

# Curs bàsic d'Anàlisi de dades amb Stata

Santiago Pérez Hoyos



UNITAT  
D'ESTADÍSTICA I  
BIOINFORMÀTICA



# Contingut

- Sessió 1
  - Introducció al Stata
  - Gestió d'arxius amb Stata
  - Manipulación de datos con Stata
  - Exercici pràctic
- Sessió 2
  - Estadística descriptiva
  - Grandaria Mostral
  - Exercici pràctic
- Sessió 3
  - Estimació i Contrast d'Hipòtesi
  - Correlació i Regressió
  - Exercici pràctic
- Sessió 4
  - Regressió lineal
  - Regressió logística
  - Anàlisi de supervivència

# Sessió 1

- Introducció al Stata
  - **Característiques generals**
  - **Menús**
  - **Ajuda**
  - **Forma de treball en Stata**
- Gestió d'arxius en Stata
  - **Entrada de dades**
  - **Obrir i desar dades**
  - **Combinar dades**
- Manipulació de dades amb Stata
  - **Definir i etiquetar variables**
  - **Transformar i recodificar variables**
  - **Crear noves variables**
  - **Control de duplicats**
- Exercici pràctic

# Introducción

- Stata programa estadístico disponible para diversos sistemas operativos
- Fácil manejo de datos con mucha versatilidad para combinar y generar nuevos datos
- Numerosos tipos de análisis estadísticos sencillos y complejos con posibilidad de modificarlos y añadir nuevos métodos elaborados por los usuarios
- Muy utilizado en ambientes epidemiológicos
- Puede trabajar por menú, pero es mejor trabajar por comandos que se ejecutan al instante, dispone de una ayuda exhaustiva y completa y fácil de generar funciones o trabajar con programas que ejecuten varias ordenes a la vez

# Extensiones de los ficheros de Stata

- .dta: Ficheros de datos en formato STATA
- .log: Fichero de texto con resultados
- .do: Fichero con instrucciones STATA
- .ado: Ficheros con macro/funciones de Stata
- .gph: Ficheros de gráficos

# Algunas cosas que hay que saber

- Stata distingue entre mayúsculas y minúsculas. No es lo mismo **var1** que **Var1**
- El directorio por defecto es **c:\data**
- Los comandos pueden ser acortados a 3 primeras letras
- Se debe actualizar el Stata de vez en cuando  
**update all**

# Ventanas de STATA

The screenshot shows the STATA 18.0 interface with the following components and annotations:

- Historial (Left Panel):** Labeled "INSTRUCCIONES EJECUTADAS" in purple. It shows a list of commands executed, with the first command being `use "e:\18949893d\Nuevo Equipo V...`.
- Command Window (Bottom):** Labeled "INSTRUCCIONES" in purple. It contains the command `use "e:\18949893d\Nuevo Equipo VHIR10 Dropbox\Santi Perez-Hoyos\UEB-compartida\__FORMACIO\Formacio_externa\Pere_Virgili\Stata\cohortvih.dta"`.
- Main Window (Center):** Displays the STATA startup screen with the text "STATA 18.0 SE-Standard Edition" and "Statistics and Data Science". A yellow arrow points from the "RESULTADOS" label to this window.
- Variables Panel (Right Panel):** Labeled "VARIABLES DISPONIBLES EN LA BASE DE DATOS" in purple. It lists available variables and their labels. A yellow arrow points from the "RESULTADOS" label to this panel.

