



**MSc in Official Statistics**  
**Statistical Computing:**  
**The RDBMS approach to Databases**

**Andrew Westlake**

Survey & Statistical Computing

63 Ridge Road, London N8 9NP, UK

+44 (0) 20 8374 4723

AJW@SaSC.co.uk (E-Mail)

[www.SaSC.co.uk](http://www.SaSC.co.uk)

# Statistical View of Data

- 'Data matrix', as in SPSS
  - » Columns are Variables
  - » Rows are data values for a respondent (Cases)

1 : province 1 Visible: 23 of 2

	province	area	cluster	h_no	l_no	q03	q04	q05	q06
1	1	2	1	1	1	1	1	1	1
2	1	2	1	1	2	2	1	1	2
3	1	2	1	1	3	3	1	1	2
4	1	2	1	1	4	3	1	1	1
5	1	2	1	1	5	3	1	1	1
6	1	2	1	1	6	3	1	1	1
7	1	2	1	1	7	3	1	1	1
8	1	2	1	2	1	1	1	1	2
9	1	2	1	2	2	3	1	1	1
10	1	2	1	2	3	4	1	1	2
11	1	2	1	2	4	3	1	1	2
12	1	2	1	2	5	3	1	1	1
13	1	2	1	2	6	3	1	1	2
14	1	2	1	2	7	3	1	1	1
15	1	2	1	3	1	1	1	1	1
16	1	2	1	3	2	2	1	1	2
17	1	2	1	3	3	3	1	1	1
18	1	2	1	3	4	3	1	1	1

Data View / Variable View

Case counter area SPSS Processor is ready

# Statistical Variables

- Data Dictionary for details
  - » Variable and Value labels, etc., in SPSS
  - » Statistical example of Metadata

The screenshot shows the SPSS Data Editor window for a file named \*HM.sav. The main window displays a variable dictionary table with columns for Name, Type, Width, Decimals, Label, Values, Missing, and Co. A dialog box titled 'Value Labels' is open in the foreground, showing a list of value labels for a variable. The list includes: 1 = "Head", 2 = "Wife or Husband", 3 = "Son or Daughter", 4 = "Son-in-Law or Daughter-in-Law", 5 = "Grand Child", 6 = "Parent", 7 = "Parent-in-Law", 8 = "Brother or Sister", 9 = "Other Relative", 10 = "Adopted/Foster/Step Child", 11 = "Not Related", and 98 = "Don't Know".

	Name	Type	Width	Decimals	Label	Values	Missing	Co
1	province	Numeric	1	0	PROVINCE	{1, NWFP }...	9	8
2	area	Numeric	1	0	TYPE OF AREA	{1, Major Urban}...	9	8
3	cluster	Numeric	3	0	CLUSTER	None	999	8
4	h_no	Numeric	2	0	HOUSEHOLD NUMB	None	99	8
5	l_no	Numeric	2	0	Line Number	None	99	8
6	q03	Numeric	2	0	Relationship to the He	{1, Head}...	99	8
7	q04	Numeric	1	0				
8	q05	Numeric	1	0				
9	q06	Numeric	1	0				
10	q07	Numeric	2	0				
11	q08	Numeric	1	0				
12	q09	Numeric	1	0				
13	q10	Numeric	2	0				
14	q11	Numeric	1	0				
15	q12	Numeric	1	0				
16	q13	Numeric	2	0				
17	q14	Numeric	1	0				
18	q15	Numeric	2	0				
19	place	Numeric	5	0				

**Value Labels**

Value Labels

Value:

Label:

Add

Change

Remove

1 = "Head"  
2 = "Wife or Husband"  
3 = "Son or Daughter"  
4 = "Son-in-Law or Daughter-in-Law"  
5 = "Grand Child"  
6 = "Parent"  
7 = "Parent-in-Law"  
8 = "Brother or Sister"  
9 = "Other Relative"  
10 = "Adopted/Foster/Step Child"  
11 = "Not Related"  
98 = "Don't Know"

