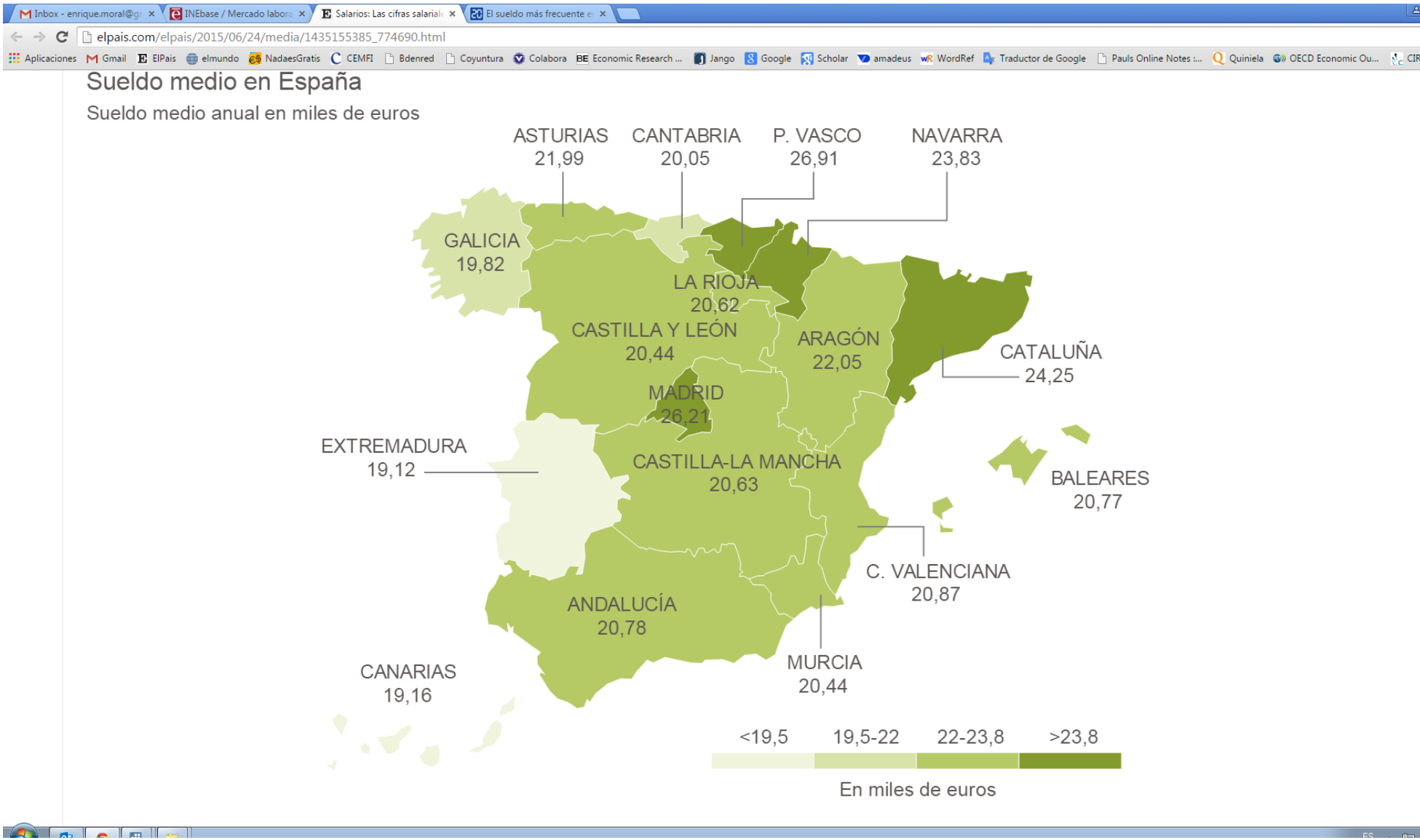
Introductory topics — Statistics

Quantitative Methods

Enrique Moral-Benito

Lesson 1: 26 September, 2016

Wages across Spanish regions in 2013



Population and samples

- Suppose we observe the wages of 1000 Spanish workers in the year 2013.
- We just have a collection of 1000 numbers that we want to analyze.
- For instance, we may want to look at the mean wage.
- However, this is not the figure we are after.
- In general, we are interested in the mean wage in the **population** of workers (around 20 million workers).
- The problem is that we do not observe that population, but a **sample** drawn from it.
- Ideally, the sample should be a **random sample**, i.e., it needs to be representative of the population and not biased in a particular direction.