**RESULTADOS**

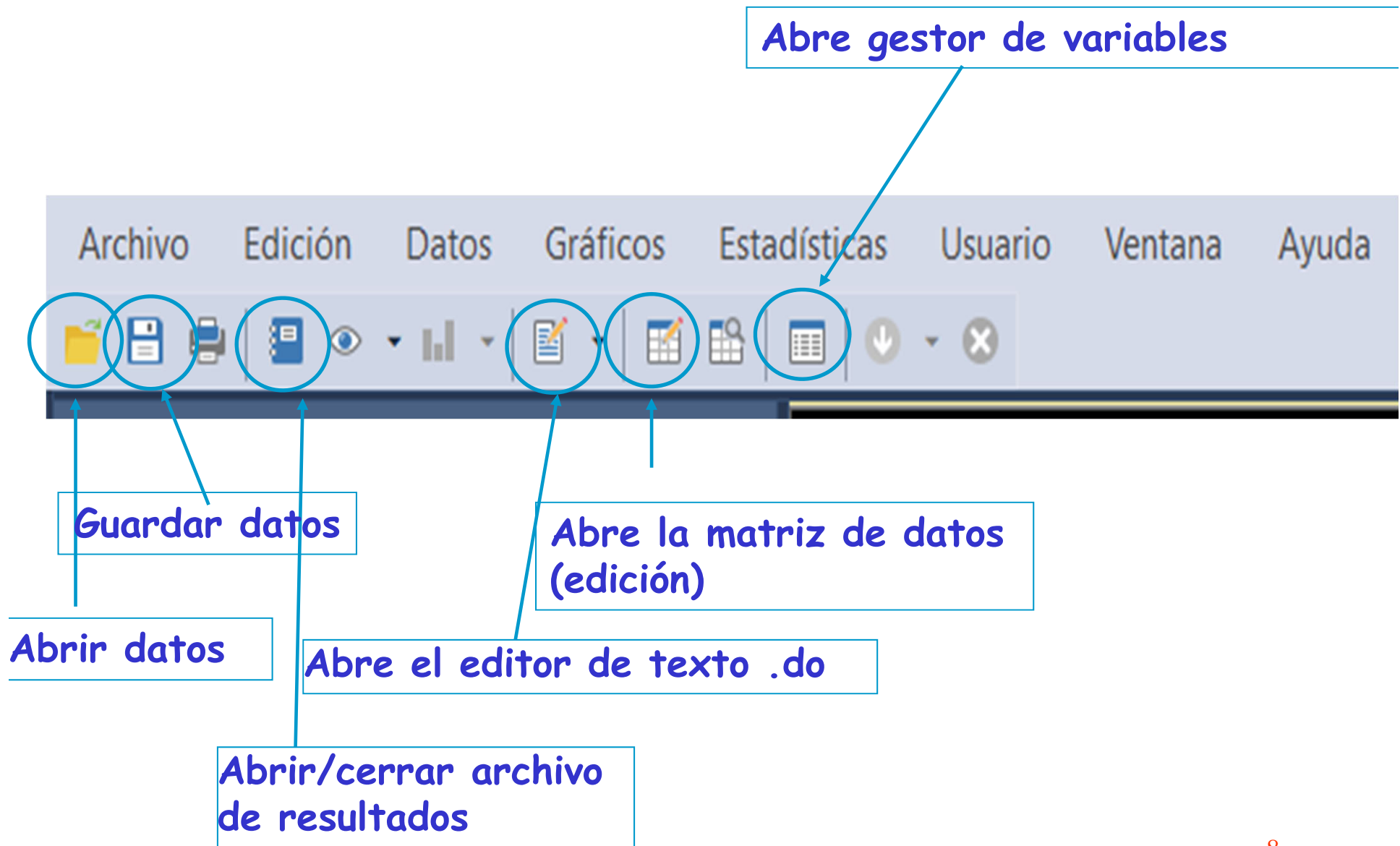
Nombre	Etiqueta
ident	
sex	sexe
dob	date of birth
expcateg	exposure category
firstpos	date first hiv+
lastneg	date last hiv-
serodate	date estimated serocon
neghow	determination hiv-
aids	aids diagnosis
aidsdate	date initial aids diagno
lastclin	date last known to be a
died	subject died
dieddate	date of death

Propiedades	
Variables	
Nombre	
Etiqueta	
Typo	
Format	
Etiqueta de valor	
Nota	
Datos	
Marco de datos	default
Nombre de archivo	cohortvih.dta
Etiqueta	
Notas	
Variables	32
Observaciones	447
Tamaño	55.00K