OK

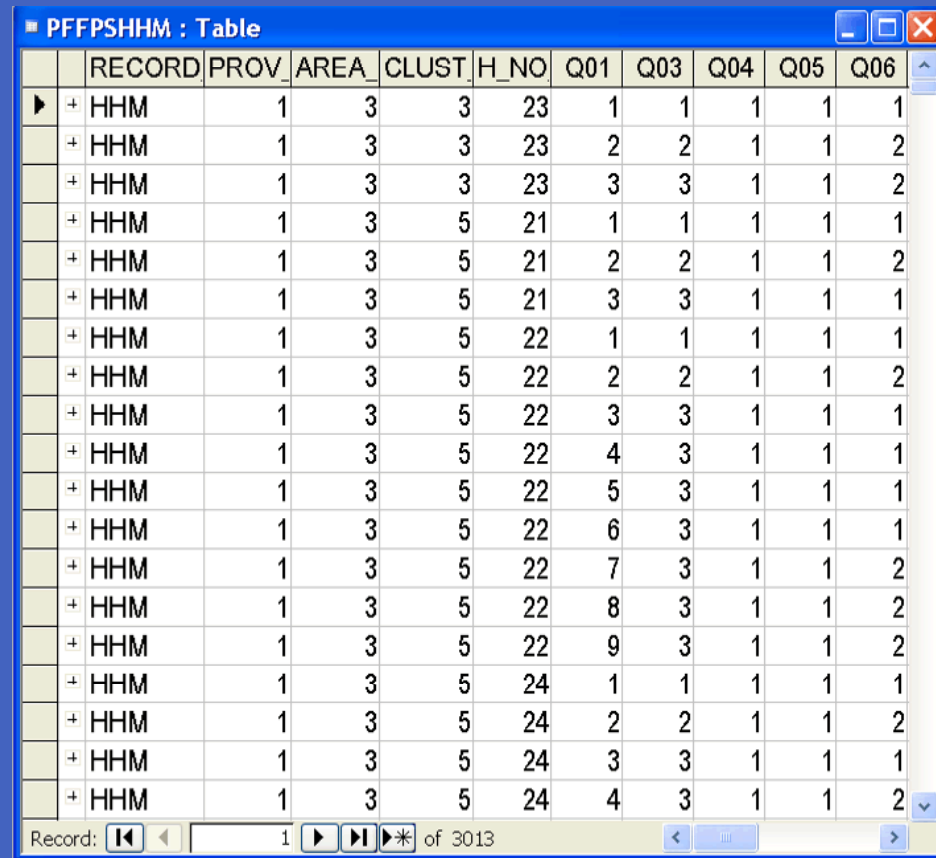
Cancel

Help

SPSS Processor is ready

# RDBMS Dataset

- Very similar to Data Matrix
  - » Same structure, different names
  - » Dataset is a Table
  - » Columns are Attributes (or Fields)
  - » Rows are Records



	RECORD	PROV	AREA	CLUST	H_NO	Q01	Q03	Q04	Q05	Q06
▶ +	HHM	1	3	3	23	1	1	1	1	1
+	HHM	1	3	3	23	2	2	1	1	2
+	HHM	1	3	3	23	3	3	1	1	2
+	HHM	1	3	5	21	1	1	1	1	1
+	HHM	1	3	5	21	2	2	1	1	2
+	HHM	1	3	5	21	3	3	1	1	1
+	HHM	1	3	5	22	1	1	1	1	1
+	HHM	1	3	5	22	2	2	1	1	2
+	HHM	1	3	5	22	3	3	1	1	1
+	HHM	1	3	5	22	4	3	1	1	1
+	HHM	1	3	5	22	5	3	1	1	1
+	HHM	1	3	5	22	6	3	1	1	1
+	HHM	1	3	5	22	7	3	1	1	2
+	HHM	1	3	5	22	8	3	1	1	2
+	HHM	1	3	5	22	9	3	1	1	2
+	HHM	1	3	5	24	1	1	1	1	1
+	HHM	1	3	5	24	2	2	1	1	2
+	HHM	1	3	5	24	3	3	1	1	1
+	HHM	1	3	5	24	4	3	1	1	2

Record: 1 of 3013



# RDBMS Variables

- Variable definition through a table
- Value labels can be supported through forms

PFPSHHM : Table

Field Name	Data Type	Description
RECORD_TYP	Text	
PROV_HHM	Number	
AREA_HHM	Number	
CLUST_HHM	Number	
H_NO_HHM	Number	
Q01	Number	
Q03	Number	
Q04	Number	
Q05	Number	
Q06	Number	

Field Properties

General    Lookup

Field Size	Double
Format	
Decimal Places	Auto
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Indexed	No
Smart Tags	

A field name can be up to 64 characters long, including spaces. Press F1 for help on field names.

Dictionary : Form

Record    **HHM**    Household Members

From	Field	Type	Start	Length	DP	Class	Missing	NA	Verify	Ranges
	<b>Q03</b>	N	13	2	0	N	99		N	1 - 11, 98

Relationship to the Head

1 Head	Test if Field Q01 = Value 01
2 Wife or Husband	Or if Field Q03 = Value 01
3 Son or Daughter	And if Field Q03 # Field Q01
4 Son-in-Law or Daughter-in-Law	Error message is "E31:Q03,Q01: HoH should be line 01 " when test is True
5 Grand Child	
6 Parent	
7 Parent-in-Law	
8 Brother or Sister	
9 Other Relative	

Record:    ⏪    ⏩    43    ⏪    ⏩    ⏭    of 368



# Contrasts - Statistical Packages

- Strengths
  - » Statistical Methods
  - » Integration of Statistical meta-data
  - » Missing Value treatment
  - » Familiar terminology
- Weaknesses
  - » Handling multiple datasets, e.g. hierarchies
  - » Data editing - cases and variables
  - » Security and Auditing

# Contrasts - RDBMS

- Strengths
  - » Standardised data manipulation and linking
    - Integrated Null concept
  - » Standardised access from other systems, including Statistical Packages
  - » Integrated Programming, Security, Integrity, ...
  - » Rich Data Design Methodologies
- Weaknesses
  - » No statistical methods, limited aggregation functionality
  - » No statistical meta-data, missing value
    - Some functionality can be programmed, but not standard
  - » No treatment of macro data
    - Various extensions, but not standard

# Roles of RDBMS and Statistical Packages

- Don't choose between, use strengths of both
  - » RDBMS for data storage and management
  - » Statistical package for analysis and presentation
- Rich RDBMS concepts
  - » Useful way to think about all data
  - » Particularly valuable for complex situations
- Example of PFFPS
  - » Pakistan Fertility and Family Planning Survey
  - » Also used in Exercises

