STATISTICAL LABORATORY

Lecture 1: Fundamental Concepts of Statistics

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

allenlui@iseg.ulisboa.pt

https://sites.google.com/view/allenlui

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Course Overview & Program

Contact Information

Instructor: Yiu Lim LUI, Allen
E-mail: allenlui@iseg.ulisboa.pt

Course Details

- **Degree:** Applied Mathematics for Economics and Management
- **Timing:** 1st Year / 1st Semester (2026/2027)
- **Software Package:** Excel

Program Structure

1. Fundamental Concepts of Statistics
2. Exploratory Data Analysis
3. Organizing and Summarizing Data
4. Association and Relationships Between Variables
5. Index Numbers
6. Time Series Analysis

Bibliography & References

Main References

- Triola, M. F. (2022). *Elementary Statistics* (14th ed.). Pearson.
- Silvestre, A. L. (2007). *Anlise de Dados: Estatstica Descritiva*. Escolar Editora.

Complementary References

- Diez, D., Barr, C., & Cetinkaya-Rundel, M. (2019). *OpenIntro Statistics* (4th ed.).
- Flury, B., & Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Chapman & Hall.
- Newbold, P., Carlson, W. L., & Thorne, B. M. (2013). *Statistics for Business and Economics* (8th ed.). Pearson.
- Winston, W. L. (2016). *Microsoft Excel 2016 Data Analysis and Business Modeling*. Microsoft Press.

LECTURE 1: FUNDAMENTAL CONCEPTS OF STATISTICS

What is Statistics?

- **Definition:** Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting data to draw meaningful conclusions.
- **Core Purpose:** Transform raw numbers into actionable knowledge and insights.

Application Domains

- **Economics & Management:** Market trends, inflation, consumer behavior. Support decision-making under uncertainly, identify patterns, and predict trends.
- Healthcare & Medicine: Clinical trials, treatment effectiveness.
- Engineering & Tech: Quality control, algorithm training, risk modeling.

Meanings of "Statistics"

Scientific Discipline

The comprehensive theoretical and applied field studying data methods.

Measure

Numerical measures that describe sample characteristics (e.g., sample mean $\bar{x}$, standard deviation s).

Data Collections

Aggregated quantitative data in specific fields (e.g., employment statistics, health stats).

Steps of a Statistical Study

1. Problem Definition
- Define research questions clearly.

2. Measure Definition
- Identify variables and measurement scales.

3. Data Collection
- Select appropriate sampling or experimental design.

4. Processing & Summary
- Use tables, charts, and descriptive tools.

5. Statistical Inference
- Generalize sample results to population.

6. Study Report
- Report findings, implications, and limitations.

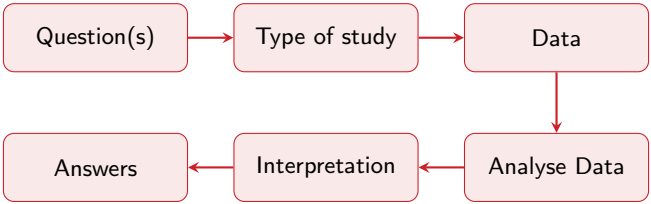
Purpose: Transform raw data into meaningful information.

Steps of a Statistical Study: Example

Suppose that we have a student call David.

- He wants to analyze campus student habits - specifically correlating study hours with GPA outcomes using statistical methods. (1 and 2)
- Sample: Random survey of 250 university peers. (3)
- Use tables and charts to visualize students' study hours and GPA. (4)
- Analysis: Hypothesis testing & regression models. (5)
- Outcome: Clear reporting on student performance trends. (6)

Steps of a Statistical Study: Process Flow



Core Objectives of Statistical Analysis

- **Data Summarization & Reduction:** Condensing massive datasets into meaningful tables, graphs, and descriptive metrics (mean, median, variance).
- **Generalization & Inference:** Making reliable predictions or estimations about an entire population based on sample observations.
- **Relationship Identification:** Uncovering associations, correlations, or potential causal mechanisms between variables.
- **Dimensionality Reduction:** Simplifying complex multi-variable datasets into fewer, highly informative components (e.g., Principal Component Analysis).
- **Classification & Discrimination:** Developing rules to assign entities into pre-defined categories based on measured traits.
- **Data Clustering:** Grouping unlabelled observations based on similarity metrics to reveal natural underlying structures.