VARIABLES  
DISPONIBLES EN LA  
BASE DE DATOS

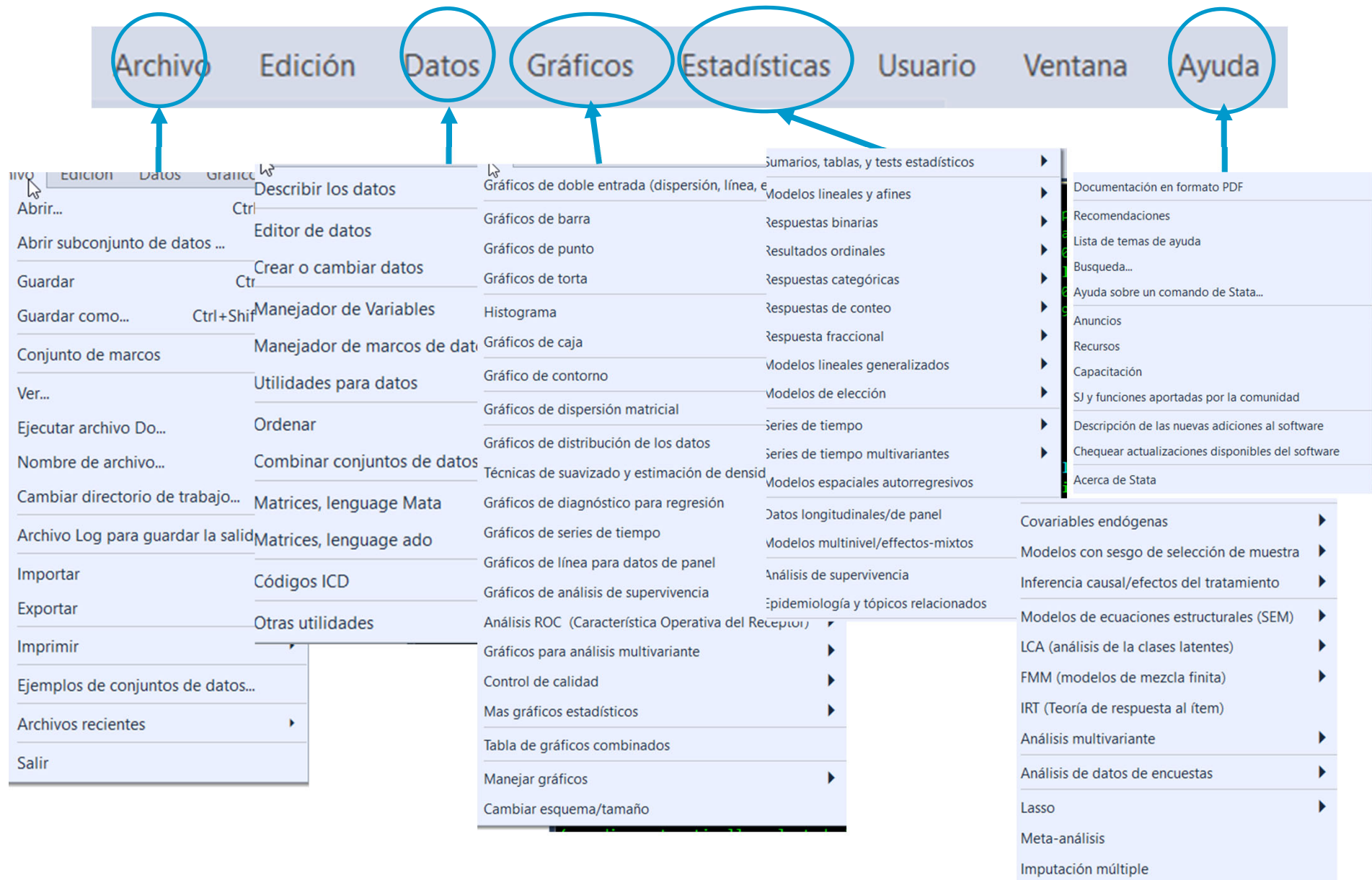


# Barra de botones de Stata



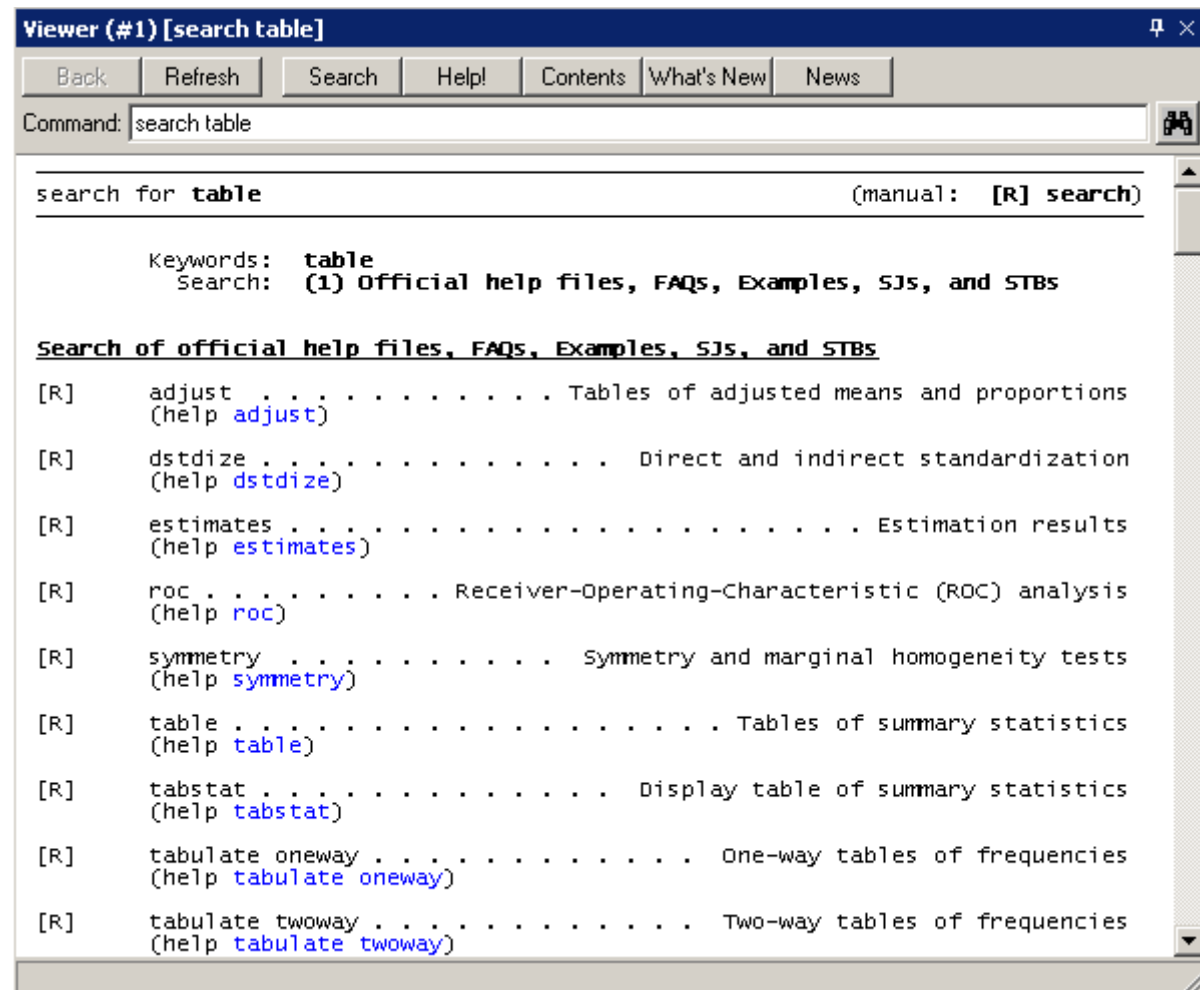


# Menu de Stata



# Ayuda

- Help comando  
**help table**

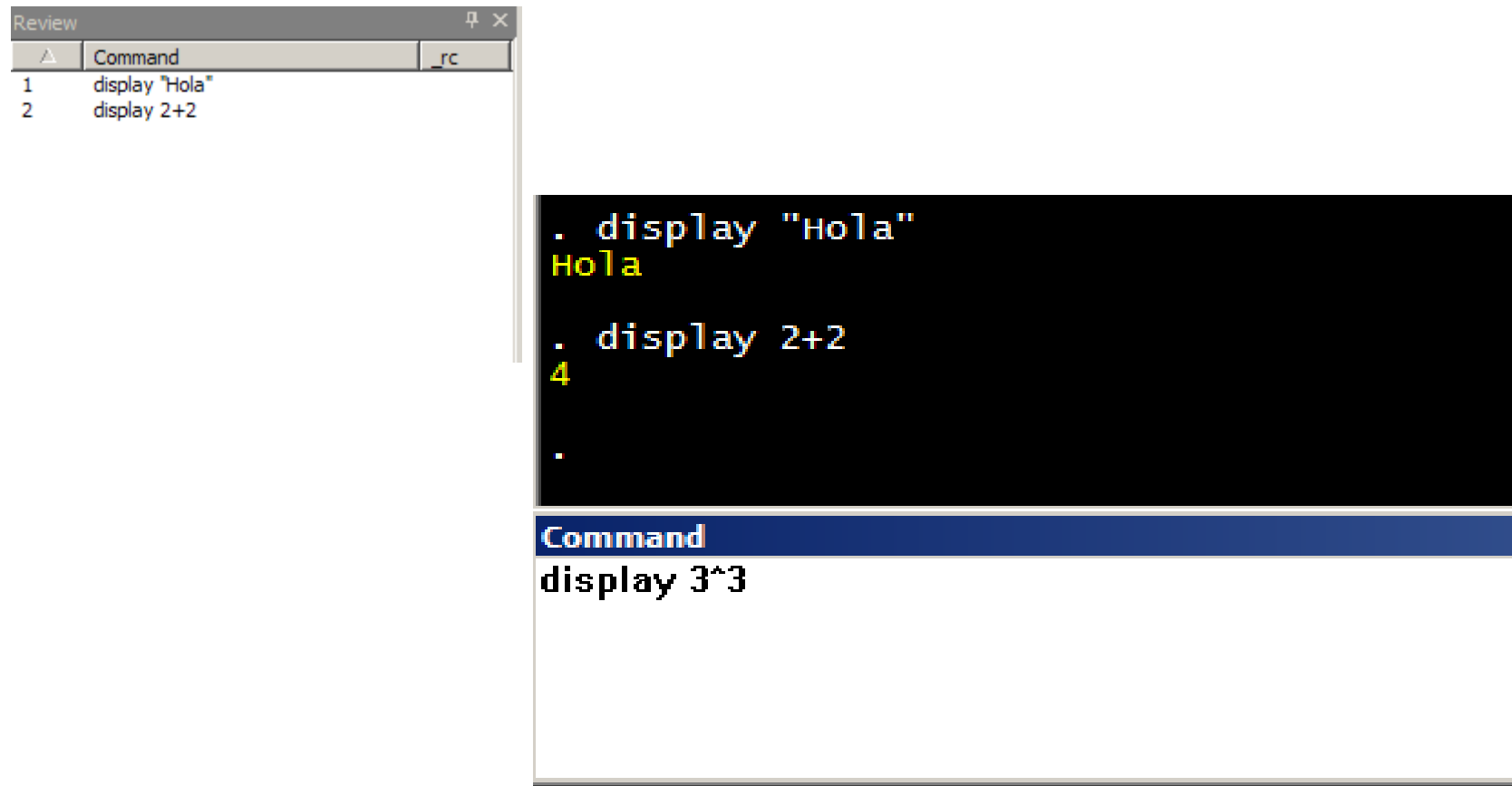


# Modos de trabajar en Stata

- Escribiendo instrucciones en la línea de comandos ejecutando una a una y viendo el resultado por pantalla sin guardarlo
- Escribiendo varias instrucciones en un fichero .do y ejecutándolas en lote
- Es la forma óptima de trabajar

