Research Seminar

Theme C

Quantitative Data Sources and Statistical Packages.

Ethical issues on data use in research

Vítor Escária

(+ Paulo Madruga and Carlos Farinha)

OCT 2014

LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

Goal

Discuss the access to and use of quantitative data, the potentialities of statistical packages, and ethical issues on quantitative data use when carrying on empirical research.

... based on our own experience

Topics

- Empirical research
- Access/preparation of data
- Data sources
- Statistical packages
- Ethical issues

Empirical research

- **Definition of a *topic* – based on a theoretical model/ the literature....**
- **Definition of the hypotheses to test – the *thesis***
- ***Access and preparation of data***
  - **Collection of data and data set construction**
  - **Measurement methods and instruments**
- ***Test of hypotheses* – data analysis and model building**
  - **Univariate, bivariate or multivariate analysis**
- **Presentation of results**

Empirical research

- **Empirical research is subsidiary to an *idea*, is carried out to answer a question, which is the centre of the thesis**
  - **We don't carry on empirical research just to do empirical research**
- **Empirical research enables to test whether an idea is true, to answer a concrete question**
  - **The increase of minimum pensions helps to reduce poverty?**
  - **A marketing campaign raises the number of consumers of a given product?**
  - **The action of the central bank reduces de cost of funding for companies?**

Access/Data preparation

- **Secondary information – produced by entities of the Statistical System/ other entities**

- **Primary information – direct collection**
  - **Enquiries**
  - **Case studies**

Data access: Most common sources of secondary statistical information

- **Statistical entities**
  - **National:**
    - **INE, Bank of Portugal, DGO, etc...**
  - **Internacional**
    - **Eurostat, ECB, European Commission, OECD, United Nations, World Bank, IMF, WTO, ILO, etc...**
- **Other**
  - **Datastream, Bloomberg, Reuters, Dun & Bradstreet**

Data access

Secondary data from statistical sources

- **Some checks:**
  - **metadata**
  - **statistical classifications**
  - **goals and data to use – issues:**
    - different sources for the same variable
    - original *vs* normalised values
    - international comparisons
    - breaks in series

Data access

Direct collection/ Enquiries

- **Some checks:**
  - **population and sample selection**
  - **enquiry methodology**
  - **pilot enquiry and simulation of analysis**
  - **type of questions: open vs closed**
  - **answers coding**
  - **closed questions: measurement scales**
  - **monetary and time costs to apply an enquiry**

Types of Data

- **With respect to the temporal dimension/units observed**
  - **Time series**
  - **Cross sections**
  - **Panels**
- **With respect to the type of statistical unit**
  - **Microdata**
  - **Aggregate data**

Data analysis and model building
Types of analysis

- **Static analysis**
  - **One observation point**
- **Comparative statics**
  - **Two observation points – two independent observations**
- **Dynamic analysis**
  - **More than one observation point and ability to measure flows**

Statistics Sources

- **Free access**
  - **Portal INE**
  - **Portal Bank of Portugal (BPSTAT, etc)**
  - **Portal Eurostat**
  - **European Commission: DG ECFIN (AMECO, KLEMS, BACH, …)**
  - **Portal OECD**
  - **WTO (World Trade)**
  - **IMF, World Bank, etc**
  - **CMVM/Euronext**

Statistics Sources

- **Available in ISEG terminals**
  - **Datastream**
  - **BANKSCOPE – information on over 23,000 banks**
  - **CHELEM – world trade, macroeconomic data and balance of payments**
  - **OSIRIS – Information on listed companies**

Statistics Sources

- **Available for research (protocol)**
  - **Protocol ex-MCTES/INE – access to microdata**

  NOTE:

  To access data it takes time: contracts / protocols / waiting time…=>

  Need to consider this at an early stage of the research.