Meaning of "Statistics"

Scientific Discipline

- Systematic study of data collection, analysis, and interpretation.

Measure

- Numerical summaries describing dataset characteristics.
- **Examples:** mean, variance, percentages.

Data

- Synonym for numerical information in specific areas.
- **Examples:** health, industrial, employment statistics.

Types of Statistics: Descriptive vs. Inferential

Descriptive Statistics

- Summarizes observed data.
- Uses tables, graphs, and calculations.
- Makes no conclusions beyond the available sample.
- *Example:* Average exam grade in Class A is 74%.

Inferential Statistics

- Extrapolates sample data to a broader population.
- Incorporates probability theory and uncertainty.
- Involves hypothesis testing and estimations.
- *Example:* Predicting election outcomes based on poll data.

IMPORTANCE OF SAMPLING STUDIES

Types of studies: census vs sampling (advantages and disadvantages)

Sampling

- More efficient, less costly, and less time-consuming than surveying the entire population.

Census

- Covers entire population, but is expensive, time-consuming, and often impractical.

Example

- Surveying 1,000 households from a city (sampling) vs. All households in the entire country (census).

Population vs. Sample

Population (Universe)

The complete collection of all individuals, items, or measurements whose characteristics are being studied.

- Finite (e.g., current ISEG students) vs. Infinite (e.g., future coin tosses).
- Real vs. Theoretical populations.

Sample

A selected subset of the population used to collect data and make inferences about the whole population. (Example: all customers of a company (population) vs 50 selected customers (sample))

Parameters vs. Statistics

Concept	Population Parameter	Sample Statistic
Definition	Fixed numerical value of population	Calculated value from sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	P or p	$\hat{p}$
Size	N	n

"Parameters are generally unknown fixed constants; statistics are random variables whose values vary from sample to sample."

Random Experiments and Sample Space

Random Experiment

A process or trial with uncertain outcomes, even when repeated under identical conditions.

Key Terminology

- **Sample Space (Ω or S):** The set of all possible outcomes.
- **Event:** A specific subset of outcomes within the sample space.

Examples

- Rolling a six-sided die: $S = \{1, 2, 3, 4, 5, 6\}$.
- Patient treatment outcome: $S = \{\text{Improved}, \text{Unchanged}, \text{Deteriorated}\}$.

Why Sample? Census vs. Sampling

Census

A full survey where data is collected from every single unit in the target population.

- **Drawbacks:** High financial cost, time-consuming, practically impossible for fluid or infinite populations, non-response difficulties.

Sampling

Studying a carefully chosen representative portion of the population.

- **Advantages:** Cost-efficient, faster output, allows deeper measurement rigor per unit.

Intuition: The Soup Analogy

Exploratory Taste

Tasting a small spoon of soup while cooking is taking a sample. Deciding if that spoon needs salt is an **exploratory analysis**.

Inference to Pot

Concluding the entire pot needs salt is an **inference**. This inference is valid only if the pot was well-stirred before taking the spoon!

Key takeaway: A huge spoon from an unstirred pot yields biased results. A small spoon from a well-stirred pot gives an accurate measurement.

Sampling Bias: Voluntary Response & Non-Response

Voluntary Response Bias

Occurs when sample inclusion depends on individual self-selection.

- Sample disproportionately attracts people with strong positive or negative feelings.
- *Example:* Online news quick polls, open rating forums.

Non-Response Bias

Occurs when selected non-respondents systematically differ in key traits from those who fill out the survey.

Convenience Sampling Hazards

Convenience Sample

Selecting individuals who are easily accessible to the researcher.

