Statistical Reasoning

Week 2

Sciences Po - Louis de Charsonville

Spring 2018

Week 2 : Data and Survey designs

Survey Designs

Datasets and Variables

- Datasets
- Variables

Research Paper

Data exploration in Stata

Survey Designs

To answer a question about the subjects of a population, we need data (at least on a sample of obs.)

Different types of data

- Observational vs. Experimental : collected from surveys (interviews, forms) or experiments (treatment group vs control group).
- Cross-sectional vs. Panel vs. Time Series
- Comparative vs. Case studies

In this course, we will use **non-experimental cross-sectional** data.

Structures of data - 1/2

Cross-sectional Data

- A **snapshot** of a sample at a **single point in time.**
- Cross-sectional surveys capture data about each unit of the sample once.
  - Ex : one round of the European Social Survey

Rolling cross-sections

- Capture the same data on various samples of units (i) at **different points in time** (t).
- The observed units (i) differ for each (t), but the collected data remains the same (same questions).
  - Ex : several rounds (years) of the ESS

Structures of data - 2/2

Panel/Longitudinal Data

- Data is collected at **several points in time (t)** for **the same individuals (i)**.
- Cross-sectional data = a photograph ; Panel-data = a film.
- Panel data have many advantages, especially for improving causal reasoning ;
  - But require more sophisticated procedures not covered in this course.

Time-series data

- Data is collected on several characteristics of a few (i) at many points of time (t) *rightarrow* **a few and constant (i) at numerous (t)**
- To analyse the evolutions of the relationships between the characteristics of (i) over time ;
  - Other sophisticated procedures not covered in the course.

Cross-sectional data

- First step : Have a look on the **codebook** of your survey
  - Provides information on sampling and survey methods, details on questions, variables, etc.
- Second step : **browse** the dataset with Stata.
  - Each row : all data about one observation (one unit).
  - Each column : all values of responses to one variable (characteristic).
  - Each cell : value of one variable for one unit.
  - Missing data : represented with a dot "." in Stata.

Sampling Weights

- Sampling weights may be required for some datasets when some groups are over/under represented in the study sample ;
- The use of weights enables us to make the descriptive statistics representative of the population, even though the sample is not ;
- Data codebooks give some information about the weights of sampled units in the population ;

Observations and Variables

- A **variable** is a characteristic that can be coded and which has more that one single value.
  - Ex : gender, income, age.
- An **observation** is a unit for which data is observed.
  - Ex : individual 1 in the survey.

Example from the World Values Survey

- The observations are the 61 062 individuals in the dataset for whom data has been collected (-count-) ;
- There are 378 variables in the dataset (-display c(k)-) ;
- Example of a variable : age, whose values range from 15 to 99 (-sum age-)

Variable Types

We distinguish between **qualitative** (or **categorical** and **quantitative** variables.

- A variable is **qualitative** if the response belongs to a set of categories.
  - Format in Stata may be string or numerical
- A variable is quantitative if the response takes a **numerical** value (age, weight, etc.).

Qualitative variables

- **Nominal** : it cannot be ordered.
  - Ex : political parties (1=Socialist Party ; 2=Republican party ; 3=Green party ; 4=far right party ; 5=far left party)
- **Binary** : take the value 0 or 1. This is a subset of nominal variable.
  - Ex : gender (0= male ; 1=female
- **Ordinal** : categories can be ordered.
  - Ex : satisfaction level (1=Not at all ; 2=Somewhat ; 3=Satisfied ; 4=Very satisfied).

Basic information on a variable

- Quantitative variables : Mean can be used
- Nominal variables : Mean cannot be used, just frequencies or percentages.
- Ordinal variables :
  - Mean can be used if we have many observations.
    - Ex : the average level of satisfaction
  - Frequencies/percentages are preferable if we have few observations
    - Especially when the variable refers to a classification built from a quantitative variable

Dependend and Independent Variables

- This distinction refers to the research question.
- The **dependent** (or "explained", or "of interest") variable is the phenomenon which is tested by the researcher : we want to see how it is linked to other variables, how it reacts to a change in other variables, to what extent it is explained by other variables ;
- These other variables are the **independent** (or "explicative") variables.
- Most generally, there is one dependent variable and several independent variables. It will be the case in your research project.

Quantitative Research

- The most common way of sharing results : scientific journal article ;
- Peer-reviewed process before publication.
  - But this does not prevent you from being critical ! (**Bias of publication**).
- Some examples of journals : American Sociological Review, American Journal of Political Science, World Development.
- In economics, top 5 is indentified as : Econometrica, Quaterly Journal of Economics, American Economic Review, Journal of Political Economy, Review of Economic Studies.
- Website to find/read articles : Google Scholar, JSTOR (Sciences Po grants new a free access !).

Basic structure of an article

- Abstract
- Introduction
- Literature Review
- Data
- Methods
- Results
- Discussion and conclusion
- References

Getting started

- Explore the various datasets to get ideas
- Formulate a specific research question and make assumptions about the relationship between a dependent variable and several independent variables.
  - Ex : Democratic attitudes in Africa (WVS 2005), Anti-Immigration attitudes and Exposure to Minorities (EVS 2008), Can Money Buy Happiness ? (WVS 2005), Attitudes towards Homosexuality in Europe (EVS 1999), Conservative Values in Hungary (ESS 2002), Institutional Trust, Political Participation and Religiosity in the Muslim World (WVS 2005), The Perception of Environmental Issues in the Emerging Powers : China and India (WVS 2005)

$\Rightarrow$ Final output : a research paper (structured as in previous slide).

Requirements

By Week 4 :

- Choose one or two partners
- Browse some empirical quantitative articles of interest to you
- Decide on a topic and your dataset
- Prepare a research proposal including : the question, assumptions, dataset, and the dependent and independent variables (use the template until "dataset")
- Make sure the variables exist in the chosen dataset !

Technical requirements

- The dependent variable must be quantitative
- Cross-sectional data only
- The analysis may be a case study or comparative

you must use the datasets provided

Understand the data

- Before analysis, you have to understand your dataset and its variables
- Use the codebook as a first step to know how the data was collected !
- Second step -and key stages : data exploration and data management ;

Dataexploration

- Understand the variables : what they measure and how they are coded ;

Datamanagement :

- You may need to rename, recode, label variables and create new variables (ex : categories of earnings) ;
- You may also need to subset data from a given year or subsample ;

Useful STATA commands for exploration

- -browse- and –describe- to identify data structure (cross section ? Time series ?) and the variables ;
- -lookfor- followed by keywords to look for variables of interest
- Be careful about the syntax !
- -codebook- and –fre- to know how the variable is coded
- -rename- to make the variables easier to understand and type in the command window
- -gen- to create a new variable
- -recode- to change the format of an existing variable (especially to make a qualitative variable categorical)
- -label var- to give a label to a variable