Warsaw Summer School 2023, OSU Study Abroad Program

Fundamentals of Research Design: Populations and Samples, Variables and Their Values,

Data organization, Part I

The Stages of Social Research

- **1) Specify research goals. What you want to investigate and why?**
- **2) Review the literature. Place your question in the context**
- **3) Formulate hypotheses. Provide a theoretical model (a set of propositions). Chose variables and specify hypotheses.**
- **4) Measure and record. (A) Define population and select sample. B) Develop instruments. C) Describe data.**
- **5) Analyze the data: Test hypotheses. Draw conclusions**
- **6) Invite scrutiny. Make decisions about the fit of data and theory. Results are communicated to an audience.**

  (Confirm or reject your initial theory)

The Nature of Social Research

Social Researchers Test hypotheses

- **A hypothesis is a prediction about the relationship between variables. It is usually based upon theoretical expectations about how things work.**

- **At minimum any hypothesis involves two variables.**
- **When causality is involved, we have independent variable(s) (IV) and a dependent variable (DV).**

  X → Y

- **What are independent and dependent variables? Presumed cause and effect notion.**

Why Do We Test Hypotheses?

- **Hypothesis testing is a foundation of science.**
- **In statistical inference, hypotheses generally take one of the two forms: substantive and null.**
- **A *substantive hypothesis* represents an actual expectation. E.g.: higher education increases the likelihood of upward mobility.)**
- **To decide whether a substantive hypothesis is supported by the evidence it is necessary to test a related hypothesis, called the *null hypothesis*.**

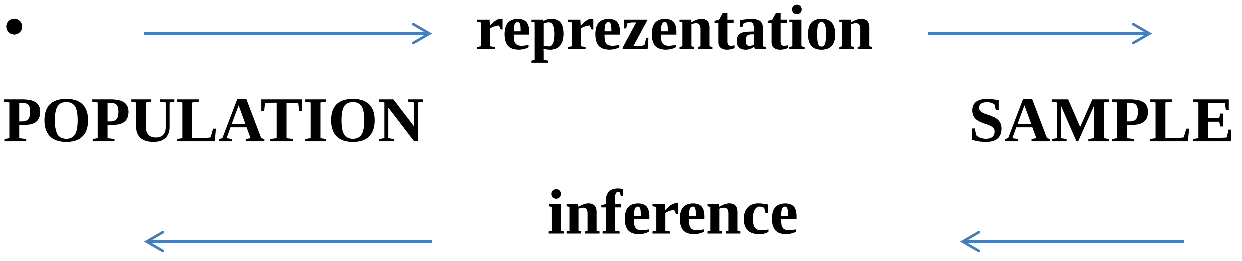
(E.g.: education has no effect on upward mobility.)

Surveys as Data Sources for Hypotheses Testing

A <u>survey</u> is a method of gathering information from a *sample* of a population, through contacts with respondents.

2 Functions of statistics:

- **- to <u>describe</u> the sample data**
- **- to <u>infer</u> from a sample about the population**



- **reprezentation**

POPULATION **SAMPLE**

inference

A Framework for Statistical Work

Units of observation/analysis (cases)

Variables: data characterizing units of observation

A **variable** (age) is a measurable characteristic that differs across the units of observation (individuals).

The observations (years) are the values of the variables for each unit.

Each variable assumes a set of some definite values.

A full measurement procedure specifies values for each variable across all units of observation.

In scientific inquiry we rely on operational definitions to specify concepts.

Matrix form of data, X_{ij}, i = unit of analysis (1,…,N), j = variable (1,…,K)

Cases	Variables					
	Age j = 1	Gender j = 2	Education j = 3	…	Political Party j = K-1	… j = K
i = 1	21	0	15	…	1	…
i = 2	27	1	16	…	2	…
i = 3	18	1	12	…	0	…
i = 4	23	1	16	…	1	…
i = 5	34	0	21	…	2	…
….	…	…	…	…	…	…
i = N - 1	17	0	11	…	2	…
i = N (last person)	36	1	17	…	3	…

Measurement

A good measuring device must meet the condition of exhausting the possibilities of what it is intended to measure.

Mutually exclusive means that each observation fits one and only one of the scale values (categories).

Missing values. Lack of information. Erroneous information. Non-interpretable information

---

Labels

- **Variable names *vs.* variable labels, and variable values**

The variable name is a mnemonic.

The variable name is a descriptive phrase, usually only a few words long, that captures the essence of what the variable is about. Variable label is a short description of the content.

