Introduction to Statistical Methods in Political Science

Lecture 1: Introduction to Data and Statistics

Ignacio Urbina

Definitions

The Study of Statistics: Some Definitions

What is statistics?

- Statistics is the study of collecting, analyzing, interpreting, and presenting data.
- Involves mathematical techniques to make inferences about a population from a sample.

Why is it important?

- Uncertainty is inherent in real-world phenomena. Statistics provides tools to manage and quantify uncertainty.
- Critical for data-driven decision-making across disciplines.

Population and Sample

Population:

- A population is the complete set of all possible observations or measurements of interest.
  - Example: All residents in a country when studying public health policies.

Sample:

- A sample is a subset of the population selected for analysis.
  - Example: A group of 1,000 residents surveyed to infer public opinion on policy.

Estimator

- An estimator is a statistical method used to estimate a population parameter based on sample data.
- A more abstract definition: *An estimator is any mathematical formula (function) computed using measures collected from a sample*.
  - Example: The sample mean used as an estimator of the population mean.

Statistical Inference

- Statistical inference is the process of drawing conclusions about a population based on sample data.
  - Example: Inferring the likely values of the average income of a population based on a sample.

Dataset

- A dataset is a structured collection of data, typically organized in a tabular format.
- Rows represent individual observations, and columns represent variables.
- Example: A dataset of economic indicators across countries.

Observations

- An observation is a single data point or record in a dataset.
- Each row in a dataset typically corresponds to one observation.
- Observations are instances of measurements or responses.

Variables

- A variable is a characteristic or attribute that can take on different values.
- Each column in a dataset typically represents a variable.
- Variables are classified based on the type of data they represent.

Example Dataset

- **Sample Dataset:**

Respondent	**Age**	**Income**	X_1	X_2	**State**	**Education**
1	34	55000	1	7	NY	High school
2	29	60000	0	1	NJ	Master's
3	45	70000	1	3	MA	High school
4	40	65000	0	4	MA	High school
5	38	62000	1	2	NY	Bachelor's
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Types of Variables

- Variables can be classified as **categorical** or **continuous**.
- Understanding the type of variable is essential for choosing the correct statistical method.
- The type of variable determines the appropriate summary and analysis techniques.