Why it Fails

- Easily reachable people rarely reflect the general population profile.
- Results cannot be generalized statistically beyond the sample group.
- *Example:* Surveying only student friends in the university cafeteria to study national income dynamics.

Historical Case Study: 1992 UK General Election

The Event

Major polling agencies (Gallup, MORI, NOP) conducted extensive pre-election polls predicting a comfortable victory or hung parliament for Labour Party over Conservatives.

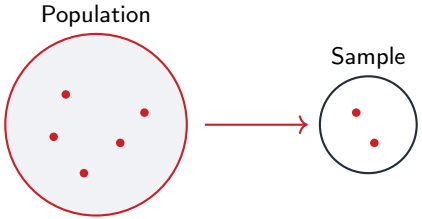
What Went Wrong?

- **Non-Response Bias ("Shy Tory Factor"):** Conservative voters were systematically less likely to participate in door-to-door or phone surveys.
- **Quota Sampling Flaws:** Pollsters used quota sampling (matching age/gender profiles) based on outdated decade-old census data, missing key demographic shifts.
- **Outcome:** The Conservatives won a comfortable majority-a 8.5% discrepancy that forced global polling methods to adopt random probability sampling and weighting

Probability Sampling: Random Sampling

Random Sampling

Each individual unit in the population has an equal probability of selection, and every possible sample of size n has an equal chance of being chosen.



- Eliminates selection bias.
- Serves as the mathematical foundation for standard statistical theory.

Random Variables vs. Empirical Variables

Random Variable (X)

A theoretical mathematical rule/function that assigns real numerical values to outcomes of a random experiment.

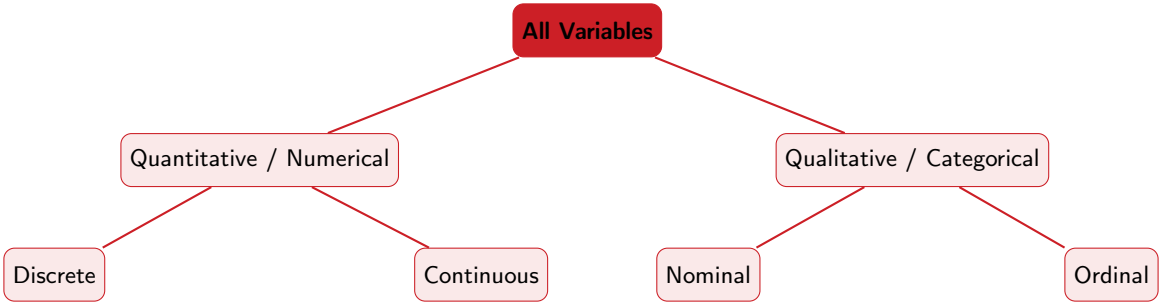
- Denoted by capital letters (X, Y, Z).
- Subject to probabilistic distributions.
- Examples: weekly sales and number of defective items.

Empirical Variable (x)

Actual observed quantitative data points collected in a study.

- Denoted by lowercase letters ($x_1, x_2, \ldots, x_n$).
- Realized numeric values.
- Examples: age, weight, and number of products sold.

Classification of Variables: Visual Structure



Numerical Variables: Discrete vs. Continuous

Discrete Numerical Variables

- Result from counting processes.
- Takes isolated, distinct values (often integers) with jumps in between.
- *Examples:* Number of bank transactions, daily website hits, family size.

Continuous Numerical Variables

- Result from measurement processes.
- Can take any real number value within a given continuum/interval.
- *Examples:* Portfolio return (%), time spent on a survey, distance, temperature.

Explanatory Variable and Explained Variable

Explanatory Variable (Independent / Predictor Variable)

- A variable used to explain or predict changes in another variable.
- Represents the potential cause, influence, or input.
- Usually placed on the **x-axis** in graphs.
- **Examples:** marketing expenditure, study hours, and product price.

Explained Variable (Dependent / Response Variable)

- The variable whose variation we want to understand or predict.
- Represents the effect, outcome, or result.
- Usually placed on the **y-axis** in graphs.
- **Examples:** sales revenue, exam score, and demand for a product.