Statistical packages

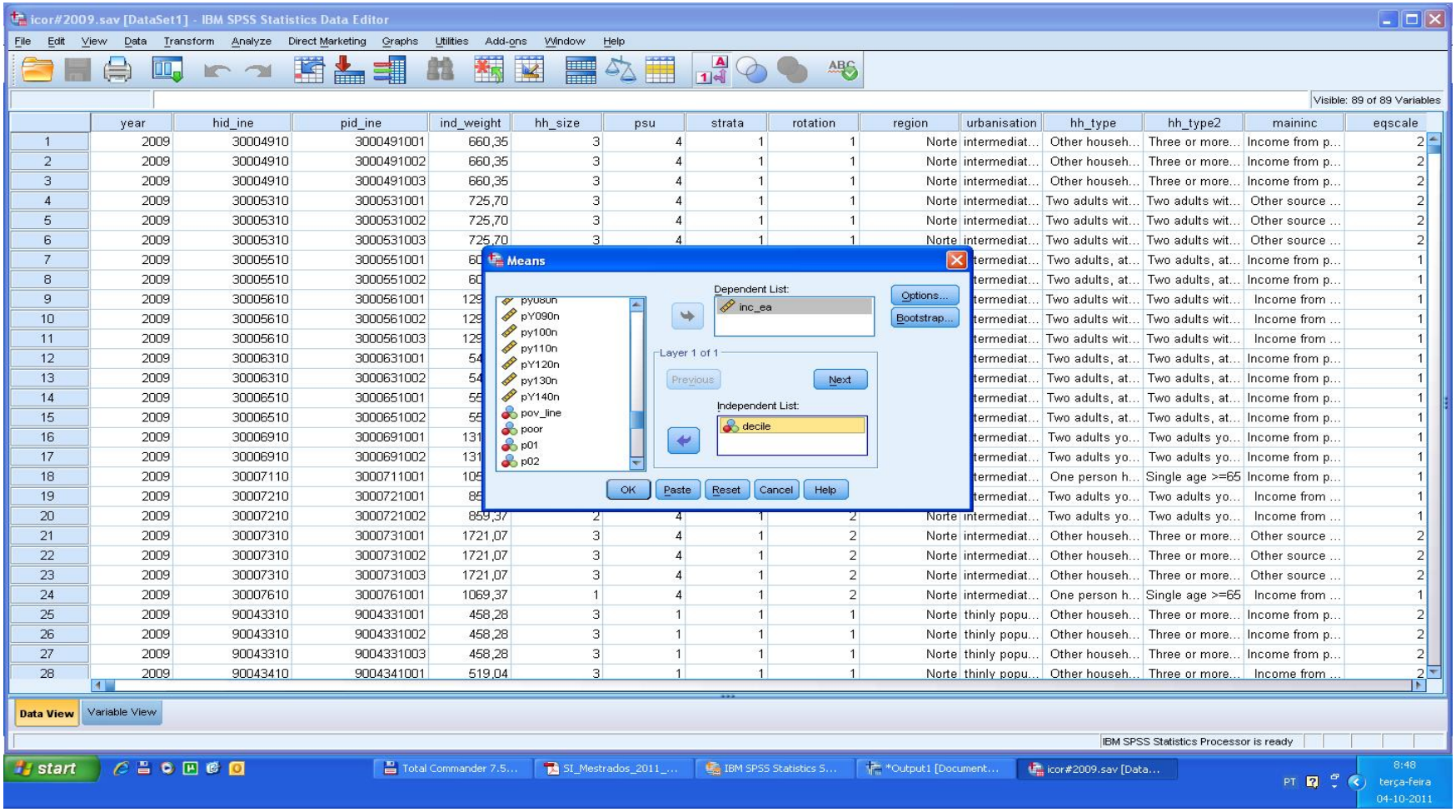
- **The choice of the package**
  - **Depends on the work to carry on and on the structure of the data**
  - **Three levels:**
    - **Excel**
    - **PASW (ex-SPSS), Stata, SAS, TSP, EViews...**
    - **Gauss**

For a discussion on the levels of popularity of the different statistical packages check: The Popularity of Data Analysis Software by Robert A. Muenchen (http://r4stats.com/popularity)

Statistical packages

- **Work with SPSS/STATA/SAS**
  - **All have an user interface based in a system of menus, a data sheet and an output window**
  - **Most have another interface that enable the user to write and run procedures using commands**
    - **Many of the most powerful commands are only available this way**
    - **Usually there is some possibility of interaction between the menu system and the programming interface – allows to use the menus to create command lines**

