

# Statistical Reasoning

## Week 3

Sciences Po - Louis de Charsonville

Spring 2018

Research projects

Data Management

Definitions

Principles of Data Management

# Research projects

## By next week :

- ▶ Your project partner(s)
- ▶ The topic, research question, main hypotheses.
- ▶ The data and the dependent variable (which must be quantitative)

Fill the Google form.

## Assignment submission

Email the assignment by midnight, Do-File + Word document to :  
louis.decharsonville@sciencespo.fr

# Data Management

# What to do with data ?

---

## Data exploration and management

- ▶ An essential step before going further with analysis ;
- ▶ Use the Stata commands : lookfor describe codebook recode rename

## Univariate statistics

- ▶ To describe single variables and their distributions ;
- ▶ `tab` `fre` `sum` `hist`

## Bivariate statistics

- ▶ To describe / model the association between 2 variables
- ▶ Double entry table (cross-tabulation), regression, etc.

## Multivariate statistics

- ▶ To describe / model the associations between 3 or more variables

# What methods for which variables?

---

Methods	Qualitative variables	Quantitative variables
Univariate statistics	<ul style="list-style-type: none"><li>• Frequencies</li><li>• Percentages</li><li>• Modal category</li><li>• Mean, median (<i>for ordinal variables only</i>)</li></ul>	<ul style="list-style-type: none"><li>• Mean</li><li>• Mode</li><li>• Median and quartiles</li><li>• Range and interquartile range</li><li>• Standard deviation</li></ul>
Bivariate statistics	<ul style="list-style-type: none"><li>• Cross-tabulations</li><li>• Cramer's V</li><li>• Logistic or multinomial regression</li></ul>	<ul style="list-style-type: none"><li>• Correlation <i>if independent var. are quantitative</i></li><li>• Simple regression</li></ul>
Multivariate statistics	<ul style="list-style-type: none"><li>• Logistic or multinomial regression</li></ul>	<ul style="list-style-type: none"><li>• Multiple regression</li></ul>

# For your paper

---

- ▶ In the **first draft** (Week 5) : Describe your independent and depend variables of interest with **univariate statistics**.
- ▶ In the **second draft** (Week 9) : Explore associations between variables with **bivariate statistics**, especially between your dependent variable and several or all independent variables.
- ▶ In the **final paper** (Week 12) : Everything before + explore the associations between several variables at the same time with **multivariate analyses** (regression models) + a “Results” section.



## Data formatting

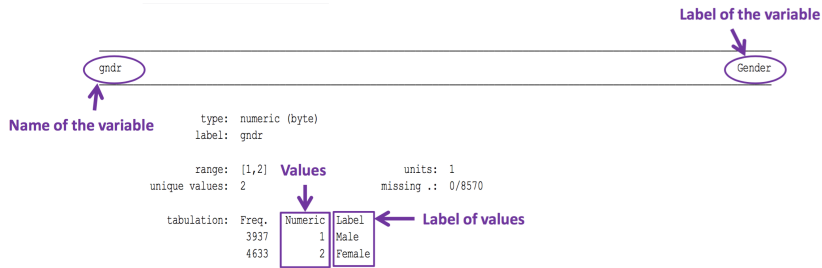
- ▶ From raw data to .dta.
- ▶ Good news : everything has been done for you !

## Data subsetting

- ▶ *Definition* : Restrict your observation to a subsample of your original dataset
- ▶ In order to **make data cross-sectional** (limit data to one year) ;
- ▶ Or to limit your observations to specific category/categories you are interested in (ex : South American countries only).

# Renaming and Labelling variable and values

- ▶ Use `describe var` to check the name, label, storage type, values and labels of values of your variable;
- ▶ Use `rename var` to give a simpler name to your variable;
- ▶ Use `label var` to give a description to your variable;
- ▶ Use `label define` and `label values` to give a description to each value of your categorical variables;



# Encoding variables, replacing values

---

*Not always necessary, depending on your data.*

## Dealing with string variables : encode

- ▶ String variables means their values are formatted as characters and not as numeric values ;
- ▶ To use statistical methods such as frequencies or percentages, Stata needs them to be numeric use `encode var`
- ▶ Ex : `encode gndr, gen(gender)`
- ▶ From now, use `gender` and not `gndr` in your commands.

# Missing values

---

- ▶ Missing values may be coded with arbitrary numbers or letters, sometimes indicating a reason why there is no value (ex : the respondent refuses to answer) ;
- ▶ These numbers will be taken by Stata as numeric values instead of missing values, and calculations will be biased ;
- ▶ Use `replace` to replace missing values by “.”

# Recoding variable

---

## Create a new variable

- ▶ To create a new variable from an existing one by assigning new categories, more relevant (according to you at least);
- ▶ The original variable still exists in case you need it;
- ▶ Always double check after a recode to make sure it has been correctly done (cross-tab the original and the new variables);

### Example :

Old variable `agea`, new variable `agecat`, with labels of values :

```
recode agea (min/24=1 "15-24") \\\  
(25/34=2 "25-34") (35/44=3 "35-44") \\\  
(45/54=4 "45-54") (55/64=5 "55-64") \\\  
(65/max=6 "65-98"), gen(agecat)
```