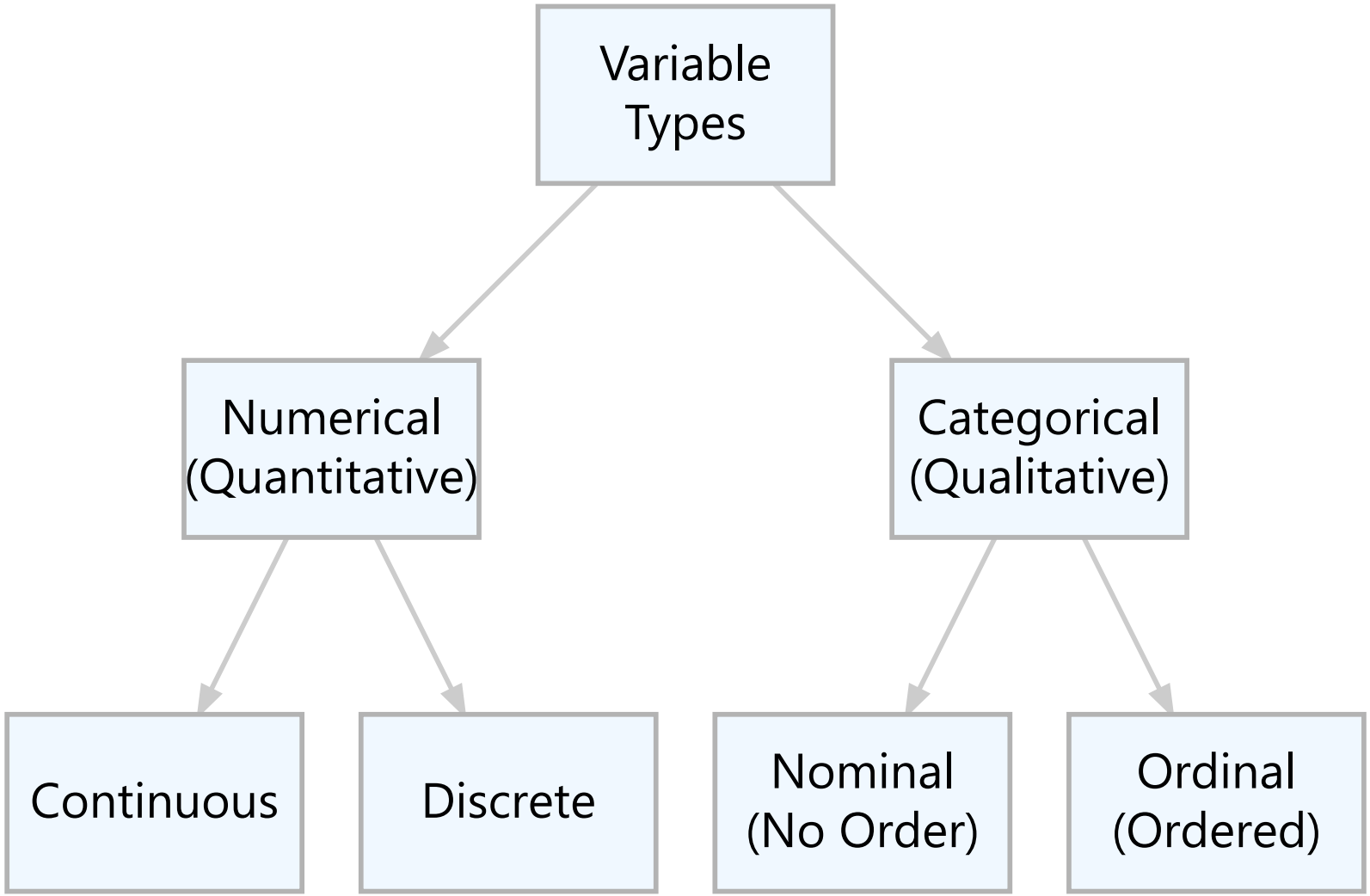
Categorical (Qualitative) Variables

- Categorical variables represent distinct categories or groups.
- Can be nominal (no order) or ordinal (ordered categories).
- Require specific transformations before statistical analysis (i.e., we need to 'code' them into numbers before analysis).
- Examples:
  - **Nominal**: No natural ordering among the categories. Example: Blood type (A, B, AB, O).
  - **Ordinal**: Categories have a natural order. Example: Educational level (High school, Bachelor's, Master's, PhD).

Numerical (Quantitative) Variables

- Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much.'
- Numeric variables can be continuous or discrete.
- Examples:
  - **Continuous**: Can take any value within a range, including fractions and decimals. Example: Height, weight.
  - **Discrete**: Can only take non-negative whole numbers. Example: Number of children, number of cars owned.

Summary: Types of Variables



Example: Types of Variables

Can you correctly identify the types of variables included in this dataset?

- **Sample Dataset:**

Respondent	**Age**	**Income**	X_1	X_2	**State**	**Education**
1	34	55000	1	7	NY	High school
2	29	60000	0	1	NJ	Master's
3	45	70000	1	3	MA	High school
4	40	65000	0	4	MA	High school
5	38	62000	1	2	NY	Bachelor's
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Statistical Principles Involved in Research Studies

Defining an Empiricist Research Question

Research Question:

- A question that seeks to explore observable and measurable phenomena.
- Must be testable through empirical data and direct observation.
- Aims to contribute to existing knowledge, test theories, or solve practical problems.

Example:

- **Poorly Defined:** "Does exercise affect health?"
- **Well-Defined:** "What is the impact of a 30-minute daily aerobic exercise regimen on the cardiovascular health markers (e.g., blood pressure, cholesterol levels) of adults aged 30-50 over six months?"

Components of a Well Defined Research Question

- **Specificity:** Clearly defines variables and their relationships or differences, ensuring focus and clarity.
- **Operationalization:** Concepts are defined in measurable terms, facilitating precise data collection.
- **Feasibility:** The question is practical to investigate considering available resources, time, and ethical standards.
- **Relevance:** Addresses significant issues or gaps within the field.
- **Novelty:** Contributes new insights or perspectives to existing research.

Example: Population and Research Question

Research Question: What is the impact of a universal basic income (UBI) on economic stability and poverty reduction, specifically among low-income households?

Population of Interest: Low-income households across the entire country, with consideration of regional, demographic, and economic diversity.