# Escribiendo instrucciones en línea

- Se puede utilizar como una calculadora
- Los comandos ejecutados previamente se pueden recuperar utilizando la tecla RePàg o clickando sobre el en la ventana de comandos



The screenshot displays a software interface with two main components. On the left, a 'Review' window with a tab labeled 'Command' shows a list of executed commands: '1 display "Hola"' and '2 display 2+2'. On the right, a larger black window shows the output of these commands: '. display "Hola"' followed by 'Hola' in yellow, and '. display 2+2' followed by '4' in yellow. Below this black window is a blue header bar labeled 'Command', and underneath it, the text 'display 3^3' is visible, indicating the next command to be entered or executed.

```
Review
┌ Command ─┴─ rc ─┐
1 display "Hola"
2 display 2+2

. display "Hola"
Hola
. display 2+2
4
.

Command
display 3^3
```

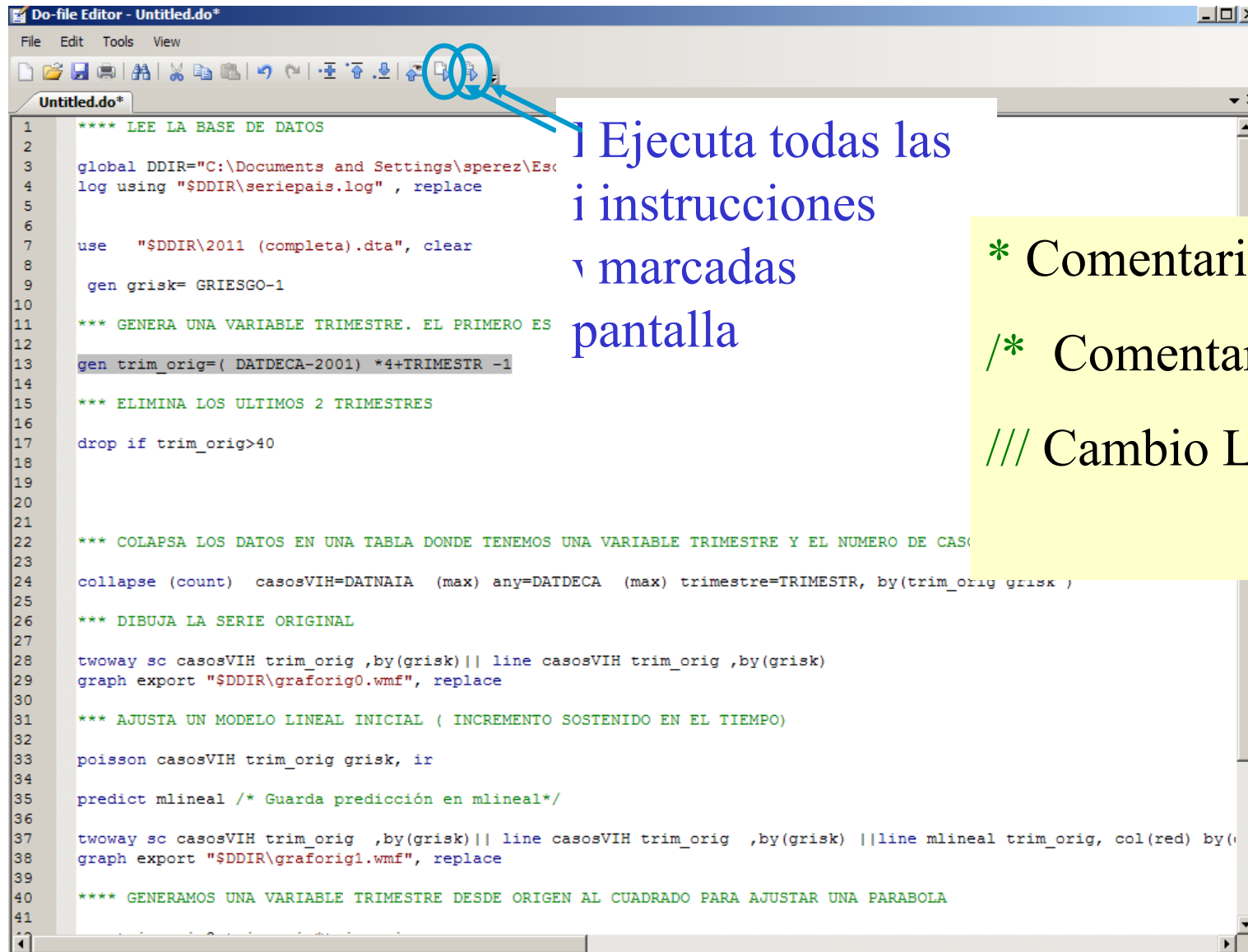
# Estructura de los Comandos

**comando**     *lista de variables*  
**condición (if) , opciones**

♦Ejemplos:

```
tabulate grupedad sexo , row col  
gen edad_15=edadsero-15  
drop if cd4>500  
xi:poisson iam i.estrés i.sexo, exp(perany)
```

