

# Seminário de Investigação

Tema C:

Bases de Dados Quantitativas e utilização de Software Estatístico.

Questões éticas da utilização da informação na elaboração do TFM

Vítor Escária  
(+ Paulo Madruga e Carlos Farinha)

OUT 2014

# Objectivo da Sessão

*O acesso e utilização de informação quantitativa, as potencialidades do software estatístico e questões éticas na utilização de informação quantitativa quando se faz trabalho empírico em economia ...*

*... a partir das nossas experiências práticas*

# Tópicos

- Trabalho empírico em economia
- Obtenção/ preparação dos dados
- Fontes de informação
- Softwares estatísticos
- Questões éticas

# Trabalho Empírico

- Definição de ***objecto de interesse*** – do modelo teórico, da literatura ....
- Definição de hipóteses a testar – a ***tese***
- ***Obtenção e preparação*** de dados
  - Recolha e sistematização de dados
  - Instrumentos e métodos de medida
- ***Teste das hipóteses*** – confirmação - Modelação e análise de dados
  - análises univariadas, bivariadas, multivariadas
- Apresentação dos resultados

# Trabalho Empírico

- **O trabalho empírico é subsidiário de uma ‘ideia’, da resposta a uma questão que é o centro da tese.**
- **Não fazemos trabalho empírico para fazer trabalho empírico**
- **O trabalho empírico serve para comprovar uma ideia, para responder a uma questão concreta**
  - **O aumento das pensões mínimas ajuda a reduzir a pobreza?**
  - **Uma certa campanha publicitária permite ganhar novos clientes para um produto?**
  - **A ação do banco central reduz os custos de financiamento das empresas?**

# Obtenção/preparação dos dados

- **Informação secundária - produzida por entidades do Sistema Estatístico/ outras entidades**
- **Informação primária - recolha directa**
  - Inquéritos
  - Estudos de caso

# Obtenção dos dados : Fontes de informação estatística secundária mais comuns

- **Entidades sistema Estatístico**
  - **Nacionais:**
    - INE, Banco de Portugal, DGEEP, DGO, etc...
  - **Internacionais**
    - Eurostat, OCDE, Comissão Europeia, Nações Unidas, Banco Mundial, FMI, etc...
- **Outras fontes**
  - Datastream, Bloomberg, Reuters, Dun & Bradstreet

# Dados de fontes estatísticas

- **Alguns cuidados a ter:**
  - **metainformação**
  - **nomenclaturas**
  - **objectivos do estudo e os dados a utilizar - exemplos**
    - Fontes diferentes para a mesma variável
    - Valores originais vs valores normalizados
    - Comparações internacionais
    - Quebras de séries

Obtenção dos dados

# Recolha directa/Inquéritos

- **Alguns cuidados a ter:**
  - **selecção da população a inquirir e da amostra**
  - **metodologia de inquirição**
  - **realização de um teste ao inquérito e de uma simulação de apuramentos**
  - **tipos de questões: abertas vs fechadas**
  - **codificação de respostas**
  - **questões fechadas: escalas de medida**
  - **custos monetários e em termos de tempo da realização**

# Tipos de dados:

- **Quanto ao perfil de observação das unidades e dimensão temporal**
  - Temporais
  - Seccionais
  - Painel
- **Quanto ao tipo de unidade estatística observada**
  - Microdados
  - Agregados

# Modelação e análise de dados:

## Tipos de análises

- **Análises de estática**
  - análise da situação num período e/ou comparação entre diferentes unidades estatísticas
- **Análises de estática comparada**
  - comparação da situação em diferentes períodos e/ou em diferentes unidades estatísticas
- **Análises de dinâmica**
  - dados agregados
  - acompanhamento de trajectórias com microdados

# Bases de dados

- **Acesso Livre**
  - Portal INE
  - Portal Banco de Portugal (BPSTAT, etc)
  - Portal Eurostat
  - Portal OCDE
  - Comissão Europeia: DG ECFIN (AMECO, KLEMS, BACH, ...)
  - OMC (Comercio internacional)
  - FMI, Banco Mundial, etc
  - CMVM /Euronext

