POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti
(Based on DSS Materials)

Lecture 03 | Observations, Variables, and Means

Plan for Today

- What are data/datasets?
    - what is an observation?
    - what is a variable?

- Types of variables based on content
    - character vs. numeric variables
    - binary vs. non-binary variables

- Average or mean of a variable
    - how to compute it?
    - how to interpret it?

Before we start

Announcements:

- I hope you had a great break! Next one is President's Day.
- Quizzes and Participation:
  - Start at week 03. I will give full marks on Quiz 1 for everyone on week 03. You're welcome :)
- How was your Lab last week?
  - We have a great line up of TAs in this class.
- Github page: https://github.com/umbertomig/POLI30Dpublic

Before we start

Recap:

- We learned the definitions of Theory, Scientific Theory, and Hypotheses.

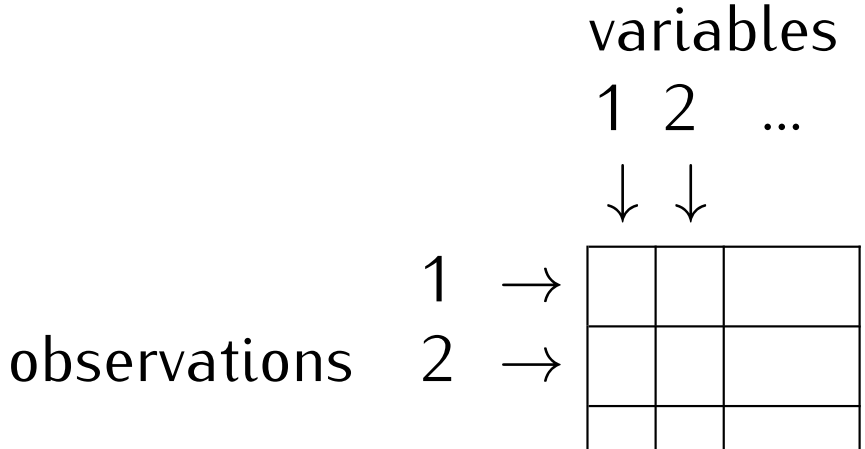
Great job!

- Do you have any questions about these contents?

Political Science Data

What are Data/Datasets?

- To test theories, we need data. What is data? What are datasets?
- Datasets capture the characteristics of a particular set of individuals or entities:
  - students, classrooms, schools, etc.
- Datasets are typically organized as **dataframes** where rows are observations and columns are variables



Example of a Dataframe

Each column is a variable

Each row is an observation (i=1,2,3,….10)

i	first_name	test_score
1	ana	80
2	elena	75
3	maria	99
4	juan	67
5	diego	89
6	carlos	80
7	olivia	70
8	jorge	86
9	adolfo	92
10	marta	83

What is an observation?

- It is the information collected from a particular entity or individual in the study
- The **Unit of observation** of the dataset defines the individuals or the entities that each observation in the dataset represents
  - if the Unit of observation is students, each row in the dataset represents a different student
- We usually refer to an observation by the row number in the dataset, which we denote as i
  - what is the first observation (i=1) in the dataframe above?

What is a variable?

- A variable contains the values of a changing characteristic for the various individuals or entities in the study
- Every column of data in a dataset is a variable
  - if the Unit of observation is students, each variable captures a specific characteristic of the students, for all the students in the study
- We usually refer to a variable by its name
  - *first_name*, *test_scores*

Notation

- When defining new variables, we represent a variable and its contents in the following format:

$$X = \{10, 5, 8\}$$

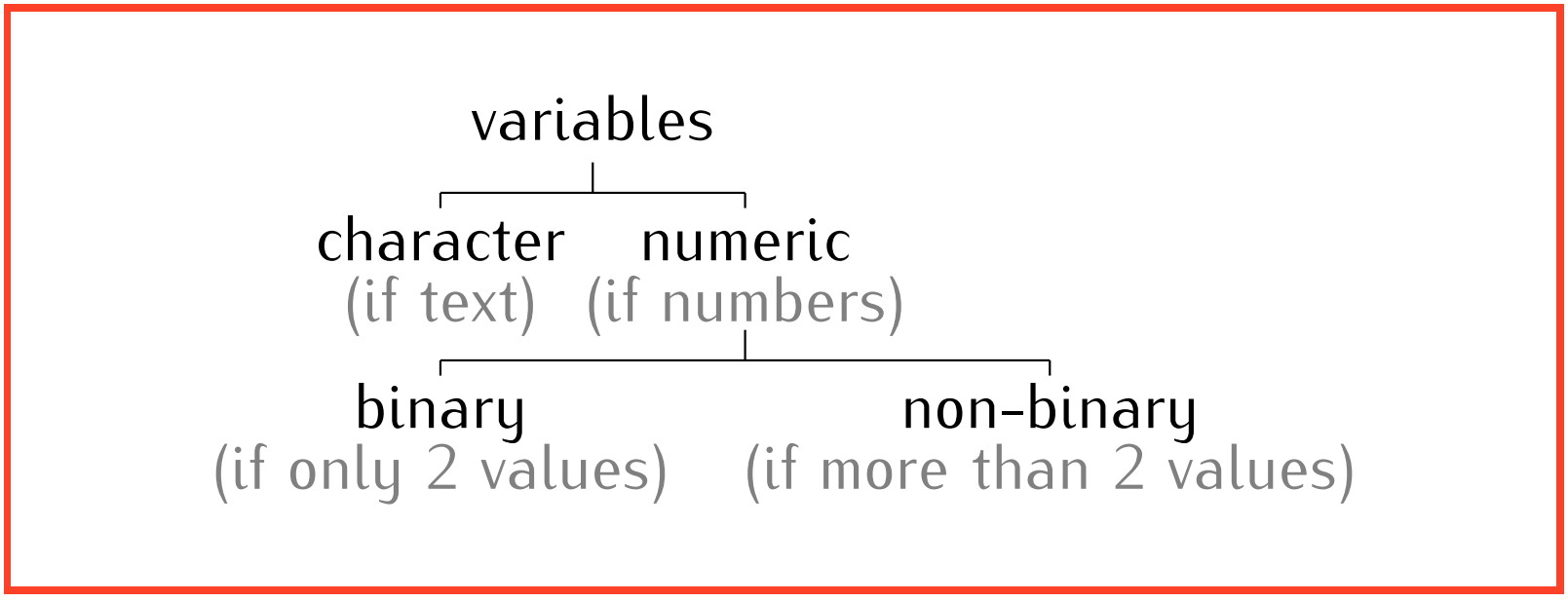
- On the left-hand side of the equal sign, we identify the name of the variable:
  - what is the name of the variable here?
- On the right-hand side of the equal sign and inside curly brackets, we have the content of the variable: multiple observations, separated by commas
  - what are the observations in X?

