# Introduction to Statistics - Young Researchers Fellowship Program

## Lecture 1 - Introduction to Statistics & Tabular Data Logic

Daniel Sánchez Pazmiño

Laboratorio de Investigación para el Desarrollo del Ecuador

September 2024

# What is statistics?

# What is statistics?

- A **methodology** for collecting, analyzing, interpreting, and presenting numerical information.
- A statistic is often referred to as a **numerical fact** or a **piece of data** which describes a particular characteristic of a group of individuals.
  - In the field of statistics, we typically don't refer to individual data points as statistics.
- In several fields, statistics is used as an aid to decision making under uncertainty.
- In a research context, statistics will be needed to understand phenomena, make predictions, and test hypotheses emerging from theory.
- *Statistics is the systematic investigation of the correspondence of theory with the real world*

# Data in statistics

# Data in statistics

- Because statistics is concerned with information, **data** is often the starting point of any statistical analysis.
- No clear definition of data can possibly satisfy everyone, but we can think of data as a collection of **facts** to be analyzed.
  - Data is **plural** for **datum**.
- A **dataset** is a collection of data points, which can be organized in a **table**, often about a specific topic, purpose, experiment, study, or context.

# Broad types of data

- Typically, statistics makes a distinction between two broad types of data:

1. **Quantitative** data, which is numerical in nature, meaning it can be measured and expressed in numbers.
   - Discrete data: whole numbers (e.g., number of students in a class).
   - Continuous data: real numbers (e.g., height, weight).
2. **Categorical** or "qualitative" data, which is non-numerical in nature, meaning it cannot be directly measured or expressed in numbers.
   - Nominal data: categories without order (e.g., colors).
   - Ordinal data: categories with order (e.g., levels of satisfaction).

# How a dataset might look like

Table 1: Example dataset from the Ecuadorian Superintendencia de Compañías, Valores y Seguros (SUPERCIAS) companies' directory.

| RUC | Fecha Constitucion | Provincia | Capital Suscrito |
|---|---|---|---|
| 0993266272001 | 25/06/2020 | GUAYAS | 800 |
| 1793131182001 | 19/03/2021 | PICHINCHA | 1000 |
| 0993381738001 | 03/05/2023 | GUAYAS | 500 |
| 0993374043001 | 28/07/2022 | GUAYAS | 800 |
| 1792649765001 | 11/01/2016 | PICHINCHA | 10800 |

- What type of data can be identified for each column in the dataset?

# Types of datasets

- Datasets can be classified into different types based on the nature of the data they contain.

1. **Cross-sectional data**: data collected at a single point in time.
   - Example: a survey conducted in 2024.
2. **Time series data**: data collected over time at regular intervals, for a single entity.a
   - Example: monthly sales data from 2020 to 2024.
3. **Panel data**: data collected over time for multiple entities.
   - Example: monthly sales data from 2020 to 2024 for multiple companies.

- Often, we also hear about repeated or pooled cross-sectional data, which is a combination of cross-sectional and time series data; we observe multiple cross-sections at different points in time.

# Tabular data logic

# Tabular data logic

- A dataset is typically organized in a **table** with rows and columns.
- Datasets often collect characteristics of individuals or entities, which are typically referred to as **elements**.
  - Elements are not necessarily the observations in a dataset, elements are those entities or individuals for which we hold information.
- When data is *tidy*, a table structure typically allows for easy identification of the following elements:
  - We will talk more about **tidy data** concept in the R companion module.

1. **Variables**: columns in the table, which represent a characteristic of an element.
2. **Observations**: rows in the table, which represent a collection of variable values for a single element.

# Elements are not observations

- Sometimes, elements in a dataset (for the SUPERCIAS dataset, companies) are not the same as observations.
- We may observe multiple observations for a single element, which is why we need to be clear about the distinction between elements and observations.
- When we observe multiple observations for a single element, we typically refer to this as **repeated measures** or **panel data**.
- It is in this context when it comes in handy to difference between long and wide format datasets.

# Long vs. wide format

- Long vs. wide format refers to the way data is organized in a table.
- In the **long format**, each row represents a single observation, and each column represents a variable.
- In the **wide format**, each row typically represents a single element. Columns may represent variables, but also repeated measures or time points of the same variable.

# Example of long vs. wide format - business creation per province

- Long format: each row represents a single observation (business creation per province per year).

Table 2: Long format business creation per province (SUPERCIAS)

| Province | Year | Number of businesses created |
|---|---|---|
| ORELLANA | 2013 | 82 |
| AZUAY | 2023 | 1267 |
| IMBABURA | 2012 | 97 |
| COTOPAXI | 2024 | 174 |
| EL ORO | 2010 | 176 |

- Notice that if a province creates businesses in multiple years, we will have multiple rows for the same province.

# Example of long vs. wide format - business creation per province

- Wide format: each row represents a single element (province), and columns represent variables (business creation per year).

Table 3: Wide format business creation per province (SUPERCIAS)

| Province | 2013 | 2023 | 2012 | 2024 | 2010 |
|---|---|---|---|---|---|
| ORELLANA | 82 | NA | NA | NA | NA |
| AZUAY | NA | 1267 | NA | NA | NA |
| IMBABURA | NA | NA | 97 | NA | NA |
| COTOPAXI | NA | NA | NA | 174 | NA |
| EL ORO | NA | NA | NA | NA | 176 |

- We will never have multiple rows for the same province in the wide format.

# Sources of data

# Where do we get data for statistical analysis?

1. Experimental studies
2. Observational studies
3. Secondary sources (existing datasets)