Statistical Reasoning

Week 1

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Course objectives

- Gain a **conceptual** and **pratical** understanding of quantitative methods applied to research in social sciences.
- Feel **comfortable with numbers**.
- Be familiar with **statistical concepts**, **survey data**, **regression analysis**
- Learn **STATA**, a widely used statistical software.

Topics covered

- Datasets and Survey Designs
- Variables
- Distributions
- Estimation
- Correlation and Comparisons
- Regression

Course Requirements

- No previous requirements in statistics
- A genuine interest in **social sciences** !

Recommended readings

- Agresti, A. and Franklin, C. (2013), Statistics : the Art and Science of Learning from Data.
- Briatte, F. and Peteve, I. (2012), Stata Guide : A Student Guide to Statistics with Stata.

How to succeed ?

Use and use and use STATA again !

- On the *Drive* of the course : all datasets are provided.
- Follow the *do-file* in class and reviews it after class, **practice is the key to success !**
- Download files and datasets *before* each class.

Evaluation

A research project

- Formulate a research question
- Choose the data among the datasets provided
- Perform analysis using STATA
- Report the results in a research paper

Evaluation

What's expected ?

- a **do-file** : make it clean, organised and meticulously commented.
- a **short research paper** :
  - Follow template's guidelines
  - Clear and concrete writings, no jargon, no long sentences. Your readers are busy and impatient.
  - A good paper is not a travelogue of your search process. Don't report the hundred things you have tried but that didn't work.

Timeline & Due dates

- Research proposal due by *February* 20^{th}.
- First draft due by *March* 6^{th}.
- Second draft due by *April* 3^{rd}.
- Final paper due by *April* 24^{th}.

Why this course ?

- Describe, hopefully understand and maybe suggest some **explanations of social phenomena**.
- ⚠ Quantitative methods are **one tool** among *others*. **Critical thinking** and **modesty** are required.
- Goal of the course : be able to establish relationships / correlations between variables and **critically read articles** which do it.

"There are three kinds of lies : lies, damned lies, and statistics."

— B. Disraeli

Describing social phenomena : generalizing and comparing

- Individual → Global
- Use data to find are average characteristics of a whole population.
- Group comparison within a population (Ex : is there a bias against women in income levels ? To what extent do obesity rates differ between social classes ?)

⇒ **Identify overall patterns and trends and describe the data through numbers and graphs**

Predicting social phenomena : identifying statistical relationships

- Determine whether two phenomenas are related.
  - Link between lung cancer and smoking, gender and income.
- Investigate **causal** relationship.
  - is smoking *cause* more lung cancer ?

Population vs Sample

- A *population* refers to all the possible units of the group we are interested in to answer a question. The units of interest are called *subjects*.
  - The units are generally people but not always : they may be countries, firms, etc.
- Populations are generally too large to collect information on all their members (too costly, too time-consuming), so we use samples to answer our question (except Census).
- A *sample* is the subset of is the subset of the population we are able to observe, for whom we have data. One unit of the sample is called an *observation*.

Sampling issues

There are different sampling methods :

- **Random sampling** : *random* means that every member of a population has an equal chance of being selected into a sample.
- **Representative sampling** : match the larger population on specific characteristics.
- **Convenience sampling** : selection is done on willingness to participate, ease-of-access, etc.

$\Rightarrow$ The major benefit of random sampling is that any differences between the sample and the population from which the sample was selected will not be systematic.

Descriptive versus inferential statistics

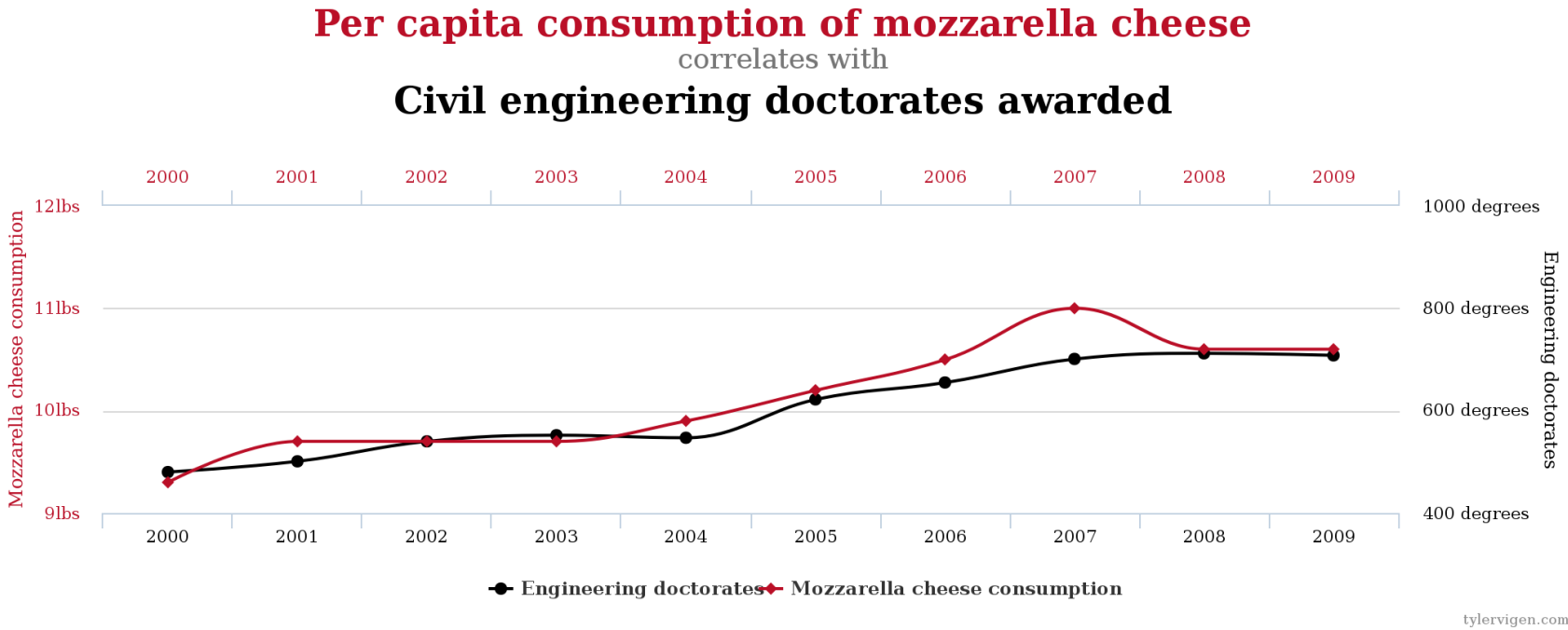
- **Descriptive statistics** refer to methods for summarizing the data
- **Inferential statistics** refer to methods of making predictions or conclusions about a population, based on data concerning a sample of this population.