# Fichero Do



Do-file Editor - Untitled.do\*

File Edit Tools View

Untitled.do\*

```
1 **** LEE LA BASE DE DATOS
2
3 global DDIR="C:\Documents and Settings\sperez\Escritorio"
4 log using "$DDIR\seriepais.log" , replace
5
6
7 use "$DDIR\2011 (completa).dta", clear
8
9 gen grisk= GRIESGO-1
10
11 *** GENERA UNA VARIABLE TRIMESTRE. EL PRIMERO ES EL SEGUNDO
12
13 gen trim_orig=( DATDECA-2001) *4+TRIMESTR -1
14
15 *** ELIMINA LOS ULTIMOS 2 TRIMESTRES
16
17 drop if trim_orig>40
18
19
20
21
22 *** COLAPSA LOS DATOS EN UNA TABLA DONDE TENEMOS UNA VARIABLE TRIMESTRE Y EL NUMERO DE CASOS
23
24 collapse (count) casosVIH=DATNAIA (max) any=DATDECA (max) trimestre=TRIMESTR, by(trim_orig grisk )
25
26 *** DIBUJA LA SERIE ORIGINAL
27
28 twoway sc casosVIH trim_orig ,by(grisk)|| line casosVIH trim_orig ,by(grisk)
29 graph export "$DDIR\graforig0.wmf", replace
30
31 *** AJUSTA UN MODELO LINEAL INICIAL ( INCREMENTO SOSTENIDO EN EL TIEMPO)
32
33 poisson casosVIH trim_orig grisk, ir
34
35 predict mlineal /* Guarda predicción en mlineal*/
36
37 twoway sc casosVIH trim_orig ,by(grisk)|| line casosVIH trim_orig ,by(grisk) ||line mlineal trim_orig, col(red) by(grisk)
38 graph export "$DDIR\graforig1.wmf", replace
39
40 **** GENERAMOS UNA VARIABLE TRIMESTRE DESDE ORIGEN AL CUADRADO PARA AJUSTAR UNA PARABOLA
41
```

Ejecuta todas las  
instrucciones  
marcadas  
pantalla

\* Comentarios

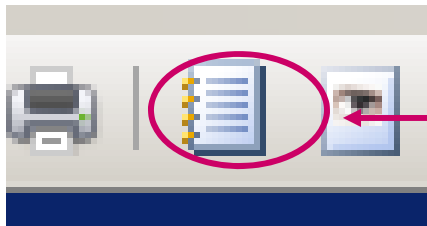
/\* Comentarios \*/

/// Cambio Linea

# Guardar resultados

- Todos los resultados se pueden guardar un fichero de resultados.
- Por defecto se graban en formato .smcl y sólo se ven desde el visor
- Si se quiere ver en otro formato se debe de usar el formato texto

```
log using nombrefichero           [abre fichero .smcl]
log using nombrefichero, replace  [reemplaza fichero]
log using nombrefichero, append   [añade a fichero]
log using nombrefichero.log , text [empieza fichero texto]
....
log off                           [pausa fichero texto]
log on                            [reinicia fichero texto]
log close                         [cierra el fichero de resultados]
```



Activa y desactiva  
fichero que guarda  
resultados



# Gestión de bases de datos

- Se pueden introducir datos directamente con el editor de Stata
- Mejor cargar fichero transferido con Stattransfer o grabarlo como .dta por otro programa (i.e. SPSS)
- Se puede cargar por el menu o con sintaxis
- Se puede importar directamente desde excel o access usando ODBC

```
use nomfichero, clear
```

- Excel

```
import excel "C:\EST.xls", sheet("GeneralFV1$") clear  
datestring lower firstrow
```

- .CSV

```
import delimited "C:/EST.csv", clear  
delimiter("comma") varnames(1)
```

# Gestión de bases de datos

- Para grabar ficheros se usa el menu o la sintaxis

```
save nomfichero, replace
```

- Exportar ficheros a excel

```
odbc insert, dsn("Excel Files;DBQ=C:/EST.xls")  
table("GeneralFV1$") create quoted
```

```
outsheet using "C:/EST.xls", delimiter(";") replace
```

- TRUCOS

Definir el directorio de trabajo, para no tener que escribir cada vez la ruta

```
global Ddir "C:\GEMES\Datagemes_2011\sandoval\"  
use "$DDir\Sandoval_2011.dta", clear
```

```
. . .
```

```
. . .
```

```
save "$DDir\Sandoval_2011.dta", replace
```

Ó cambiar de directorio

```
cd C:\GEMES\Datagemes_2011\sandoval
```

# Inspeccionar datos

- Para ver que todo es correcto se puede ver la estructura de los datos  
**describe**  
**describe** var1 var2...
- Se puede tener una pequeña descripción que permite ver si hay datos extraños y una pequeña descripción (frecuencia valores, medias, etc.)  
**codebook**  
**codebook** var1 var2
- Si todo es correcto ya estamos en condiciones de empezar el análisis estadístico propiamente dicho