Notation

$$X = \{10, 5, 8\}$$

- To represent each observation we use X_i
  - where i stands for the observation number
  - the subscript i means that we have a different value of X for each value of i
  - what is X_3?
- The total number of observations is denoted as n
  - what does n equal to here?

Types of Variables Based on Content



Character vs. Numeric

- **Character variables** contain text
  - *first_names*={ana, elena, maria, ...}
- **Numeric variables** contain numbers
  - *test_score*= {80, 75, 99, ...}

Numeric: Binary variables (AKA dummy variables)

- **Binary variables** can take only two values: 1s and 0s
- They represent the presence/absence of a trait:
  - 1 if individual i has the trait
  - 0 if individual i does not have the trait
- Example: $voted = \{1, 0, 0, 1, 1, 1, 0\}$ where

$$voted_i = \begin{cases} 1 \text{ if individual } i \text{ voted} \\ 0 \text{ if individual } i \text{ didn't vote} \end{cases}$$

- Can you think of another example?

Numeric: Non-binary variables

- **Non-binary variables** can take more than two values
  - *distance*={1.452, 2.345, 0.298}
  - *dice_roll*={2, 4, 6}
- Can you think of another example?

Mean of a Variable

Average or Mean of a Variable: How to Compute it?

- Sum the values across all observations and divide the result by the total number of observations

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

- $\overline{X}$ (pronounced X-bar) stands for the average of X
- $\sum_{i=1}^{n} X_i$ stands for the sum of all X_i (observations of X) from i=1 to $i = n$, meaning from the first observation of the variable X to the last one ($\sum$ is Greek letter sigma)
- X_i stands for a particular observation of X, where i denotes the position of the observation and n is the total number of observations in the variable

- Example: if X={10, 4, 6, 8, 22}, then:
  - $n = ?$
  - $\overline{X} = ?$
  - Let's compute it!

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$$= \frac{10+4+6+8+22}{5} = \frac{50}{5} = 10$$

Average or Mean of a Variable: How to Interpret it?

- First, we need to figure out the quantity in which the value is measured
  - Whenever interpreting numeric results, you should make it clear whether the number is measured in points, percents, miles, kilometers, etc.
  - This is called the **unit of measurement**

Unit of Measurement of the Mean of a Variable

interpretation of the mean of a variable

if variable is non-binary:
as an average, in the same unit of measurement as the variable

if variable is binary:
as a proportion, in % after multipying the result by 100

- When the variable is **non-binary**, the mean should be interpreted as an average in the same unit of measurement as the values in the variable

- Example: if $X=\{10, 4, 6, 8, 22\}$ and measured in miles
  - $\overline{X} = ?$
  - what type of variable is X (binary or non-binary)?
  - shall we interpret $\overline{X}$ as an average or a proportion?
  - unit of measurement of $\overline{X} = 10$?

- When the variable is **binary**, the mean should be interpreted as a proportion, in % after multiplying the result by 100

- Why?
  - Because the mean of a binary variable is equivalent to the proportion of the observations that have the characteristic identified by the variable (i.e., that meet a criterion)

- Example: if X={1, 1, 1, 0, 0, 0}, then:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}$$
$$= \frac{1+1+1+0+0+0}{6} = \frac{3}{6} = 0.5$$

  - what type of variable is X (binary or non-binary)?
  - shall we interpret $\overline{X}$ as an average or a proportion?
  - interpretation of $\overline{X} = 0.5$ (including units)?
    - 50% of the observations are 1s, that is, have the characteristic identified by X (0.5x100=50%)
  - note that the fraction $\frac{3}{6}$ is equivalent to the proportion of the observations that are 1s

- The proportion of observations in a variable that meet a criterion is calculated as:

$$\frac{\text{number of observations that meet criterion}}{\text{total number of observations}}$$

- Example: if X={1, 1, 1, 0, 0, 0}, the proportion of observations in X that are 1s is:
  - $\frac{3}{6}$= 0.50
  - to interpret the result of this fraction as a percentage, we multiply the decimal by 100 (0.50x100=50%)
  - interpretation: 50% of the observations in X are 1s

Summary

- **Today's Class:**
  - Data/datasets
  - Observations and variables
  - Character vs. numeric variables
  - Binary vs. non-binary variables
  - Computing and interpreting means
- Next class:
  - Causal effects
  - Randomized experiments
  - Difference-in-means estimator

Questions?

See you in the next class!