Parameter versus Statistic

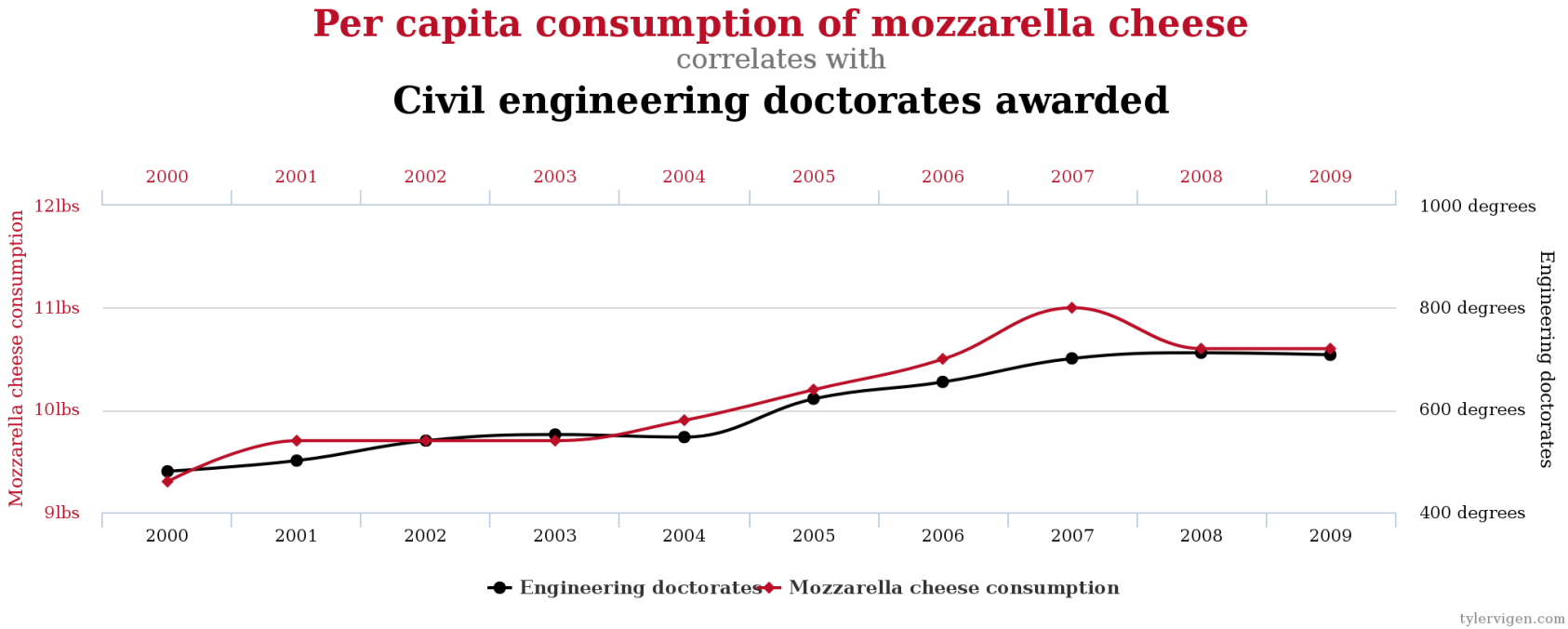
- A **statistic** is a numerical summary of a **sample** taken from a population
  - Ex : 59% of surveyed people on January 14th and 15th said they would like the French wing to organize primary elections. The survey was conducted with 1011 persons representative of the French population over 18 ; the margin of error is of 2,5%
- A **parameter** is a numerical summary of the **population** ;
  - Ex : The percentage of all French people in favour of these primary elections falls within 2,5% of the survey's value of 59%, that is, between 56,5% and 61,5%.

$\Rightarrow$ **True parameter** values are almost always **unknown**, so we use sample statistics to *estimate* the parameter values.

Correlation versus Cause



Correlation versus Cause



Correlation is not causality

The fact that two variables are correlated, does not mean that there is a causal link between two variables. This is called a **spurious relationship**.