# Gestión de bases de datos

- La base de datos se ordena con  
**sort** *var1 var2...*  
**gsort** *-var1 +var2...*
- Las características de la base de datos se miran con  
**describe** [lista nombre variables y etiquetas]  
**codebook** *var1* [lista nombre, etiquetas y datos descriptivos]
- La tabla de datos se puede ver con  
**browse**  
**browse** *var1 var2*
- Y se puede ver y modificar con  
**edit**  
**edit** *var1 var2...*
- Los datos se listan con  
**list**  
**list** *var1 var2...*

# Gestión de bases de datos

- Para borrar variables  
`drop var1 var2`
- Para borrar casos  
`drop if condición`
- Para mantener variables  
`keep var1 var2`
- Para mantener casos  
`keep var1 var2 if condición`
- Repite comandos en un subconjunto de datos  
`by var1, sort: comando stata`
- TRUCO

*Genera un indicador del número de medición por paciente*

`by paciente, sort: gen nvisita=_n`

*Mantiene el primer caso de cada paciente*

`by paciente, sort: keep if _n==1`

`_n` = Número de registro

`_N` = Número total de casos

# Gestión de bases de datos

- Para añadir casos a un fichero existente

```
use nomfile1, clear  
append using nomfile2  
save nomfile1+2, replace
```

- Para añadir variables a un fichero existente

```
use nomfile1, clear  
sort variableclave  
merge 1:1 variableclave using nomfile2, sort  
merge m:1 variableclave using nomfile2, sort  
merge 1:m variableclave using nomfile2, sort
```

Añade una variable interna `_merge` que se codifica como sigue

---

1	<b>master</b>	La observación aparece sólo en el fichero 1 (master)
2	<b>using</b>	La observación aparece sólo en el fichero 2 (using)
3	<b>match</b>	La observación aparece en los dos ficheros

```
keep if _merge==3  
drop _merge  
save nomfilevar1+2, replace
```

# Gestión por menú de la variables

The screenshot shows a software interface for managing variables. The main window, titled '2 - Manejador de variables', contains a table of variables. A search bar at the top left allows filtering variables. A secondary bar above the table allows grouping by column. The table has columns for #, Nombre, Etiqueta, Tipo, Formato, Etiqueta de valor, and Notas. The variable 'ident' is selected, and its properties are shown in a sidebar on the right titled 'Propiedades de las variables'. The sidebar includes fields for Nombre, Etiqueta, Tipo, Formato, Etiqueta de valor, and Notas, each with a corresponding input field and a 'Manejar...' button. At the bottom of the sidebar are navigation buttons: '<', '>', 'Resetear', and 'Aplicar'. The status bar at the bottom indicates 'Activo' and 'Vars: 29 CAP NUM'.

#	Nombre	Etiqueta	Tipo	Formato	Etiqueta de valor	Notas
	FERRITIN	cifra de ferritina previa al in...	double	%10.0g		
	NHC	Número de historia clínica	byte	%8.0g		
	CENTRO		byte	%8.0g	CENTRO	
	NHC_CORT		int	%8.0g		
	ident	iniciales del paciente	str4	%4s		
	UPN	UPN	byte	%8.0g		
	TPH_DATE	fecha de trasplante	long	%dD_m_Y		
	TRANSFER	cifra de transferrina previa ...	int	%8.0g		
	TRANSFE	saturation de transferrina p...	byte	%8.0g		
	SIDEREMI	cifra de sideremia preTPH	double	%10.0g		
	TBIC_PRE	capacidad total de union a ...	double	%10.0g		
	PCR_PRE	Cifra de proteína C reactiva ...	double	%10.0g		
	VSG_PRE	VSG pre TPH (solo si PCR n...	int	%8.0g		
	QUELACIO	Se realizó quelacion pre TP...	byte	%8.0g		
	QUELA_FA	Con que fármaco se queló?	str1	%1s		
	QUELA_PO	Se realizó quelacion post T...	byte	%8.0g		
	QUELA_A	Con que fármaco se queló?	str1	%1s		

Propiedades de las variables

Nombre: ident

Etiqueta: iniciales del paciente

Tipo: str4

Formato: %4s [Crear...]

Etiqueta de valor: [Manejar...]

Notas: Sin notas [Manejar...]

< > [Resetear] [Aplicar]

Activo Vars: 29 CAP NUM

. label variable NHC "Número de historia clínica"



# Gestión de variables

- Para renombrar variables

```
rename nomvarviejo nomvarnuevo
```

- Para etiquetar la base de datos

```
label data "contenido de la base"
```

- Para etiquetar variables

```
label var var1 "etiqueta de la variable"
```

- Para etiquetar valores

```
label define nomormato valor1 "etiql" valor2 "etiql2"
```

```
label val variable nomformato
```

- Para asignar formato a las variables

```
format varlist %fmt
```

- Truco (código para cambiar nombre variables a minúsculas)

```
unab listavar:*
```

```
foreach var of varlist `listavar' {
```

```
  cap ren `var' `=lower("`var`")'
```

```
}
```