Categorical Variables: Nominal vs. Ordinal

Nominal Variables

- Data categorized into mutually exclusive groups with no natural or logical ordering.
- Mathematical operations $(+, -, >)$ are meaningless.
- *Examples:* Eye color, marital status, nationality, payment choice.

Ordinal Variables

- Categories have a distinct natural ranking or hierarchy.
- Exact differences between ranks are not fixed or quantitatively measurable.
- *Examples:* Education level, customer satisfaction rankings (1 to 5), credit rating.

Levels of Measurement Summary

Nominal Scale

Categories only (e.g., Gender, Country).

Ordinal Scale

Categories with order (e.g., Likert scale, Class rank).

Interval Scale

Ordered numbers, equal intervals, no absolute zero (e.g. 0 °C $\neq$ no temperature).

Ratio Scale

Ordered numbers, equal intervals, true absolute zero (e.g. 0 Income = no income).

Special Variable Nuances: Telephone Codes & Dread Scale

Case 1: Telephone Country Code

- **Question:** Is a phone country code numerical or categorical?
- **Analysis:** Though written as digits (e.g., 351, 34), calculating an average area code is meaningless.
- **Classification:** Categorical Nominal.

Case 2: Customer Satisfaction Scale (1 to 5)

- **Classification:** Categorical Ordinal (e.g., 1 = Very Dissatisfied, 5 = Very Satisfied).
- **Practical Application:** Although the gaps between choices aren't strictly equal, social scientists and managers often treat Likert scales as numerical to compute sample means for survey analysis.

Icebreaker Exercise: Analyzing Pokemon Data

The Pokedex Dataset

Imagine collecting data on 100 Pokemon. Classify each attribute:

1. **Type:** Electric, Fire, Water...
2. **Evolution Stage:** Basic, Stage 1...
3. **Base HP:** e.g., 35, 60, 250
4. **Height (m):** e.g., 0.4 m, 1.7 m
5. **Pokedex #:** e.g., #0025 (Pikachu)

Variable Types

- **Type:** Categorical, Nominal
- **Stage:** Categorical, Ordinal
- **HP:** Numerical, Discrete
- **Height:** Numerical, Continuous
- **#:** Categorical, Nominal *(It's an ID!)*

Formal Representation of Data Matrix

- n **observations of 1 variable:**

$$(x_1, x_2, x_3, \ldots, x_n)$$

- n **observations of 2 variables:**

$$[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$$

- n **observations of** p **variables (Data Matrix):**

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Data Collection

- **Sources:** surveys, interviews, administrative databases, sensors, and company records.
- **Processes:** coding, recording, and validating data.
- **Data Coding:** transforming answers into numerical or categorical codes.
  Data Recording: storing information systematically (Excel, SPSS, SQL).
  Data Validation: checking consistency, completeness, and accuracy before analysis.

Types of Data: Cross-Sectional, Time Series, Panel

1. Cross-Sectional Data

Observations collected for multiple units at a single point in time.

Ex: GDP of 30 EU countries in 2024.

2. Time Series Data

Observations collected for a single unit across equal time intervals.

Ex: Monthly sales revenue over 5 years.

3. Panel Data

(Longitudinal) Combines cross-sectional and time series: tracking multiple units over time.

Ex: Annual income of 500 households (2018–2024).

Exercise 1.1: Mortgage Dataset

A mortgage company randomly samples accounts of their time-share customers. State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?

a. The original purchase price of a customer's time-share unit
b. The state (or country) of residence of a time-share owner
c. OA time-share owner's satisfaction level with the maintenance of the unit purchased (1: very dissatisfied to 5: very satisfied)
d. Number of times customer payment was late.

Source: Newbold et al. (2013)

Exercise 1.1: Solutions

Solutions

a. **Numerical, Continuous (Ratio):** Monetary amounts take real values on a continuum with absolute zero.

b. **Categorical, Nominal:** Qualitative geographic names without intrinsic ordering.

c. **Categorical, Ordinal:** Qualitative scale with clear natural rank ordering.

d. **Numerical, Discrete (Ratio):** Integer count value starting at zero.