# Bases de dados

- **Disponíveis terminais ISEG**
  - **Datastream/Reuters**
  - **BANKSCOPE – informação sobre 23,000 bancos**
  - **CHELEM – dados comércio internacional , agregados macroeconómicos e balança pagamentos**
  - **OSIRIS – informação sobre empresas cotadas**

# Bases de dados

- **Disponíveis mediante protocolo**
  - **Protocolo ex-MCTES/INE – acesso microdados**

## **NOTA:**

Obter dados já existentes pode levar tempo: contratos / protocolos / tempo de espera...=> Necessidade de resolver a questão numa fase preliminar da dissertação.

# Software estatístico

- **Escolha de software**

- depende do trabalho a desenvolver e da estrutura da informação estatística que a utilizar
- Três níveis:
  - Excel
  - PASW (ex-SPSS), Stata, SAS, TSP, EViews...
  - Gauss

Para uma discussão dos níveis de popularidade dos vários programas consultar: **The Popularity of Data Analysis Software** by Robert A. Muenchen ( <http://r4stats.com/popularity> )

# Software estatístico

- **Trabalhar com SPSS/STATA/SAS**
  - Todos tem um interface com o utilizador geralmente assente num sistema de menus, numa folha de dados, onde são apresentados os dados, e uma janela de output
  - Todos tem igualmente um outro interface que permite ao utilizador escrever e introduzir rotinas, submete-las e obter os resultados
    - Muitos dos comandos mais potentes e/ou menos usuais somente podem ser passados para o programa por esta via
    - Geralmente têm um sistema de interacção entre o sistema de menus e a lógica de programação de rotinas - É possível utilizar o sistema de menus para escrever parte das rotinas

# Software estatístico

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a data table with 28 rows and 16 columns. The columns are: year, hid\_in, pid\_in, ind\_weight, hh\_size, psu, strata, rotation, region, urbanisation, hh\_type, hh\_type2, maininc, and eqscale. The data table is partially obscured by a 'Means' dialog box.

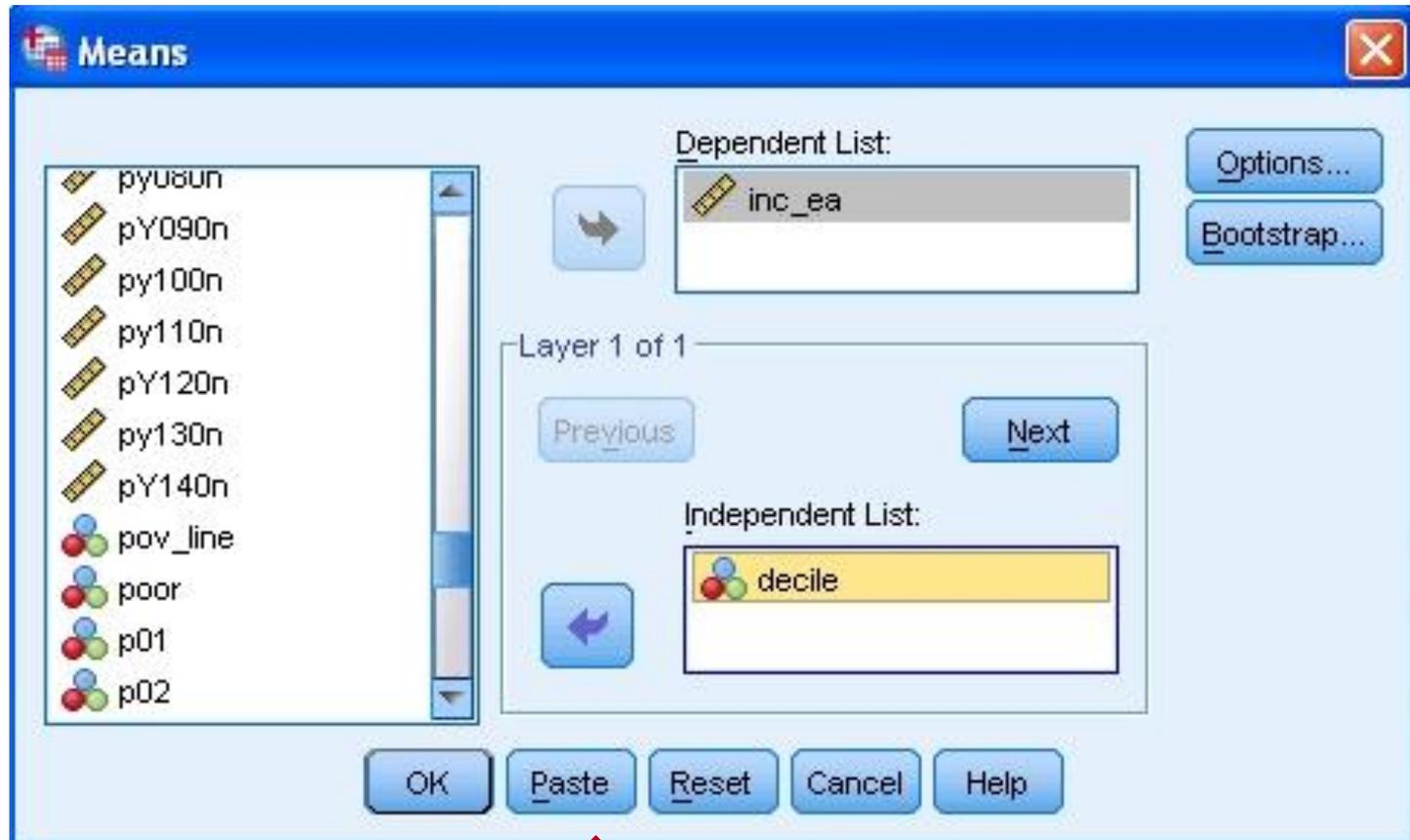
The 'Means' dialog box is open, showing the following configuration:

