Research Design and Data Processing

Lectured by: Dr. Xinzhi ZHANG

12 June, 2018

@CUCN Data-driven Journalism Workshop

Agenda

- Research Design
  - Conceptualization
  - Operationalization
  - Measurements
- Data processing
  - Codebook creating
  - Coding

Data-Driven Storytelling

- “Unlike other visual media – such as still photography and video – data visualization is deeply rooted in measurable facts.”
- Data can be the source of data journalism, or it can be the tool with which the story is told — or it can be both (Data Journalism Handbook, chpt 4, paragraph 3)

Cases

- 200 countries in 200 years [Link]

Conceptualization

- A concept is a term that expresses an abstract idea formed by generalizing from particulars and summarizing related observations from social realities.
- Conceptualization is the process of refinement and specification of abstract concepts.
- In data-driven journalism, just like social science research, conceptualization is also a crucial step from daily observations to “datafication.”

Operationalization

- Development of specific research procedures that will result in empirical observations representing those concepts in the real world.
- Concepts should be observed from concrete incidents in the real world.

Variables

- The empirical counterpart of a concept is called a variable.
- A variable can be measured or manipulated in research.
- A variable has more than one possible values.

Concept	Variable	Value
Gender (People's social sex)	Gender	Female Male
Age (How many years for which a person has been born)	Age	0, 1, 2, 3, ……18, 19, 20, ….
Internet use (the frequency of using the internet)	Internet use	Never 0, 1, 2, …, 24 hours per day

Measurement

- Measurement: the scale we use to quantify a variable.

Levels of Measurement

- The measurements could be divided into 4 levels, according to the relationship among the value points that appear for a measurement.

Types of measurement	Variables	Examples
Nominal	Whose attributes have only the characteristics of exhaustiveness and mutual exclusiveness	Gender (female vs. male)
Ordinal	Whose attributes can be rank-ordered along some dimensions	Socioeconomic status, class (high, medium, low)
Interval	Whose attributes are rank-ordered and have equal distances b/w adjacent attributes	IQ
Ratio/Continuous	Whose attributes can be quantified continuously	Age

Data types: two dimensions

- Whether the values are **discrete** or **continuous**?
  - Discrete: headcounts
  - Continuous: rainfalls (mm) per month
- Whether the values are **ordered** or **unordered**?
  - Ordered: the values are comparable (larger/smaller, higher/lower)
  - Unordered: the values only indicate differences (an apple vs an orange; a tiger vs a lion, cucn and nju)

Data types classification – by John C. Hart, UIUC

- table from John C. Hart @ UIUC.

Criteria for Measurements

- Reliability
    - Consistency/stability: measuring something in a consistent and stable manner
- Validity
    - Internal validity: Are the conclusions drawn from a study accurate for the group of people studied?
    - External validity: Can the conclusions drawn from a study be applied to other people, places, or times?

Data (Pre-)Processing (coding*)

- The purpose of data processing is to transfer collected data (i.e., respondents’ answers) into ***machine-readable*** form.
- After that, one can use mathematics and statistics to analyze the data with statistical packages.

- * coding here in Chinese should be translated as 编码, not 编程, which may be expressed as programming.

Machine-readable Data File

- “Spreadsheet”
- Case: a collection of values that belong to a unique subject (unit) in the data file.
    - Example: a person, a news article, a country…
- Variable: a logical grouping of attributes, which describe characteristics or qualities of an object.
    - Example: Age, race, weight, name, scores on a test, and time measured….
- Value: represents the observed attribute of a specific variable of a case
    - Example: 25 years old, Asian, 120 pounds, A..
    - Scale: The possible values the variable can assume form the scale for measuring the variable.

Coding

- Close-ended questions
- Open-ended questions
  - Type in respondents’ answers
  - Group them, assign numerical values to each group if you need to analyze them later.

Codebook

- In the data processing, one needs to create a codebook first.
- A codebook is a document that describes the locations of variables and lists the assignments of codes to the attributes composing those variables.
- A codebook is the primary guide used in the coding process.
- A codebook is the guide for locating variables and interpreting codes in the data file during analysis.

Codebook

- A codebook at least includes the followings:
  - variable name
  - variable label: the description of the variable, usually the question on the questionnaire
  - value definition: you assign a number to each value of the variable: exclusive and exhaustive
  - Define missing values

Coding

- When we define values, we assign numbers to each possible value.
- Each value and each assigned number has a correspondence.
- These numbers are just the “names” for peculiar answers. They don’t have numerical meanings.
    - Assign peculiar numbers
    - Define those values

18 years old ➔ 18 19 years old ➔ 19 20 years old ➔ 20 Refused to answer ➔ ???	Male ➔ 1 Female ➔ 2 Blank ➔ ??	Disagree ➔ 1 Neither disagree nor agree ➔ 2 Agree ➔ 3 Don’t know ➔ ???

Coding

- Missing values should be defined.
- Examples:
    - Don’t know ➔ -100
    - Refusal ➔ -101 (-100)
    - Blank ➔ -102 (-100)
- Be consistent!

Codebook: An example

- - end of the session -