Exercise 1.2: Supermarket Survey

Visitors to a supermarket in Singapore were asked to complete a customer service survey. Are the answers to the following survey questions categorical or numerical? If an answer is categorical, give the level of measurement. If an answer is numerical, is it discrete or continuous?

a. Have you visited this store location before?
b. Rate customer service level today from 1 (Very Poor) to 5 (Very Good).
c. How much money did you spend in store today?

Source: Newbold et al. (2013)

Exercise 1.2: Solutions

Solutions

a. **Categorical, Nominal:** Yes/No, no natural order.

b. **Categorical, Ordinal:** Rating scale 1-5, ordered categories.

c. **Numerical, Continuous (Ratio):** Money spent, decimal values possible, true zero.

Exercise 1.5: London Tourism Survey

A number of questions were posed to a random sample of visitors to a London tourist information center. For each question below, describe the type of data obtained.

a. Are you staying overnight in London?

b. How many times have you visited London previously?

c. Which London attractions have you visited today?
   - Tower of London
   - Buckingham Palace
   - Big Ben
   - Covent Garden
   - Westminster Abbev

d. How likely are you to return within 12 months (Unlikely, Likely, Very Likely)?

Source: Newbold et al. (2013)

Exercise 1.5: Solutions

Solutions

a. **Categorical, Nominal:** Yes/No, no inherent order.

b. **Numerical, Discrete:** Count of visits, zero possible, only integer values.

c. **Categorical, Nominal (Multiple Response):** Each attraction is a yes/no question; categories with no order.

d. **Categorical, Ordinal:** Likert-type scale (e.g., very unlikely $\rightarrow$ very likely), ordered categories.

What is an empirical study?

1. Based on direct observation or data collection.
2. Evidence-based: uses real data from experiments, questionnaires, interviews, or observation.
3. Analytical purpose: identify patterns, relationships, or effects.
4. Systematic approach: rigorous collection, processing, and interpretation.
5. Replicable: methods can be repeated by other researchers.
6. Example: Measuring teaching methods/ impact on student performance.

Empirical Research in Economics and Management

- Data sources: surveys, experiments, administrative records, and financial databases.
- Importance:
  - Understanding market behavior
  - Supporting decision-making
  - Evaluating policies
  - Forecasting trends

Empirical Work Procedure

Steps:

1. Define research question.
2. Collect relevant data.
3. Organize and process data.
4. Apply descriptive and inferential statistical methods.
5. Interpret and report results.

Sources of Statistical Information

Information Hierarchy

- **Primary Data:** Collected directly by researchers for the specific project (e.g., surveys, lab experiments).
- **Secondary Data:** Pre-existing data collected by external entities (e.g., national statistics offices, corporate financial filings).
- **Tertiary Data:** Synthesized summaries, reference encyclopedias, and compilations.

Data Processing Steps

Data Coding $\longrightarrow$ Systematic Recording $\longrightarrow$ Validation & Cleaning

Databases in Economics & Management

Socio-Economic Databases

- **Eurostat:** Official European Union statistics.
- **World Bank Open Data:** Global development metrics.
- **INE (Portugal):** National statistical datasets.
- **OECD Data:** Economic and structural policies.

Financial & Corporate Databases

- **Bloomberg / Refinitiv:** Real-time financial market data.
- **S&P Capital IQ / Orbis:** Corporate reporting & accounts.
- **Yahoo Finance:** Market indices and stock tickers.

Lecture 1 Summary

Key Concepts Covered

- Definitions of Statistics, Population, Sample, Parameter, Statistic.
- Classification of variables (Numerical vs Categorical; Discrete, Continuous, Nominal, Ordinal).
- Random sampling sampling biases.
- Observational Studies vs. Experimental Design Principles.

Next Lecture: Exploratory Data Analysis & Graphical Representations.

THANKS!

Questions & Discussion

allenlui@iseg.ulisboa.pt