- Dependent List: inc\_ea
- Independent List: declle
- Layer 1 of 1
- Buttons: Options..., Bootstrap..., Previous, Next, OK, Paste, Reset, Cancel, Help

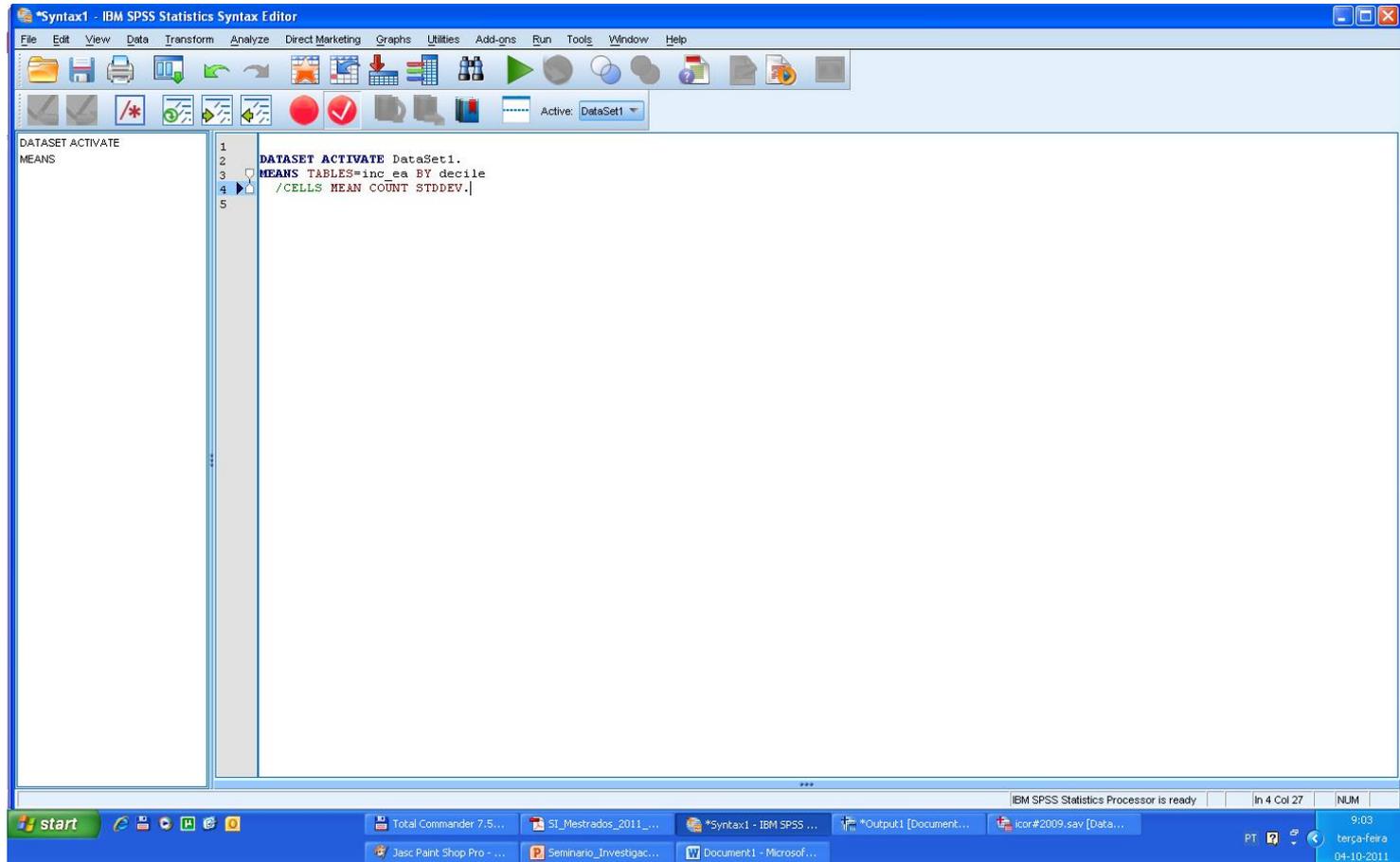
The data table shows the following data for the first 10 rows:

row	year	hid_in	pid_in	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	3	4	1	1	Norte	intermediat...	Two adults, at...	Two adults, at...	Income from p...	1
8	2009	30005510	3000551002	60	3	4	1	1	Norte	intermediat...	Two adults, at...	Two adults, at...	Income from p...	1
9	2009	30005610	3000561001	125	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Income from ...	1
10	2009	30005610	3000561002	125	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Income from ...	1

# Software estatístico



# Software estatístico



# Software estatístico

```
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.
```

# Software estatístico

- **Vantagens da utilização de “ficheiros de comandos”**
  - Uma vez aprendida a linguagem de programação poupa tempo - é mais simples alterar um elemento no programa e mandar executá-lo que repetir toda a sequência do sistema de menus.
  - O programa permite perceber a lógica de tratamento dos dados que foi seguida e a modelização efectuada. O(s) programa(s) permite(m) identificar a estratégia implementada para a parte empírica do trabalho.
  - O mesmo programa pode ser aplicado/adaptado a outros projectos de investigação ou a outras bases de dados.

**NOTA:É possível encontrar muitos procedimentos, rotinas e outros recursos passíveis de serem utilizados no nosso trabalho empírico ‘on-line’**

# Questões éticas

- **Regras de conduta na investigação:**
  - **Propriedade intelectual**
  - **Os resultados obtidos têm de poder ser verificados pela comunidade científica.**
  - **Os resultados apresentados têm de poder ser duplicados por outros investigadores.**
  - **As diferentes hipóteses assumidas quanto ao tratamento dos dados tem de ser apresentadas e justificadas.**

# Questões éticas

- **Regras de conduta quanto à utilização de dados:**
  - **Limitar o uso de dados ao objectivo solicitado**
  - **Mencionar sempre quem cedeu os dados (e a versão com que se está a trabalhar)**
  - **Não utilizar os dados para fins comerciais ou outros não estabelecidos**
  - **Respeitar as regras de confidencialidade e de anonimização**
  - **Destruição dos dados no fim do período estabelecido**