Data collection

- The main reason why we generally have samples and not data on the full population is that collecting data is costly.
- For instance, the INE surveys thousand of workers for the EES but think of the cost of surveying 22 millions of workers each year.
- Moreover, collecting a random sample is difficult in practice.
- Therefore, we often assume that the samples we observe are random draws from the population of interest.
- Is this assumption reasonable?

The 1936 US Presidential Election

- A. Landon (Republican) against incumbent F. Roosevelt (Democrat).
- The Literary Digest poll was the largest and most expensive poll ever conducted, with a sample size of 2.4 million people.
- George Gallup used a much smaller sample of about 50,000 people.
- **Which one predicted Roosevelt's victory (62% vs. 38%)?**

The 1936 US Presidential Election

- A. Landon (Republican) against incumbent F. Roosevelt (Democrat).
- The Literary Digest poll was the largest and most expensive poll ever conducted, with a sample size of 2.4 million people.
- George Gallup used a much smaller sample of about 50,000 people.
- **Which one predicted Roosevelt's victory (62% vs. 38%)?**
- Literary Digest prediction was Landon 57% against Roosevelt's 43%.
- George Gallup predicted Roosevelt 56% against Landon's 44%.
- **Why?**

The 1936 US Presidential Election

- A. Landon (Republican) against incumbent F. Roosevelt (Democrat).
- The Literary Digest poll was the largest and most expensive poll ever conducted, with a sample size of 2.4 million people.
- George Gallup used a much smaller sample of about 50,000 people.
- **Which one predicted Roosevelt's victory (62% vs. 38%)?**
- Literary Digest prediction was Landon 57% against Roosevelt's 43%.
- George Gallup predicted Roosevelt 56% against Landon's 44%.
- **Why?**
- Literary Digest used a list of 10 million names based on telephone directories.
- In 1936, telephones were much more of a luxury than they are today.
- Such a list is guaranteed to be slanted toward middle and upperclass voters excluding lower income voters.
- The Literary Digest sample was not a random sample of the US population.

Types of data

- We generally work with **observational** data as opposed to **experimental** data.
  - What is the effect of training programs to the unemployed?
  - Observational: training programs not randomly assigned (e.g. real world).
  - Experimental: training programs randomly assigned (e.g. medical trial).

- We can have **cross-sectional** data, **time-series** data, or **panel** data.
  - Cross-sectional data refers to several individuals at the same point in time.
  - Time-series data refers to the same individual at different points in time.
  - Panel data refers to several individuals at different points in time.

Describing data

- If we open the EES data file corresponding to the year 2002:

worker	year	sex	wage	birth	tenure	...
1	2002	6	61489	1965	12	...
2	2002	1	68215	1961	2	...
3	2002	1	40099	1955	27	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
186761	2002	1	16669	1945	4	...
186762	2002	6	10100	1980	2	...
186763	2002	1	31306	1959	17	...

- Given the large number of workers it is difficult to learn something from visual inspection.
- STATA (or Excel) provides the appropriate tools for exploiting this collection of numbers.

Describing data with Excel

- Let's open the file `ees2002.xlsx` and compute the following figures in the sample:
  - The average annual wage.
  - The maximum and minimum annual wage.
  - The average hourly wage.
  - The number of men.
  - The proportion of women.
  - The proportion of temporary workers.
  - The proportion of public workers.
  - The average wage by sex.
  - The average wage of public and private workers.
  - The average wage of Spanish and African workers.
  - The average wage by sex and type of workday.
  - The average wage by age group (less than 30 years and more than 30 years).
  - The average wage by CCAA.

Describing data with STATA

- Let's open the file `ees2002.dta` and compute the following figures in the sample:
  - The average annual wage.
  - The maximum and minimum annual wage.
  - The average hourly wage.
  - The number of men.
  - The proportion of women.
  - The proportion of temporary workers.
  - The proportion of public workers.
  - The average wage by sex.
  - The average wage of public and private workers.
  - The average wage of Spanish and African workers.
  - The average wage by sex and type of workday.
  - The average wage by age group (less than 30 years and more than 30 years).
  - The average wage by CCAA.

Describing data with Excel and STATA

- Now have a look at the files `ees2006.xlsx` and `ees2006.dta`.
- Compute the same descriptives for the year 2006.
- Analyze the evolution of these figures between 2002 and 2006.

Describing data with Excel and STATA

- Now have a look at the files `ees2006.xlsx` and `ees2006.dta`.
- Compute the same descriptives for the year 2006.
- Analyze the evolution of these figures between 2002 and 2006.

- Which alternative is more prone to errors?

The Reinhart-Rogoff Scandal