiment, interview	Examples: census reports, CBS publications, journals, websites, company records

3. Methods and Problems of Data Collection

3.1 Collection of Primary Data

Method	Description	Merits	Demerits
Direct personal investigation	Investigator meets respondents in person	Accurate, original, flexible	Costly, limited area, personal bias
Indirect oral interview	Ask third parties who know the facts	Useful when direct contact is difficult	Answers may be biased
Questionnaire (mailed / online)	Written questions sent to respondents	Cheap, wide coverage	Low response, no clarification
Schedule (enumerators)	Trained enumerators fill in forms by asking	High response, accurate	Costly, needs training
Observation	Watching and recording behaviour	Direct, no respondent bias	Cannot capture opinions
Experiment / telephone / web survey	Controlled trials, phone or online forms	Fast, cause-effect study possible	Limited coverage, technical bias

3.2 Collection of Secondary Data

Secondary data come from **published sources** such as government reports (Central Bureau of Statistics, Nepal Rastra Bank), international agencies (UN, World Bank, IMF), journals, newspapers, research reports and trade associations, and from **unpublished sources** such as office files, diaries and internal records. Before use, the data must be scrutinized for **reliability** (who collected it and how), **suitability** (definitions and units match the need) and **adequacy** (coverage and time period).

3.3 Problems in Collection of Data

- **Non-response** — respondents do not answer or cannot be found.
- **Response and memory bias** — false, forgotten or exaggerated answers; sensitive questions.
- **Faulty questionnaire** — ambiguous, leading or long questions.
- **Sampling frame problems** — incomplete or outdated lists of units.
- **Enumerator bias and errors** — untrained or dishonest data collectors, recording mistakes.
- **Cost, time and administrative difficulty** — funding, geography, lack of skilled staff.
- **Ethical and privacy issues** — confidentiality, consent, data protection.
- **Inconsistent definitions** of terms and units in secondary data, missing or outdated data.

Note: A pilot survey (pre-test) on a small group helps to detect problems in the questionnaire before the main survey.

Key Takeaways

- ✓ Statistics is both numerical data and the science of collecting, analysing and interpreting them; it is descriptive or inferential.
- ✓ Its scope covers business, government, IT, science and social research, but it cannot study individual cases and can be misused.
- ✓ Data may be qualitative (nominal, ordinal) or quantitative (discrete, continuous) and come from primary or secondary sources.
- ✓ Primary data are collected by interviews, questionnaires, schedules, observation and experiments; secondary data must be checked for reliability and suitability.
- ✓ Non-response, bias, poor questionnaires and cost are the usual collection problems.

— *End of Unit 1* —