# Formato de variables

%fmt	description	example
-----		
Right-justified formats		
<b>%#.#g</b>	<b>general numeric format</b>	<b>%9.0g</b>
%#.#f	fixed numeric format	%9.2f
%#.#e	exponential numeric format	%10.7e
%d	default numeric elapsed date format	%d
<b>%d...</b>	<b>user-specified elapsed date format</b>	<b>%dM/D/Y</b>
<b>%#s</b>	<b>string format</b>	<b>%15s</b>
Right-justified, comma formats		
%#.#gc	general numeric format	%9.0gc
%#.#fc	fixed numeric format	%9.2fc
Leading-zero formats		
%0#.#f	fixed numeric format	%09.2f
%0#s	string format	%015s
Left-justified formats		
%-#.#g	general numeric format	%-9.0g
%-#.#f	fixed numeric format	%-9.2f
%-#.#e	exponential numeric format	%-10.7e
%-d	default numeric elapsed date format	%-d
%-d...	user-specified elapsed date format	%-dM/D/Y
%-#s	string format	%-15s
Left-justified, comma formats		
%-#.#gc	general numeric format	%-9.0gc
%-#.#fc	fixed numeric format	%-9.2fc
Centered formats		
%~#s	string format (special)	%~15s
-----		

# Creación de variables

- Para generar nuevas variables (*ver help functions*)

```
gen nomnuevavar = expresión
```

```
gen nomnuevavar = expresión if condiciónlogica
```

- Para reemplazar valores en una variable existente

```
replace nomvar = expresión if condiciónlogica
```

- Operadores lógicos

Igual (**==**), mayor (**>**), mayor o igual (**>=**), menor (**<**), menor o igual (**<=**), diferente (**!=**)

- *TRUCOS*

```
gen var_sino=(var1==valor) /* Genera variable 0=no 1=si */
```

```
gen data_nac = mdy(mes,dia,any) /* crea variable fecha */
```

```
gen num_ident = _n /* 1 número identificación por caso*/
```

```
gen random= uniform() /* n° aleatorios entre 0 y 1 */
```

```
gen varnum= real(vartexto) /* convierte var texto en n° */
```

```
gen varnoblancos= trim(vartexto) /*elimina texto en blanco */
```

```
gen seletxt= substr(vartexto,pos,len) /*selecciona texto de
```

longitud len desde la posición pos en la variable texto

i.e. **substr**("12/11/2012",1,2)=12

**substr**("12/11/2012",4,2)=11 \*/

# Creación de variables

- Para recodificar variables existentes

**recode variable sint1 sint2..., generate(varnueva)**

+-----+		
<i>sintaxis</i>	<i>Ejemplo</i>	<i>Significado</i>
+-----+		
# = #	3 = 1	3 recodifica a 1
# # = #	2 . = 9	2 Y . recodifican a 9
#/# = #	1/5 = 4	De 1 a 5 recodifican a 4
<i>nonmissing</i> = #	<i>nonmiss</i> = 8	resto de no perdidos a 8
<i>missing</i> = #	<i>miss</i> = 9	resto de perdidos a 9
<i>else</i> = #	<i>else</i> = 9	resto a 9
+-----+		

- Convertir variables en números cuando el contenido es numérico  
**destring variable\_txt, replace**
- Generar variable numérica con etiquetas a partir de variable texto  
**encode var\_txt, gen(var\_num) label**

# Creación de variables

- Para repetir por subgrupos de datos

**by** *vargrupo*: **gen** *nomnuevavar* = *expresión*

**by** *vargrupo*: **comando análisis**

- Genera variables especiales

**egen** *nomvar* = *función(argumentos)* , *opciones*

**anycount**(*varlist*), **values** (*numlist*)    [cuenta apariciones de valores en variables]

**cut**( *varname*), **at**(#, #, ..., #)    [categoria en grupos]

**group**(*var1 var2 ....*)    [combina 2 variables]

**rowmean**(*varlist*)    [calcula la media de las variables de la lista]

**rowmax**(*min/total* )( *varlist*)    [elige el maximo(minimo/suma)]

- Para generar variables dummy o ficticias

**tabulate** *variablecat*, **gen**(*vardummy*)

- Para definir valores perdidos

**mvdecode** *var*, **mv** (*valor*)

**mvdecode** *\_all*, **mv** (*valor*)

# Control de duplicados

- Para identificar casos duplicados

**duplicates report** *variables* [tabla casos duplicados]

**duplicates list** *variables* [lista casos duplicados]

**duplicates drop** *vars, force* [elimina casos duplicados]

**duplicates tag** *vars , gen(nvar)* [marca casos duplicados]

- Para duplicar casos

**expand # if** *condición*

# Girar ficheros

- Convertir columnas en filas

```
reshape long inc , i(id) j(year)
```

- Convertir filas en columnas

```
reshape wide inc , i(id) j(year )
```

(long form)

(wide form)

i		..... <u>x_ij</u> .....		
id	sex	inc80	inc81	inc82
-----				
1	0	5000	5500	6000
2	1	2000	2200	3300
3	0	3000	2000	1000

i	j		<u>x<sub>ij</sub></u>
id	year	sex	inc
-----			
1	80	0	5000
1	81	0	5500
1	82	0	6000
2	80	1	2000
2	81	1	2200
2	82	1	3300
3	80	0	3000
3	81	0	2000
3	82	0	1000