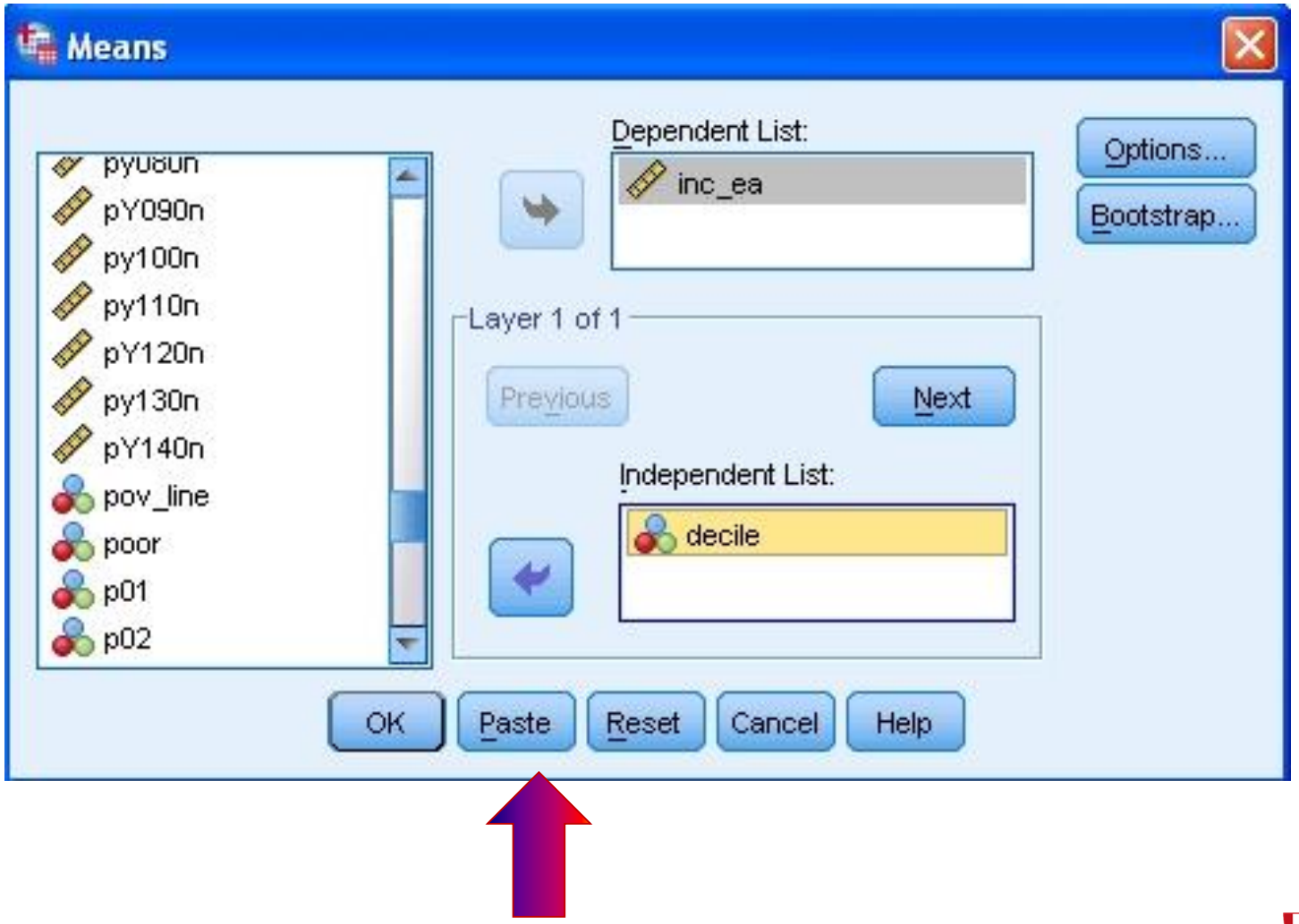
Statistical packages



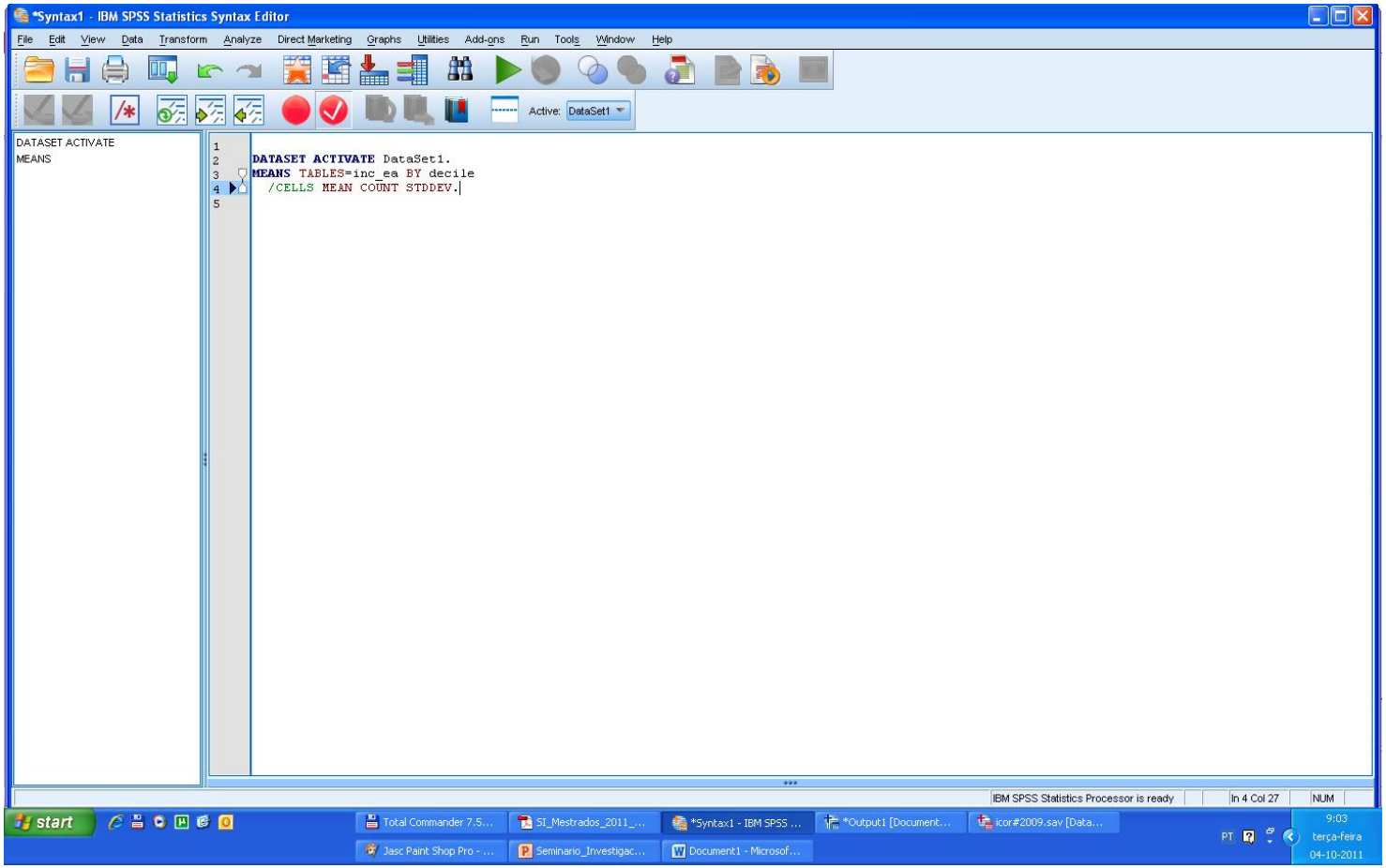
	year	hid_ine	pid_ine	ind_weight	hh_size	psu	strata	rotation	region	urbanisation	hh_type	hh_type2	maininc	eqscale
1	2009	30004910	3000491001	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
2	2009	30004910	3000491002	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
3	2009	30004910	3000491003	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
4	2009	30005310	3000531001	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
5	2009	30005310	3000531002	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
6	2009	30005310	3000531003	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
7	2009	30005510	3000551001	60						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
8	2009	30005510	3000551002	60						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
9	2009	30005610	3000561001	129						termediat...	Two adults wit...	Two adults wit...	Income from ...	1
10	2009	30005610	3000561002	129						termediat...	Two adults wit...	Two adults wit...	Income from ...	1
11	2009	30005610	3000561003	129						termediat...	Two adults wit...	Two adults wit...	Income from ...	1
12	2009	30006310	3000631001	54						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
13	2009	30006310	3000631002	54						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
14	2009	30006510	3000651001	55						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