Associated and Independent Variables

- **Associated Variables**: Variables that show some relationship with each other. The association can be **positive** (both increase and decrease together), **negative** (one increases while the other decreases), or **non-linear**.
- **Independent Variables**: Variables that are not associated and do not influence each other.

Example

What is the association between the total number of community activities in a neighborhood and the level of social trust among its residents?

Explanatory and Response Variables

- **Explanatory Variable**: The variable suspected of affecting the other. Often considered the cause.
- **Response (Outcome) Variable**: The outcome or effect being measured.
- **Note**: *Association between variables does not imply causality*, even if an explanatory-response relationship is identified.

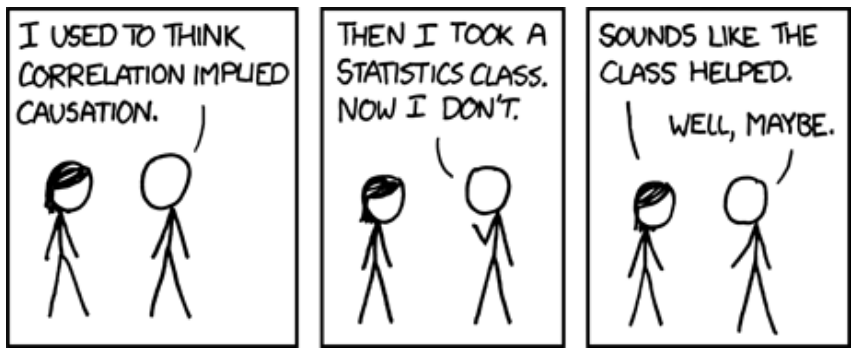


Figure: Correlation (Source: XKCD)

Observational vs. Experimental Studies

Aspect	Observational Study	Experimental Study
Definition	Data collected without manipulating variables	Involves manipulating one or more variables
Purpose	Identify associations between variables	Establish cause and effect relationships
Causation	Cannot determine causality	Can determine causality with proper design

Confounding Variables and Statistical Bias

Confounding Variables

Definition: Unaccounted variables that influence both the explanatory and response (outcome) variables.

Impact: Without controlling for confounders, researchers may **misattribute** the association between variables, incorrectly inferring a direct relationship.

Statistical Bias

Definition: Systematic errors in data collection, analysis, interpretation, or review processes that can **distort** findings. **Example:**

Confounding Bias is a distortion that alters the true association between an explanatory and an outcome variable due to a third factor that is independently associated with the exposure and the outcome (Source).

Example: Assessing the Impact of Math Tutoring

Research Context:
A high school is implementing a new structured math tutoring program that provides students with an additional 2 hours of tutoring per week. Participation is voluntary. The goal is to determine whether this program positively affects students' performance in standardized math tests over the course of one academic year.

Task:
Identify and describe **one variable** that could cause **confounding bias** when assessing the association between **participation in the math tutoring program (X)** and **standardized math test scores (Y)**.

Experiments in Social Science

Context:
Social science experiments help us understand cause-and-effect relationships by manipulating a specific variable (the *treatment*) under controlled conditions.

Experimental Treatment (Definition):
The specific intervention or condition applied exclusively to the **treatment group**. This allows researchers to measure its impact on the response (outcome) variable compared to the **control group** (untreated).

Principles of Experimental Design

- **Control**: Accounting for confounding factors to ensure observed effects are due to the treatment.
- **Randomization**: Randomly assigning subjects to treatment and control groups to eliminate selection bias.
- **Replication**: Repeating the experiment or using a sufficiently large sample to ensure the results are reliable.
- **Blocking**: Grouping subjects with similar characteristics to reduce variability and better isolate the treatment effect.

Blinding in Studies

- **Single Blinding**: Participants are unaware of whether they are in the treatment or control group.
- **Double Blinding**: Both participants and experimenters are unaware of the group assignments.
- Blinding helps reduce bias, particularly the Performance and Experimenter bias.
  - E.g., in a clinical study, if participants in the control group systematically seek other treatments, there could be performance bias.
  - Or, if researchers/clinicians treat participants differently depending on which group they are in, this could imply experimenter bias (see: Experimenter Effect).
  - **Placebo Effect**: Change in participants' outcome variable due to their belief in the treatment rather than the treatment itself.