Labels

Assigning <u>value</u> labels

- **Continuous variables usually do not need value labels. Examples: income, results of complicated tests, age, year in the labor force.**

Researchers' advise:

- **If a variable has limited number of values k (k < 10), it is better to label them all or at least a subset, independently of the level of measurement.**

Missing Values

Meaning:

- **Lack of information.**
- **Erroneous information.**
- **Non-interpretable information.**

Don’t know **as a special category – Is this missing datum?**

Levels of Measurement

The level of measurement of a variable refers to the type of information that the numbers assigned to units of observation contain.

Four levels of measurement:

- **nominal (categorical; discrete)**
- **ordinal (rank-order)**
- **interval (distance)**
- **ratio (zero-reference)**

Level of Measurement: Nominal

<u>Nominal Variables</u> (qualitative):

observations consist of separate categories that are labeled.

For practical data processing the names are numerals, but in that case the numerical values is irrelevant (we cannot order them).

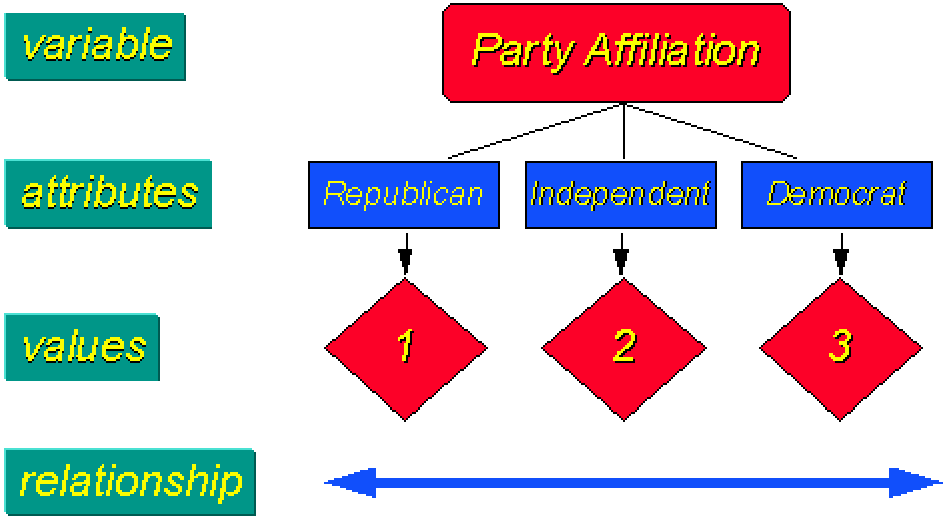
Ex:

"Dummy" (dichotomous) variables: Gender (0, 1 where 0 = female; 1 = male)

Religious affiliation (1, 2, 3, 4, 5 where 1 = Catholic, 2 = Protestant, 3 = Jewish, 4= Muslim, 5 = Other)

Party affiliation

Social Class



Any nominal variable can be recoded into a set of 0,1 variables, called also dummies

Ordinal variables:

- observations consist of separate categories that are arranged in rank order (can be ordered, but we don’t know if the distance between the steps is equal for all steps; no addition, no substraction).

Ex: Likert scales

When no. of categories = large (7/more) → treat rank-order scales as continuous

Metric Variables

Interval variables: observations consist of ordered categories, where distances between categories, called *intervals*, reflect differences in magnitude.

Ex: Celsius

Ratio variables: interval scale with the additional feature of an absolute zero point.

Ex: Income (in Zloty, Dollars, …), Education (in years)

Single Indicators and Composite Measures

An **indicator** consists of a single observable measure, such as a single questionnaire item.

Ex: What year have you been born in?

Composite measures: Scales & Indexes

- use several indicators combined, to create a new variable

Ex: attitudes toward immigrants; self-esteem scale; Notingham scale

Level of measurement and the purpose of the study

Example: Education

1. Elementary, 2. Some High School, 3. High School Completed, 4. Community College, 5. Liberal Arts College Incomplete, 6. College Completed, 7. Above College

- **Nominal Scale (Labels? See 4, 5)**

- **Ordinal Scale?**

- **Interval?**

After recoding into years of schooling: 1=8, 2=10, 3=12, 4=14, 5=14, 6=16, 7=18

Measurement

The measure should be:

- **Valid**
- **Reliable**
- **Exhaustive**
- **Mutually Exclusive**

Validity refers to the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration.

Reliability refers to the likelihood that a given measurement procedure will yield the same description of a given phenomenon if that measurement is repeated.

Reliability is the consistency of measurement.