15	2009	30006510	3000651002	55						termediat...	Two adults, at...	Two adults, at...	Income from p...	1
16	2009	30006910	3000691001	131						termediat...	Two adults yo...	Two adults yo...	Income from p...	1
17	2009	30006910	3000691002	131						termediat...	Two adults yo...	Two adults yo...	Income from p...	1
18	2009	30007110	3000711001	105						termediat...	One person h...	Single age >=65	Income from p...	1
19	2009	30007210	3000721001	85						termediat...	Two adults yo...	Two adults yo...	Income from ...	1
20	2009	30007210	3000721002	859,37	2	4	1	2	Norte	intermediat...	Two adults yo...	Two adults yo...	Income from ...	1
21	2009	30007310	3000731001	1721,07	3	4	1	2	Norte	intermediat...	Other househ...	Three or more...	Other source ...	2
22	2009	30007310	3000731002	1721,07	3	4	1	2	Norte	intermediat...	Other househ...	Three or more...	Other source ...	2
23	2009	30007310	3000731003	1721,07	3	4	1	2	Norte	intermediat...	Other househ...	Three or more...	Other source ...	2
24	2009	30007610	3000761001	1069,37	1	4	1	2	Norte	intermediat...	One person h...	Single age >=65	Income from ...	1
25	2009	90043310	9004331001	458,28	3	1	1	1	Norte	thinly popu...	Other househ...	Three or more...	Income from p...	2
26	2009	90043310	9004331002	458,28	3	1	1	1	Norte	thinly popu...	Other househ...	Three or more...	Income from p...	2
27	2009	90043310	9004331003	458,28	3	1	1	1	Norte	thinly popu...	Other househ...	Three or more...	Income from p...	2
28	2009	90043410	9004341001	519,04	3	1	1	1	Norte	thinly popu...	Other househ...	Three or more...	Income from ...	2