Sampling: Basic Principles

Sampling Methods and Sampling Bias

Sampling Method (Definition):
A sampling method refers to the process used to select individuals or observations from a population to be included in a sample.

Sampling Bias (General):

- Sampling bias occurs when the process of selecting a sample **skews the results**, making some members of the population systematically more (or less) likely to be included than others.

Sampling Methods and Sampling Bias

Types of Sampling Bias:

- **Non-response Bias:** Arises when individuals selected for the sample do not respond or are unwilling to participate, and those non-responders **differ in important ways** from those who do respond.
- **Self-Selection Bias:** Occurs when participation in a survey, study, or experiment is **voluntary**, allowing individuals to decide on their own whether to be included. As a result, those who opt-in may **systematically differ** from those who do not, leading to a non-representative sample.
- **Coverage Bias:** Happens when some members of the population are inadequately represented in the sample due to limitations in the **sampling frame**. This occurs when certain groups have **no chance** or a **lower chance** of being included, leading to a distorted view of the population.
  - **Sampling Frame:** The list or method used to identify and select individuals from the population for inclusion in the sample.

Probabilistic Sampling

- **Probabilistic Sampling (def.):** A method where each member of the population has a known, non-zero chance of being selected in the sample. This approach ensures that the sample is more representative of the population.
  - **Examples:** Simple random sampling, stratified sampling, cluster sampling.
- **Advantages:** Reduces sampling bias, allows for generalization to the population, and facilitates the use of statistical inference.

Sampling Techniques

- **Simple Random Sampling**: Every subject has an equal probability of being selected.
- **Stratified Sampling**: Population is divided into strata; a random sample is drawn from each stratum.
- **Cluster Sampling**: Population is divided into clusters (i.e., naturally forming groups that are diverse within and similar between); a random sample of clusters is selected, and all subjects within those clusters are studied.
- **Multistage Sampling**: Corresponds to taking samples in stages using smaller and smaller sampling units at each stage.
  - For example, clusters are chosen randomly, and then a random sample from within each cluster is selected.

Stratified vs. Cluster Sampling

Stratified Sampling

- **Objective:** Ensure representation across distinct subgroups
- **When to Use:**
  - Population is heterogeneous and can be divided into meaningful strata
  - Detailed analysis is required within each subgroup
  - High precision in estimates across strata
- **Example:** Survey different income brackets to study economic disparities

Cluster Sampling

- **Objective:** Reduce cost and logistical complexity for large, dispersed populations
- **When to Use:**
  - Population naturally forms clusters
  - Easier to collect data from entire clusters than from scattered individuals
  - Ideally, each cluster is diverse internally (a "mini-population")
- **Example:** Conduct a health survey by sampling entire schools as clusters

Convenience (Non-probabilistic) Sampling

- **Convenience (Non-probabilistic) Sampling (def.):** Individuals who are easiest to reach or most accessible are chosen for the sample, rather than randomly or systematically selected.
- **Potential Issues:**
  - **Bias:** May result in a sample that is not representative of the population. Findings may not be applicable to the broader population due to the non-random nature of sample selection.
  - **Examples:** Surveying people in a mall or using participants from a single location or organization.

Representativeness of a Sample

Definition: A sample is **representative** if its key characteristics (e.g., demographics, behaviors) closely match those of the overall population, allowing for valid generalizations.

Discussion:

- *Survey Vendors*: Many commercial survey vendors sell **representative samples** based on a few demographic features but not true probabilistic samples, meaning results may still be affected by sampling bias.
  - The most common form of representative sample is quota sampling.
- Even a probabilistic sample can be **non-representative** if it is too small or improperly stratified.

Terms and Concepts Covered

Terms and Concepts Covered in this Lecture

- Statistics
- Population
- Sample
- Estimator
- Statistical Inference
- Dataset, Observations, & Variables
- Categorical/Continuous Variables
- Response/Explanatory Variable
- Observational/Experimental Study
- Confounding Variables
- Bias
- Sampling Techniques
- Voluntary Response
- Simple Random Sampling / Stratified / Cluster / Multistage Sampling
- Research Question
- Population of Interest
- Associated/Independent Variables
- Blinding
- Control
- Randomization
- Replication
- Blocking
- Sampling Bias
- Convenience Sampling Bias