LISBOA SCHOOL OF ECONOMICS & MANAGEMENT

Statistical packages



Statistical packages



Statistical packages

```
SET  printback=listing messages=listing.
Title     '****              rsi_model_09#001              ****'.
* ********************************************************************************************************.
* rsi_model_09#001:
*.
*    Simulation of RSI based on SILC 2009
*.
*    Builds the individual datafile from silc files
*.
* ********************************************************************************************************.
* @cfr2011 - version 24-09-2011
** *******************************************************************************************************.
DATASET CLOSE ALL.
GET FILE='c:\temp\icor2009r.sav'/KEEP hid_ine pid_ine rb010 rb030 rb050 rb080 rb090 rb220 rb230 rb240.
DATASET NAME DataSet1 WINDOW=FRONT.

IF (rb080 ne rb010)age=(rb010-1)-rb080.
IF (rb080 eq rb010)age=0.
FORMATS age (f3.0).
VARIABLE LABELS age 'Age at the end of the income reference period'.
EXECUTE.

Rename vars (rb010 rb030 rb050 rb090 rb220 rb230 rb240 = year pid ind_weight sex pid_father pid_mother
pid_partner).
execute.
compute hid = trunc(pid/100).
variable label hid 'Household ID'.
execute.
```

Statistical packages

- **Advantages in using syntax files**
  - **Once the language is known it saves a lot of time – it is easier to change some bits of the program and run it again than to repeat all the steps**
  - **The programme allows to understand the research strategy and options made when dealing with data or model problems**
  - **The same programme may be used in different projects**

NOTE: It is possible to find many procedures available on line

Ethical issues

- **Behavioural rules on research**
  - **Discuss intellectual property frankly**
  - **Results have to be verifiable by our peers**
  - **Different hypothesis assumed must be presented an justified**

Ethical issues

- **Behavioural rules on data use**
  - **Do not use data for commercial or other non agreed uses**
  - **Always refer who has made the data available (and the version that is being used)**
  - **Respect the rules of confidentiality and anonimization**
  - **Destroy data in the end of the period agreed**